Injuries in Professional and Collegiate Sports: Identifying Athletic Predisposition and Susceptibility to Injury and Analyzing Modern Incorporation of Biomedical Technology

Peter Chapman

University of Delaware

STAT468: Senior Research Project

Rough Draft

Professor Bryan Crissinger

November 21, 2025

Introduction

Research Question

Professional and collegiate sports are of heavy importance in modern culture, especially to athletes, coaches, athletic trainers, and fans. One unfortunate downside to these two levels of athletics is the multitude of injuries that occur to athletes, especially in contact sports. It is generally understood that there is a multifactorial nature to injuries, which limits the ability to predict the likelihood of injury for a given athlete. Despite this limitation, there recently has been a large emphasis on identifying injury risk and prevention within these two levels of athletics. Naturally, this research trend has become most popular in mainstream sports such as football, basketball, soccer, and baseball. A series of related statistical research questions arise from this growing trend. First, are there predispositions to certain injuries for athletes of a certain size, position, sport, or activity? If these predispositions exist, what is the full extent to the susceptibility to injury? The second major string of questions addresses biomedical advances within these sports. From a biomedical and biomechanical perspective, what is currently being done to address injuries in professional and collegiate sports? Are the modern advancements in injury risk and prevention effective in reducing the number of injuries, especially serious incidents? This study looks at data and research from multiple sports leagues at the two levels, including the National Football League (NFL), the National Basketball Association (NBA), several major National Collegiate Athletic Association (NCAA) sports, and the English Premier League (EPL) in professional soccer.

Hypotheses

Prior to conducting research and viewing data, the hypotheses for this research project must be identified. Since there are two different series of questions asked in this research project, there are two main hypotheses. The first hypothesis relates to the first two questions about predisposition to injury and the full extent of susceptibility to injury. It is generally believed before conducting the investigation that certain internal and external factors play an important role in determining an athlete's likelihood of injury. Internal factors include anything relating to the athlete as an individual: age, flexibility, genetics, previous injuries, etc. External factors refer to anything outside of an athlete's control that can affect health on the field: playing surface type and conditions, opponent behavior, quality and range of protective equipment, etc. With these factors in mind, it is expected that there are statistically significant predispositions to injury. However, due to the multifactorial nature of injury, it is hypothesized that the full extent of potential susceptibility to injury can only be estimated and not perfectly calculated.

The second major hypothesis relates to the final two questions described above. As mentioned before, it is understood that quantitative research into injury risk and prevention has increased tremendously in previous years. As a product of this trend, coupled with the rise of technology into athletics, it is believed that there have been several advancements from the biomedical and biomechanical industries for athletic health purposes. It is expected that these advancements have been most prevalent for the mainstream sports with a large number of injuries (such as tackle football) or for a recurring injury that happens within a given sport (such as an elbow injury for a pitcher in baseball). There is less certainty in the hypothesis regarding the effectiveness of these biomedical and biomechanical advancements in athletics. However, due to the extensive research into sports health, and the exceptional growth in technological implementation into collegiate and professional athletics, it is optimistically believed that these advancements have reduced the number of injuries, particularly serious ones.

Motivations

The first goal of this research project is to help identify potential causes of injuries in significant professional sports leagues across the world. Professional sports leagues and teams have prioritized research into injury prevention, specifically looking into what factors could potentially be causing certain injuries. This research study hopes to be a part of this movement into preventing injuries by determining what may be causing them.

The second goal of this study is to evaluate the technological advancements that were recently implemented into collegiate and professional athletics. As mentioned before, many leagues and agencies have invested a lot of time and resources into research regarding the creation of tools that can help reduce specific injuries to a certain body part. This study hopes to find and use relevant data to identify how effective these developments have been in reducing the number of injuries for its targeted injury.

Literature Review and Background

Biomedical and Biomechanical Research

Before any data is collected and analyzed, research must be conducted in order to find correct biomedical information on injury prevention and risk. As mentioned before, this subject area has seen a tremendous increase in research attention over the past few years. Databases such as Academic OneFile, Google Scholar, SPORTDiscus, and DELCAT Online Catalog used to find research articles and academic pieces relating to this research topic.

Internal Factors

A study by Zhao Wentao from Sichuan Minzu College includes an investigation into the individual factors that affect the likelihood and severity of injuries in sports. Age, health status, mode, and physical condition all have a significant impact in determining the susceptibility of injury (Wentao). Individuals of older age are at increased risk of developing conditions such as arthritis, muscle atrophy, and more. These conditions subsequently increase the risk of sports injuries. Physical condition refers to metrics such as height, weight, bone density, gender, flexibility, etc. These factors, sometimes together, are all associated with different levels of susceptibility to injury (Wentao). Psychological factors such as self-esteem and self-identity are also important in the likelihood of injury. For example, individuals with depression or low self-esteem are at a greater risk of injury.

A similar study by Simo Taimela, Urho M. Kujala, and Kalevi Osterman echoes the findings of the previous study, specifically about age, physical condition, and psychological factors (Taimela, et al.). This study found other significant factors as well. Notably, an athlete's injury history is a debated topic among medical experts. This study identifies that previous injuries may not be linked to a future injury as long as the first injury was treated correctly (Taimela, et al.). If not, then certain injuries may lead to other injuries in the future, especially if the athlete has injury-prone biological characteristics. From a body size perspective, taller and/or heavier people are at a higher predisposition to injury due to a difference in center of gravity (Taimela, et al.). Those with longer limbs may also have more stress on their joints.

