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 Rough Draft

Professor Bryan Crissinger

October 17, 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 will look at data and research from multiple sports leagues at the two levels, including the National Football League (NFL), Major League Baseball (MLB), the National Basketball Association (NBA), National Collegiate Athletic Association (NCAA) football and basketball, Major League Soccer (MLS), and the National Association of Stock Car Auto Racing (NASCAR).

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.

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) 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.

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 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