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

Progress Report

Professor Bryan Crissinger

October 17, 2025

Literature Review Sources

Before any data was searched for, research was conducted in order to find correct biomedical information on injury prevention and risk. Databases such as Academic OneFile, Google Scholar, SPORTDiscus, and DELCAT Online Catalog were searched for these research articles and academic pieces.

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).

Data Sources

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) and select NCAA sports. The process of data collection, cleaning, and transformation is ongoing. Unless otherwise denoted, each data set was downloaded or pulled into Microsoft Excel for data storage. 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.

Internal Factors for NBA Players

The age, weight, and height of each NBA player in the league's history was found using the database on the NBA's official website [4]. As far as I know, 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. In data cleaning, I deleted the 'Last Attended' and 'Country' columns, which contain information about the last attended academic institution and country of origin for the player. This left the finalized data set with just 6 columns. Table 1 below shows the complete 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 1: Data dictionary for the 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. Unfortunately, I was not able to scrape the data from this website myself due to an access issue, so I had to rely on Hopkins' work. The raw data contained 27105 observations and 5 columns before data cleaning. The 'acquisitions' column was removed, as it was blank for almost all observations. For the 'Relinquished' column, any values reading 'null' were removed from the data set. A custom column called 'Serious_IL' was created in Power Query in order to create a binary classification metric with two outcomes. If a player's injury resulted in an IL (injured list) designation from their team, required surgery, or included a tear or fracture, then the variable takes on a value of 1. Otherwise, the injury is not very serious, so it takes on a value of 0. The final data set has 17560 observations and 5 columns. Table 2 below shows the complete data dictionary of the cleaned 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
Serious_IL	Binary variable that takes on a 1 if the injury was serious, 0 if not

Table 2: Data dictionary for the cleaned 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. In the data cleaning step, the 'Status' column was filtered to just find players who were officially out of the game, as opposed to questionable, probable, available, or doubtful. The 'Reason' column was filtered to include players who had an injury/illness, as opposed to players missing games due to personal reasons, suspensions, etc. A new column called 'Body Part' was created from the 'Reason' column. Using Power Query's autofill feature, the injured body parts (ankle, foot, hamstring, etc.) of the players were denoted. A future data cleaning or data analysis step could be included to prevent an injury from being counted for as many games that the player missed, as opposed to counting it just as a single injury. As of right now, the cleaned data set has 15532 observations and 7 columns. Table 3 below shows the complete data dictionary of the cleaned 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
Body Part	Injured body part for the player (ankle, foot, calf, etc.)

Table 3: Data dictionary for the cleaned 2021-2024 NBA injuries data set from Vaughn Hajra on Stat Surge.

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 100 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. Table 4 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 4: Data dictionary for the cleaned NCAA injuries data set from 'Ziya' on Kaggle.

Analysis Techniques

Multicollinearity

As mentioned in the sources above, many of the internal factors may be statistically significant in determining the likelihood of an injury occurring. A natural concern with these variables (especially height and weight) is that they may be linearly related to each other, even before the dependent variable is considered. This is a statistics and math concept known as multicollinearity. These linear dependencies in the predictor variables would impede the ability to estimate the change in the dependent variable with respect to an independent variable while holding the other predictor variables constant. After all, if two variables are related, then controlling for one in real life then acts as a control for the other.

To make these multicollinearity checks, SAS Academic OnDemand will be used to find the variance inflation factors for each coefficient. These factors, denoted as VIFs, are the amount of increase in the variance of a coefficient above the ideal orthogonal case. For a general rule of thumb for this project, a VIF above 10 will be considered an indication of multicollinearity.

Logistic Regression with Multiple Predictor Variables

As mentioned above, it is recommended in a project like this one to create a logistic regression model with multiple internal and external factors included as predictor variables. The equation below shows what this type of regression would look like from a mathematical perspective.

$$Injury_i = \alpha + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Height_i + \beta_4 PreviousInjury_i + \epsilon_i,$$

The dependent variable here is the injury outcome classification. This variable acts as a binary dependent variable: the player either suffered an injury or did not. This dependent variable can be adjusted as well. For example, the two outcomes can be switched to “serious injury” and “nonserious injury or no injury”. The predictor variables for this type of regression would be the internal or external factors that could possibly contribute to the likelihood of injury (or serious injury, dependent on the dependent variable). The research study above recommended including as many predictor variables as possible due to the multifactorial nature of injury.

To conduct this logistic regression, the programming language R will be used in Visual Studio Code. Notably, a library module or package may need to be downloaded on my version of R. Additionally, the interpretation of this logistic regression is different than that of a traditional regression. The estimated coefficients for each predictor variable represent a percentage point change in the probability that the dependent variable (injury) occurs.

Updated Timeline

Deliverables

` The timeline of deliverables for this research project course is in the process of being met. The formal proposal of the research project was submitted on Friday, September 19th, 2025. This progress report on the project, which includes more details about the data sources and a discussion of the analysis methods, will be submitted on Friday, October 17th, 2025. A full rough draft with analytical results, a summary, and a conclusion will be submitted on Friday, November 21st, 2025. The full presentation to classmates and faculty will be delivered in-person on Friday, December 12th, 2025. The final paper will be submitted on Monday, December 15th, 2025.

Individualized Timeline

As mentioned in the formal proposal, a second, individualized timeline is being used for the smaller tasks of this research project. The formal proposal's goal was for me to use the rest of September to conduct the literature review and background section of the project. The first week of October was to be used to find the data sources and data sets that will be used for the analysis section. Data cleaning, data transformation, identification of software, and clarification of analysis techniques were to be completed in the second week of October.

As it stands, I am behind in certain areas while also being ahead in others. It is clear that more data needs to be collected in terms of injury outcomes for the major leagues mentioned in the formal proposal. The research section can also use more information detailing what has been done in terms of specific biomedical and biomechanical changes within these levels of athletics. Although it should be easy, I still need to find the data relating to internal factors regarding each athlete that is identified in the data sets. I am further along with the rough draft than anticipated for this point in time. I underestimated how much of the progress report and formal proposal can actually be included in the rough draft itself. Along with this progress report, I have uploaded a "rough" rough draft that shows what I have so far at the time of submitting this progress report.

With some adjustments, the previous plans from the formal proposal can be achieved. The last few weeks of October will be used to conduct statistical analyses for each of the data sets already found for this research project. Additionally, more data will be searched for regarding the leagues mentioned in the formal proposal. The early part of November will then be spent on finishing the full rough draft, which will be submitted on Friday, November 21st, 2025. Any advice on the "rough" rough draft will be considered for this stage. The week of Thanksgiving break will be used to take the key points of the paper and create a slideshow presentation out of them. The weeks of December will be spent practicing the presentation and refining the paper into the final draft.

Links

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 Sources 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