External Factors

A study by R. Bahr and I. Holme of the Oslo Sports Trauma Research Center and the University of Sport and Physical Education also reported the internal factors of the previous two research articles. Additionally, this research study analyzed the potential contributions of external factors as well. Referred to as external risk factors, these include human factors, protective factors, sports equipment, and environment (Bahr and Holme). Human factors act as an umbrella term for any action or event from another individual outside of the injured player. For example, teammates, referees, and opponents can all affect the likelihood of getting injured. Protective equipment, such as shin guards and helmets, includes equipment specifically designed to protect human body parts (Bahr and Holme). Sports equipment includes all other equipment that is used: baseball bats, gloves, batting gloves, skis, hockey sticks, etc. The environmental factors include the weather, snow and ice conditions, field or floor type, and the maintenance of the playing surface. All of these external factors were found to have an impact on the likelihood of an athlete getting injured (Bahr and Holme).

Multifactorial Nature of Injury

This same research study highlighted a concept called the multifactorial nature of injury. This concept states that there are many factors that could be correlated with the occurrence of an injury within an athlete. The cause of an injury in organized athletics cannot be determined due to the sheer number of possibilities behind each injury (Bahr and Holme). Rather, internal and external factors are considered to be correlated with a predisposition to injuries. Because several factors should be considered, a multivariate approach is necessary in studies such as this one that try to analyze injury risk (Bahr and Holme). Being more specific, this study recommends the use of a logistic regression model with multiple predictor variables being considered (Bahr and Holme).

Methodology

Data Collection

Data sets relating to internal and external factors as well as injury outcomes were searched for and found online for the National Basketball Association (NBA), the English Premier League (EPL), the National Football League (NFL) and select NCAA sports. Unless otherwise denoted, each raw data set was downloaded or pulled into Microsoft Excel for data storage.

English Premier League Injury Outcomes

A sample of injury data from professional soccer's English Premier League, called 'Football Player Injury Data', was found from the user 'Kalpesh Kolambe' on Kaggle [5]. This user scraped the injury data using the 'worldfootballR' package in R. Player attributes were taken from the FIFA video games from 2016 to 2021. This raw data set contains 30 columns and 685 observations in total. Internal factors such as age, height, weight, and body mass index (BMI) and external factors such as minutes played, position, matches played, and the number of days injured in the previous season. There are two potential dependent variables of interest. The first is a binary variable representing whether or not the player got injured during the season. The second is the number of days that the player was injured during the season. Appendix A shows the full data dictionary for the raw English Premier League injury data set from 2016 to 2021.

NCAA Injury Outcomes

Injury outcomes for NCAA-sponsored sports were more difficult to find, as collegiate athletic departments and the NCAA itself are required by HIPAA to protect data regarding student-athletes. Fortunately, an injury-related data set was found on Kaggle that protects the identities of student-athletes. The data set, titled ‘College Sports Injury Detection” by an account called ‘Ziya’, shows an anonymous ‘Athlete ID’ instead of sharing the names of student-athletes [3]. This data set of 1000 observations contains 13 columns, including biomedical metrics such as heart rate in beats per minute and respiratory rate in breaths per minute. Other key metrics include the specific sport that was played and the type of activity that was performed during the session (jumping, sprinting, etc.). Most importantly, it contains a binary classification metric that shows whether the student-athlete suffered an injury (denoted with a 1) or stayed healthy (denoted with a 0). All observations and columns were kept in this process, so the data did not need to be cleaned or transformed in a future step. Table 1 below shows a full data dictionary for this data set.

Variable	Description and Units
Athlete_ID	Unique numerical ID for each student-athlete
Sport_Type	The type of sport the athlete is engaged in (soccer, basketball, etc.)
Session_Date	Date of the training session, written in yyyy-mm-dd
Heart_Rate_BPM	Heart rate in beats per minute
Respiratory_Rate_BPM	Respiratory rate in breaths per minute
Skin_Temperature_C	Athlete’s skin temperature in degrees Celsius
Blood_Oxygen_Level_Percent	Percentage of blood oxygen saturation
Impact_Force_Newtons	Force of impact during the training session, measured in Newtons
Cumulative_Fatigue_Index	A calculated index representing the athlete’s fatigue level
Activity_Type	Type of activity performed during the session (Sprinting, Jumping)
Duration_Minutes	Duration of the training session in minutes
Injury_Risk_Score	A calculated score representing the risk of injury
Injury_Occurred	Binary variable indicating whether an injury during the session (1 for injury, 0 for no injury)

Table 1: Data dictionary for the raw NCAA injuries data set from ‘Ziya’ on Kaggle.

External Factors for NFL Injuries

External factors such as playing surface, environmental factors, opponent behavior, and protective equipment are all most relevant in the National Football League (NFL). Each NFL stadium is a different experience for players when these external factors are considered. Teams can play indoors or outdoors, and they are free to choose the playing surface that they wish to play on. If the stadium is outdoors, then the stadium's specific location may lead to specific weather conditions and other environmental conditions. Because the NFL's external factors are most relevant, the search for NFL data sets with external factors was prioritized over other leagues. In particular, the goal was to find a data set with playing surface and number of days missed for each injury. The relationship between these two is highly debated across the NFL.

Fortunately, a data set was found online that includes the playing surface of each stadium, the body part that was injured, and the severity of the injury using binary threshold variables. The user 'JasonZivkovic' on Kaggle scraped the data from the NFL's archive of league plays [6]. The priority for this user was to find plays where significant lower-limb injuries happened. This data set contains 9 columns and 105 observations, which is admittedly a small sample size. Table 2 below shows the data dictionary for the raw data set found on Kaggle.

Variable	**Description and Units**
PlayerKey	The anonymous ID of the athlete who was injured
GameID	The ID of the game that the player was injured in
PlayKey	The ID of the play that the player was injured in
BodyPart	Body part injured during the play (knee, ankle, or foot)
Surface	Playing surface that the injury occurred on (synthetic or natural surface)
DM_M1	Binary variable representing if the player missed at least 1 day (1 if the player did, 0 if the player missed 0 days)
DM_M7	Binary variable representing if the player missed at least 7 days (1 if the player did, 0 if not)
DM_M28	Binary variable representing if the player missed at least 28 days (1 if the player did, 0 if not)
DM_M42	Binary variable representing if the player missed at least 42 days (1 if the player did, 0 if not)

Table 2: Data dictionary for the raw NFL external factors data set. Provided by Jason Zivkovic on Kaggle.

Internal Factors for NBA Players

The weight and height of each NBA player in the league's history was found using the database on the NBA's official website [4]. This data set contains every player in league history. The website says that there are 5126 observations and 8 columns in the raw data set. Outside of the aforementioned internal factors, other variables include the player's name, team, number, and position. Table 3 below shows the original data dictionary of the NBA internal factors data set provided by the league itself.

Variable	Description and Units
Player	Name of the athlete
Team	The NBA team the athlete plays for or last played for, abbreviations used
Number	Number last worn by player
Position	Primary basketball position of player (C, G, F)
Height	Height of player, feet and inches form (ex: 6-10)
Weight	Weight of player in pounds

Table 3: Data dictionary for the raw NBA internal factors data set. Provided by NBA.

NBA Injury Outcomes

Two different data sources were found for injury outcomes in the National Basketball Association (NBA). The first data set, 'NBA Injuries from 2010-2020', comes from user Randall Hopkins on Kaggle [1]. The data he found was scraped from Pro Sports Transactions, a website that contains every major transaction (including injuries) that a professional sports team made. The raw data contained 27105 observations and 5 columns before data cleaning. Table 4 below shows the complete data dictionary of the raw NBA injury data set from 2010 to 2020.

Variable	Description and Units
Date	Date of the reported injury, written in yyyy-mm-dd
Team	The NBA team the athlete plays for
Relinquished	Name of injured player
Notes	Qualitative description of injury

Table 4: Data dictionary for the raw 2010-2020 NBA injuries data set from Randall Hopkins on Kaggle.

The second data set, an NBA injury report from October 2021 to June 2024, comes from Vaughn Hajra on Stat Surge [2]. Notably, Hajra has injury reports for the entire 2024-2025 season as well, but this data set comes from the historical data at the bottom of the website. The raw data set contains 35,522 observations and 6 columns. Table 5 below shows the complete data dictionary of the raw NBA injury data set from 2021 to 2024.

Variable	Description and Units
Player	Injured player's name: last name, first name
Status	The game status of the player (e.g. out)
Reason	Description of injury
Team	NBA team that the athlete plays for
Game	Upcoming game for the player
Date	Date of the game, in mm/dd/yyyy format

Table 5: Data dictionary for the raw 2021-2024 NBA injuries data set from Vaughn Hajra on Stat Surge.

Data Cleaning and Transformation

For data cleaning and transformation, Microsoft Power Query was utilized within Excel for removing unnecessary columns, creating new columns for the analysis, merging tables, and more. Fortunately, as mentioned above in the data collection step, not all data sets needed to be cleaned and transformed.

Finalized English Premier League Data from 2016-2021

Table 6 below shows the finalized data dictionary for the English Premier League injury data set from 2016-2021. For the data cleaning step, there were several null values for the variables that required data from previous seasons. This includes minutes played in the previous season, matches played in the previous season, and average minutes played per game in the previous seasons. It is believed that when the data was scraped, the algorithm did properly handle players who joined the English Premier League midway through the time range of 2016 to 2021. These cells were filled with zeroes as the players did not participate in any English Premier League matches or receive any minutes prior to joining the league.

Most variables were removed, but a few were transformed from categorical variables into binary variables. Using the goalkeeper as a reference group, binary variables were created for whether or not a player in the study is a forward, midfielder, or defender. These ‘dummy’ variables allow for the categorical variable of position to be interpreted numerically. Internal factors kept were age and body mass index (BMI). The external factors kept were minutes played in the season, days missed due to injury in the previous season, and the aforementioned binary variables for position. The outcome variable of interest kept was the number of days the player missed due to injury.

Variable	**Description and Units**
season_days_injured	Number of days missed in the season due to injury
season_minutes_played	Number of minutes player received on the field for the current season
age	Player age in relevant season
BMI	Body mass index of player
season_days_injured_prev_season	Number of days missed due to injury in the previous season
isForward	Binary variable representing if the player is a forward or not: 1 if a forward, 0 if not
isMidfielder	Binary variable representing if the player is a midfielder or not: 1 if a midfielder, 0 if not
isDefender	Binary variable representing if the player is a defender or not: 1 if a defender, 0 if not

Table 6: Data dictionary for the cleaned and transformed 2016-2021 EPL injury data.

Finalized NCAA Injury Outcomes Data

While the data did not need to be cleaned, the data set with the NCAA injury outcomes needed some transformations for the categorical variables that are useful for interpretation. Table 7 below shows the finalized data dictionary for the sample of training sessions across the NCAA. Internal factors such as heart rate in beats per minute, respiratory rate in breaths per minute, skin temperature in degrees Celsius, cumulative fatigue index, and blood oxygen level percentage were kept. External factors such as session duration in minutes and force of impact of the session in Newtons were also kept. Using the sport type, a binary variable was created to measure if the sport played had heavy contact or not. If the sport played during the training session was football, then this variable took on a value of 1. If it was any other sport, the variable took on a value of 0. Another binary variable was created to measure if the sport played had mild contact of any form. If the sport played was soccer or basketball (two sports that have mild contact during training), then the variable took on a 1. If any other sport was played, this variable for mild contact took on a value of 0. Finally, a third binary variable was created from the intensity of the activity that was performed. If the activity type was running or sprinting, then the binary intensity variable took on a value of 1. For any other activity, the variable took on a value of 0.

Variable	Description and Units
Heart_Rate_BPM	Heart rate in beats per minute
Respiratory_Rate_BPM	Respiratory rate in breaths per minute
Skin_Temperature_C	Athlete's skin temperature in degrees Celsius
Blood_Oxygen_Level_Percent	Percentage of blood oxygen saturation
Impact_Force_Newtons	Force of impact during the training session, measured in Newtons
Cumulative_Fatigue_Index	A calculated index representing the athlete's fatigue level
Duration_Minutes	Duration of the training session in minutes
Injury_Risk_Score	A calculated score representing the risk of injury
Injury_Occurred	Binary variable indicating whether an injury during the session (1 for injury, 0 for no injury)
isHeavyContact	Binary variable indicating whether or not the sport contained heavy contact (1 for football, 0 for any other)
isMildContact	Binary variable indicating whether or not the sport contained mild contact (1 for soccer or basketball, 0 for any other)
isIntense	Binary variable indicating whether or not the activity performed was intense (1 for running or sprinting, 0 for any other)

Table 7: Data dictionary for the transformed NCAA injuries data set.

Finalized NFL Playing Surface and Injury Severity Data Set

Table 8 below shows the finalized data dictionary for the NFL data set containing the playing surface, body part, and injury severity. These threshold variables for the number of days missed were renamed, as the old variable names were not very clear. The categorical variables for the injured body part and playing surface were recreated using binary variables. If the playing surface was synthetic, then this variable takes on a value of 1. If the playing surface was natural, then it takes on a value of 0. If the injured body part was the knee, then the variable takes on a value of 1. If the injured body part was the foot or ankle, then the variable takes on a value of 0.

Variable	Description and Units

DaysMissed1	Binary variable representing if the player missed at least 1 day (1 if the player did, 0 if the player missed 0 days)
DaysMissed7	Binary variable representing if the player missed at least 7 days (1 if the player did, 0 if not)
DaysMissed28	Binary variable representing if the player missed at least 28 days (1 if the player did, 0 if not)
DaysMissed42	Binary variable representing if the player missed at least 42 days (1 if the player did, 0 if not)
isSynthetic	Binary variable representing if the playing surface of the injury was synthetic (1 if synthetic, 0 if natural)
isKnee	Binary variable representing if the body part injured was the knee (1 if knee injury, 0 if injury to ankle or foot)

Table 8: Data dictionary for the finalized NFL external factors data set.

Statistical and Data Analysis

The finalized data sets were analyzed in an attempt to determine which internal and external factors contributed most to the likelihood of an injury and the extent of an injury. As with most other research projects, the statistical significance was considered as an evaluation technique for each model and predictor variables in each model. A standard confidence level of 95% was used for this threshold. Due to the topic of this study being injuries and the health of athletes, a model or predictor variable may still be considered clinically significant in this project even if it is not statistically significant.

EPL Predictive Model for Injury Days Missed

To measure the effects of external and internal factors on injuries in professional soccer, a traditional multiple linear regression model was created using R programming. The dependent variable used was the number of days missed in the season due to injury. This variable obviously shows if a player was injured or not, but the number of days missed also allows for the true extent of the injury to be considered. Predictor variables include minutes played in the previous season, age, body mass index (BMI), days injured in the previous season, and the three binary variables representing a player’s position. The equation below shows the multiple linear regression model used for this data.

$$\begin{aligned} SeasonDaysInjured_{Player,Year} \\ &= \beta_0 + \beta_1 SeasonMinutesPlayed_{Player,Year} + \beta_2 Age_{Player,Year} \\ &+ \beta_3 BodyMassIndex_{Player,Year} + \beta_4 DaysInjuredPreviousSeason_{Player,Year} \\ &+ \beta_5 isForward_{Player,Year} + \beta_6 isMidfielder_{Player,Year} + \beta_7 isDefender_{Player,Year} \\ &+ \varepsilon_{Player,Year} \end{aligned}$$

Binary Logistic Regression Model for NCAA Injuries

The response variable of interest for the NCAA injury data set was whether or not the player suffered an injury during the training session that was conducted. This dependent variable had only two outcomes (1 for an injury, 0 for no injury) and therefore was not fully continuous. The standard multiple linear regression model was not appropriate for this problem.

Instead of a multiple linear regression model, a binary logistic regression model was implemented in Statistical Analysis Software (SAS). The equation below shows the logit transformation of the logistic regression function model for the probability of an NCAA athlete from the sample suffering an injury during the training session. The probability of interest is the likelihood of the student-athlete suffering from an injury during the training session that they are participating in. Relevant predictor variables for internal factors include heart rate in beats per minute, respiratory rate in breaths per minute, skin temperature in degrees Celsius, and blood oxygen level percent. External factors include session duration in minutes, impact level of the training session in Newtons, and the contact and intensity levels of the training session.

$$\begin{aligned} \ln\left[\frac{P(injury)_{Athlete,Session}}{1 - P(injury)_{Athlete,Session}}\right] = \\ &= \beta_0 + \beta_1 HeartRate_{Athlete,Session} + \beta_2 RespiratoryRate_{Athlete,Session} \\ &+ \beta_3 SkinTemperature_{Athlete,Session} + \beta_4 BloodOxygenLevel_{Athlete,Session} \\ &+ \beta_5 ImpactForce_{Athlete,Session} + \beta_6 CumulativeFatigueIndex_{Athlete,Session} \\ &+ \beta_7 SessionDuration_{Athlete,Session} + \beta_8 isHeavyContact_{Athlete,Session} \\ &+ \beta_9 isMildContact_{Athlete,Session} + \beta_{10} isIntense_{Athlete,Session} \end{aligned}$$

The odds were converted back to probability using the equation below. This easily allowed for interpretation, as each individual metric then had a predictive likelihood of the athlete suffering from an injury during the training session.

$$P(injury)_{Athlete,Session} = \frac{Odds_{Athlete,Session}}{1 - Odds_{Athlete,Session}}$$

Before the binary logistic regression was conducted, the predictor variables in the model were checked for multicollinearity. It was expected that a few of the variables for internal factors would be linearly related to each other before the regression was even run. For this project, a condition number above 1000 was used as the threshold for multicollinearity. If the variance inflation factor was greater than 10, then multicollinearity was considered severe among predictors. This step, run just before the regression, was also completed using Statistical Analysis Software (SAS).

NFL Playing Surface and Injury Severity Binary Logistic Regression Model

In a similar method to the NCAA injury model, a binary logistic regression model was used to determine the effects of playing surface and injured body part on the severity of an NFL injury from the perspective of days missed. The equation below shows the logit transformation of the logistic regression model. For the purposes of this regression model, the dependent variable will be the binary variable representing whether or not the player missed at least 42 days from their injury. This means that the probability of interest is the likelihood of a player missing at least 42 days from an injury. This is the most extreme injury outcome from the finalized data set, which may be of interest to the NFL and its teams as they look to reduce the number of severe injuries. As with the previous model, the odds ratio will be considered to allow for interpretations of the two predictor variables.

$$\ln\left[\frac{P(Min42DaysMissed)_{Athlete,Injury}}{1-P(Min42DaysMissed)_{Athlete,Injury}}\right]=$$

$$=\beta_0+\beta_1 isSynthetic_{Athlete,Injury}+\beta_2 isKnee_{Athlete,Injury}$$

Results and Discussion

EPL Multiple Linear Regression Results

Table 9 below summarizes the multiple linear regression results for the English Premier League model with the number of injury days missed in the season as the dependent variable. Internal factors in this regression include age and body mass index (BMI). External factors include minutes played in the season, days injured in the previous season, and the player's position. As expected from the multifactorial nature of injury, the model achieved an adjusted R-squared of just 11.39%, so the model is not highly predictive of the number of days that a player misses due to injury. Despite the model being weak, several variables could be considered significant. Minutes played in the season and days injured in the previous season are statistically significant in the model. Age, along with the binary variables representing the different positions, may also be considered clinically significant by those with medical expertise.

Variable	Coefficient	P-Value
Intercept	88.569744	0.01616
Season Minutes Played	-0.027323	<0.0001
Age	0.969747	0.10979
Body Mass Index (BMI)	-0.496373	0.75396
Days Injured in Previous Season	0.086429	0.00467
isForward	10.634257	0.26964
isMidfielder	12.337696	0.16918
isDefender	17.0408	0.05643

Table 9: Multiple linear regression results for the 2016-2021 EPL injury sample, with the dependent variable being the number of injury days missed in the season.

Using the estimated coefficients from the table, the equation below shows the predictive model with injury days missed in a season as the dependent variable. These estimates can be used to make interpretations about the effects of internal and external factors on days missed due to injury during an English Premier League game from 2016 to 2021. Surprisingly, an additional minute played in a season is predicted to decrease the number of days missed by 0.027323. If a player's age increases by one year, they are predicted to miss almost a full additional day due to injury. An increase by one in body mass index is associated with a decrease in days missed by 0.496373. An additional day missed in the previous English Premier League season is predicted to increase days missed by 0.086429. For the binary variables representing positions, the reference group consists of all goalkeepers. Being a forward, as opposed to being a goalie, is associated with 10.634257 additional days missed. Being a midfielder, as opposed to a goalie, is associated with 12.337696 additional days missed during the season. Being a defender, as opposed to being a goalie, is associated with 17.0408 additional days missed. These estimated positional coefficients suggest that defenders miss the most time, followed by midfielders, forwards, and then goalkeepers.

$$\begin{aligned}
&SeasonDaysInjured_{Player,Year}\\
&\qquad = 88.568744 - 0.027323 SeasonMinutesPlayed_{Player,Year}\\
&\qquad + 0.969747 Age_{Player,Year} - 0.496373 BodyMassIndex_{Player,Year}\\
&\qquad + 0.086429 DaysInjuredPreviousSeason_{Player,Year}\\
&\qquad + 10.634257 isForward_{Player,Year} + 12.337696 isMidfielder_{Player,Year}\\
&\qquad + 17.0408 isDefender_{Player,Year}
\end{aligned}$$

NCAA Injuries Binary Logistic Regression Model Results

Appendix B shows the full SAS results for the multicollinearity checks among the predictor variables. The variance inflation factors quickly reached extremely high values, which suggest that there are significant linear dependencies amongst the predictors in the model. In particular, strong collinearity existed between two pairs of internal factors. First, blood oxygen level percentage and skin temperature in degrees Celsius were very strongly dependent on each other. Second, heart rate in beats per minute and respiratory rate in breaths per minute were also strongly dependent on each other. Because these dependencies were so strong, the determined solution was to remove one variable from each pair. In this case, respiratory rate in breaths per minute and skin temperature in degrees Celsius were removed from the binary logistic regression model. For interpretation, respiratory rate can be considered with heart rate and skin temperature can be considered with blood oxygen level percent.

Table 10 and the equation below show the results of the NCAA binary logistic regression model performed in Statistical Analysis Software (SAS). Statistically significant factors include blood oxygen level percentage, force impact of the training session in Newtons, and cumulative fatigue index during the session. Surprisingly, the duration of the training session was not statistically significant. The binary variable for heavy contact, which represents whether or not the sport played was football, is right on the threshold for the significance level of 95%. To some, this may suggest that the binary variable for heavy contact is clinically significant.

Variable	Coefficient	P-Value
Intercept	13.315	<0.0001
Heart Rate (BPM)	-0.00279	0.3334
Blood Oxygen Percent Level	0.1196	<0.0001
Impact of Force (N)	-0.00597	0.0001
Cumulative Fatigue Index	-1.9031	<0.0001
Duration of Session (min)	-0.00162	0.6241
isHeavyContact	0.3394	0.0606

isMildContact	-0.0294	0.8452
isIntense	-0.124	0.365

Table 10: Binary logistic regression results for the NCAA injury sample, with the dependent variable being whether or not the player suffered an injury during the training session.

$$\ln\left[\frac{P(injury)_{Athlete,Session}}{1-P(injury)_{Athlete,Session}}\right] =$$

$$= 13.315 - 0.00279 HeartRate_{Athlete,Session} + 0.1196 BloodOxygenLevel_{Athlete,Session} - 0.00597 ImpactForce_{Athlete,Session} - 1.9031 CumulativeFatigueIndex_{Athlete,Session} - 0.00162 SessionDuration_{Athlete,Session} + 0.3394 isHeavyContact_{Athlete,Session} - 0.0294 isMildContact_{Athlete,Session} - 0.124 isIntense_{Athlete,Session}$$

Table 11 below shows the odds ratio estimates for each of the predictor variables included in the binary logistic regression model. If a predictor variable had no impact on the likelihood of injury, it would have a point estimate of 1. Therefore, these estimates can be used to quantify the impact of the internal and external factors on the likelihood of an NCAA student-athlete suffering an injury. An increase in heart rate by one beat per minute is predicted to decrease the probability of an injury occurring by 0.3%. A 1% increase in a student-athlete's blood oxygen level during the training session is associated with a 11.3% decrease in the probability of an injury. An increase by one Newton of force impact during the training session is predicted to decrease the likelihood of injury by 0.6%. By a large margin, the variable with the largest impact on the probability of an injury occurring was cumulative fatigue index. Bizarrely, an increase by 1 in cumulative fatigue index is associated with an 85.1% decrease in the probability of injury. Another shock was that the duration of the session in minutes did not make much difference in the likelihood of injury. An increase by one minute in training session length is associated with a 0.2% decrease in the likelihood of an injury occurring. The only variable to drastically increase the likelihood of injury was the binary variable for heavy contact. Playing a sport with heavy contact such as football, as opposed to other sports, is predicted to increase the probability of an injury occurring by 40.4%. The same increased effect was not the case for the binary variable for mild contact. Playing a sport with mild contact such as soccer or basketball, as opposed to other sports, is associated with a 2.9% decrease in the likelihood of an injury occurring during the training session. Finally, performing an intense activity such as running or sprinting, as opposed to other activities, is predicted to decrease the likelihood of injury by 11.7%.

Variable	**Point Estimate**
Heart Rate (BPM)	0.997
Blood Oxygen Percent Level	0.887
Impact of Force (N)	0.994
Cumulative Fatigue Index	0.149
Duration of Session (min)	0.998
isHeavyContact	1.404
isMildContact	0.971
isIntense	0.883

Table 11: Point estimate results for the NCAA injury binary logistic regression model, with the dependent variable being whether or not the player suffered an injury during the training session.

NFL Playing Surface and Injury Severity Binary Logistic Regression Results

Table 12 and the equation below show the results of the NFL binary logistic regression model performed in Statistical Analysis Software (SAS). At the 5% level, neither the playing surface nor the location of the lower-limb injury had a statistically significant impact on the likelihood of missing at least 42 days due to injury. However, these factors could still hold clinical significance due to this being a health and safety setting.

Variable	Coefficient	P-Value
Intercept	0.7832	<0.0466
isSynthetic	-0.1642	0.7119
isKnee	0.3992	0.3682

Table 12: Binary logistic regression results for the NFL injury sample, with the dependent variable being whether or not the player missed at least 42 days due to injury.

$$\ln\left[\frac{P(Min42DaysMissed)_{Athlete,Injury}}{1-P(Min42DaysMissed)_{Athlete,Injury}}\right] = 0.7832 - 0.1642isSynthetic_{Athlete,Injury} + 0.3992isKnee_{Athlete,Injury}$$

Table 10 below shows the odds ratio estimates for both of the predictor variables included in the NFL binary logistic regression model. Once again, if a predictor variable had no impact on the likelihood of an injury resulting in at least 42 days missed, it would have a point estimate of 1. Suffering an injury on a field with a synthetic surface, as opposed to a natural surface, is associated with a 15.1% decrease in the likelihood of the injury resulting in at least 42 days missed. This is somewhat of a surprising finding, as many NFL fans and players believe that synthetic surfaces such as artificial turf are associated with severe lower-limb injuries. As opposed to a foot or an ankle injury, a knee injury is predicted to increase the probability the player misses at least 42 days by 49.1%.

Variable	Point Estimate
isSynthetic	0.849
isKnee	1.491

Table 10: Point estimate results for the NFL injury binary logistic regression model, with the dependent variable being whether or not the player missed at least 42 days due to injury.

Conclusion and Future Research

Many conclusions can be drawn from the statistical analyses conducted within this study. For now, findings are best to be interpreted on a sport-by-sport basis. For professional soccer players such as the athletes in the English Premier League from 2016-2021, the number of minutes played on the season and the number of days missed in the previous season were statistically significant in the multiple linear regression model with the dependent variable being the number of days missed on the current season due to injury. Other variables such as the athlete's position and age may be considered clinically significant, depending on one's interpretation. Defenders are predicted to miss the most days due to injuries, followed by midfielders, forwards, and then goalkeepers.

For college athletes who play a sport in the NCAA, there are several factors that may play a part in the athlete suffering from an injury during a training session. The most important factor is whether or not the athlete plays a sport (such as football) that consists of heavy contact. Football players are at a much higher risk of getting injured during a training session than athletes in other sports. Other factors that play a part in suffering from an injury include the athlete's blood oxygen level percent, cumulative fatigue index, and the total impact of forces during the entire training session.

The analysis on the NFL injuries revealed some surprising results. Most fans and players think very poorly of synthetic playing surfaces such as turf, especially in the context of severe injuries. Despite this, the analysis revealed that suffering an injury on a synthetic surface is associated with a decrease in the likelihood of the injury requiring an absence of at least 42 days. A knee injury, as opposed to an ankle or foot injury, is predicted to greatly increase the probability of the injury requiring an absence of at least 42 days.

As expected, there were several internal and external factors that were considered significant in contributing to injuries, including severe ones. Despite this confirmation of the hypothesis, the results varied across each of the different sports that were studied. This suggests that the true conclusion about the factors that contribute to injuries in sports is dependent on the sport or activity that is being played.

Future research should be conducted in order to find the full population of injuries across all of the sports leagues that were analyzed in this study. These analyses used just small samples from the full population, so finding the full extent of the data would allow for the best results. If this data can be found, then the estimates could become more accurate, and the model would hold higher relevance in predicting injuries and their severity.

Limitations

1. Finding data for this project was difficult for a few reasons. First, many leagues are very private with their injury data, as they do not want to publicly make available data that describes how dangerous their league is. Second, many data sets were behind paywalls or restrictions that could not be bypassed. This made it difficult to find data for all leagues. The goal was to find data for more leagues, such as Major League Baseball (MLB), the National Hockey League (NHL), and the National Association of Stock Car Auto Racing (NASCAR).
2. Most of the data sets that were found tended to be samples of the full population of injuries in the given league. Interpreting the findings of these analyses requires an assumption that these samples are representative of the full population of injuries in the given sport.
3. For the NBA data set, the data transformation step did not work properly in Microsoft Excel and Microsoft Power Query. The goal was to merge the internal factors provided by the NBA with the two data sets that contained injury outcomes. This prevented the analysis of NBA injuries for this study, but future time and effort will go into this.

Acknowledgements

I would like to give special thanks to everyone who aided me at any point during the process of completing this research study. First and foremost, I would like to thank Professor Bryan Crissinger, a Statistics professor in the Department of Applied Economics and Statistics within the University of

Delaware College of Agriculture and Natural Resources. Professor Crissinger served as my faculty advisor for this project, which was submitted for STAT468: Senior Research Project, the capstone requirement for the undergraduate Statistics degree at the University of Delaware.

I would also like to thank Professor Patrick DeFeo, another professor from the University of Delaware College of Agriculture and Natural Resources. Professor DeFeo instructed STAT611: Regression Analysis in the spring of 2025, a course that was very beneficial for the regression analysis techniques used in this project. In particular, the coursework on binary logistic regressions and multicollinearity among predictors proved to be of great assistance for this project.

Finally, I would like to thank Professor Jack Davis, the Associate Director of Business Intelligence, Analytics, and Strategy at the University of Delaware Athletic Department. Professor Davis instructed SPAX402: Predictive Analysis with Athletics Data during the fall of 2025. This course provided me with experience with Microsoft Power Query, a key concept used in this project.

Sources

Literature Review Works Cited

Bahr, R, and Holme, I. “Risk Factors for Sports Injuries -- a Methodological Approach.” *British Journal of Sports Medicine*, vol. 37, no. 5, 1 Oct. 2003, pp. 384–392, bjsm.bmj.com/content/37/5/384, https://doi.org/10.1136/bjsm.37.5.384. Accessed 27 Sept. 2025.

Taimela, Simo, et al. “Intrinsic Risk Factors and Athletic Injuries.” *Sports Medicine*, vol. 9, no. 4, Apr. 1990, pp. 205–215, link.springer.com/article/10.2165/00007256-199009040-00002, https://doi.org/10.2165/00007256-199009040-00002. Accessed 27 Sept. 2025.

Wentao, Zhao. “Analysis of the Causes of Sports Injuries in Sports Training.” *Frontiers in Sport Research*, vol. 6, no. 1, 2024, francis-press.com/uploads/papers/GABe79pq5P7HyzaqpGLsIxYGW4JZwvH1lJHfA1eT.pdf, https://doi.org/10.25236/fsr.2024.060106. Accessed 27 Sept. 2025.

Data Set Links

1. https://www.kaggle.com/datasets/ghopkins/nba-injuries-2010-2018?select=injuries_2010-2020.csv
2. https://statsurge.substack.com/p/downloadable-nba-injury-datasets
3. https://www.kaggle.com/datasets/ziya07/college-sports-injury-detection
4. https://www.nba.com/players
5. https://www.kaggle.com/datasets/kolambekalpesh/football-player-injury-data?resource=download
6. https://www.kaggle.com/code/jaseziv83/an-analysis-of-nfl-injuries

Appendices

Appendix A: Data Dictionary for Raw Data Set of 2016-2021 English Premier League Injuries

Variable	**Description and Units**
p_id2	Name of player
start_year	Year of observation
season_days_injured	Number of days missed in the season due to injury
total_days_injured	Total days missed across all six seasons
season_minutes_played	Number of minutes player received on the field for the current season
season_games_played	Number of matches player participated in for the current season
season_matches_in_squad	Number of matches the player's team had in current season
total_minutes_played	The total number of minutes the player received across all six seasons
total_games_played	Total number of matches the player participated in across all six seasons
dob	Date of birth of player, shown in m/dd/yyyy format
height_cm	Player's height in centimeters from FIFA 16-21
weight_kg	Player's weight in kilograms averaged across FIFA 16-21
nationality	Player's country of origin

work_rate	Categorical variable from FIFA 16-21 grading how hard the player works during matches; high, medium, or low
work_rate_numeric	Player’s work-rate mode across FIFA 16-21 ranging from 2-4 in 0.5 intervals
pace	Player’s ‘pace’ rating averaged across FIFA 16-21
physic	Player’s ‘physical’ rating averaged across FIFA 16-21
fifa_rating	Overall grade assigned to player by FIFA 16-21; averaged across every video game player appeared in
position	Player’s soccer position: Defender, Forward, Midfielder, or Goalkeeper
position_numeric	‘Goalkeeper’: 0, ‘Defender’: 1, ‘Forward’: 2, ‘Midfielder’: 3
age	Player age in relevant season
cumulative_minutes_played	Cumulative minutes played by player in all previous seasons
cumulative_matches_played	Cumulative number of matches played by player in all previous seasons
minutes_per_game_prev_seasons	Average minutes player per game in all previous seasons
avg_days_injured_prev_seasons	Average calendar days injured in all previous seasons
avg_games_per_season_prev_seasons	Average number of games played participated in during previous seasons
bmi	Body mass index of player
significant_injury_prev_season	Whether player had significant injury in previous season; 1 for injury and 0 for no injury
cumulative_days_injured	Cumulative calendar days player was injured prior to current season
target_major_injury	Whether player had major injury in current season; True = yes, False = no
season_days_injured_prev_season	Number of days missed due to injury in the previous season

Table A.1: Data dictionary for the raw 2016-2021 EPL injury data from ‘Kalpesh Kolambe’ on Kaggle.