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

Formal Proposal

Professor Bryan Crissinger

September 19, 2025

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.

Data Sources

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. This should allow for a detailed literature review and background section of the project.

Online publications from peer-reviewed academic journals will be utilized for this part of the project. To find these publications, online databases from the University of Delaware Library will be searched. The online library includes databases such as Academic OneFile, SPORTDiscus, and DELCAT Online Catalog. These databases will most likely not contain enough information about each of the major professional and collegiate sports mentioned above. As an alternative, academic search engines such as Google Scholar will be used to find articles that fill in the missing gaps of information. Additionally, independent research teams and medical committees appointed by these leagues have created publications discussing injury and health trends in their respective sports. Since these league-appointed committees have a clear conflict of interest, these publications will be considered with hesitation.

Data Sets

The completion of the literature review and background section should ease the search for data sets online. There are a few different categories that the data sets can be organized into. First, data will need to be collected in regard to the internal factors of athletes within a given sport. This should not be much of a problem, as information such as height, weight, and age are all frequently updated in team and league records.

The next type of data set relates to the external factors that potentially contribute to an athlete's likelihood of getting injured. This category includes characteristics such as playing surface, protective equipment, weather conditions, and more. For most of the mainstream sports, data on the playing conditions of each game are recorded. Data relating to the use of protective equipment may be limited at both professional and collegiate levels. If this is the case, the findings of the literature review and background sections will become even more important.

The third and most important type of data set needed is one that contains injury outcomes for athletes within each of the aforementioned sports leagues. Fortunately, online data sources and databases such as Kaggle contain league-specific data sets for data science projects such as this one. Injury outcomes for mainstream leagues such as the NFL and NBA can be found in several places online. Additionally, many leagues maintain injury and incident databases that can be used to generate data sets. This will be especially beneficial for the leagues that may not be as mainstream as the others (MLS and NASCAR).

The fourth and final type of data set contains data relating to biomechanical and biomedical measurements in professional and collegiate athletics. This data should be very easy to identify at the professional level of sports. For example, several databases (such as BaseballReference.com) track biomechanical data for Major League Baseball pitchers, who suffer the most injuries in professional

baseball. There might be a challenge to find this data at the collegiate level, which may limit the ability to answer the second pair of research questions.

Key Concepts Involved

Qualitative and Quantitative Data

There are several key sports and statistical concepts involved in this research project. To start, the data sets found will most likely include a mix of qualitative and quantitative data to be analyzed. Quantitative data includes numerical measures such as player age, height, weight, training workload, minutes played, or number of injuries sustained over a season. Qualitative data, on the other hand, includes categorical or descriptive information such as player position, type of injury, playing surface, or equipment type. To be able to build predictive models, these variables need to be converted into numeric measurements in some way.

Risk Factors

A second key concept is identifying the full list of risk factors that could be predictors of injury. These risk factors may be different based on the sport that is being played, which means the list of potential predictors of injury changes as well. To determine the full list of injury risk factors, the literature review section will be used to identify general internal and external factors that are relevant to sports as a whole. From this general list, risk factors will be identified as statistically significant or insignificant for each of the major sports individually.

Multicollinearity

Multicollinearity checks among predictors (especially internal factors) is another step towards building predictive models for these sports. It is anticipated that some internal factors (especially height and weight) will be linearly related to each other in the data sets. Before conducting any regression analyses, collinear variables in each data set will be addressed.

Correlation vs. Causation

For any project like this one that attempts to use statistical findings for future decision-making, it is important to identify the difference between correlation and causation. It is expected that certain internal and external factors will have a statistically significant relationship with injuries. Generally speaking, this report will be very reluctant to make any claim about injury causation in reference to these factors.

Analysis Techniques

Statistical Methods

The analysis techniques will be mostly dependent on the findings from the biomedical research section of the paper. Each injury, team, and sport may require certain statistical and data science techniques to allow for analytical results that can be used in a predictive manner.

When considering a list of external and internal injury risk factors, a statistical method that comes to mind is model selection. Because each risk factor essentially acts as a predictor variable for injury

outcomes, a model selection process such as stepwise selection can be used to find the best multiple linear regression model for a specific criterion. This statistical analysis process will remove any risk factors that are not relevant regressors when it comes to predicting injury.

For injury data sets where there are only two outcomes (injured or not injured), binary logistic regressions can be used to create predictive models. This form of logistic regression will make it possible to estimate the probability of injury given a set of risk factors, such as position, sport, height, weight, etc. Odds ratios can be calculated to measure the relative importance of each risk factor in producing a heightened injury risk factor.

A third type of regression that might be used is polynomial regression, where the relationship between risk factors and injury outcomes is not perfectly linear. A common example of this relationship in sports injuries is the correlation between athletic workload and the likelihood of an athlete getting injured. As workload metrics increase, it is understood that the probability of an athlete suffering an injury accelerates rapidly.

Software and Data Analysis Tools

For a large-scale research project such as this one, software and data analysis tools will be used for the statistical analysis section. To clean, organize, and store data, Microsoft Excel will be used. To conduct the necessary statistical analyses, programming languages such as Python, R, and Statistical Analysis Software (SAS) will be used. To visualize findings of this research project, Tableau will be used to create a digital dashboard.

Timeline for Completion

Deliverables

The completion and submission of the deliverables for this research project will follow the research project timeline posted on the course syllabus. This report, the formal proposal of the research project, will be submitted on Friday, September 19th, 2025. A progress report on the project, which will include more details about the data sources and a formal 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

Past the course's calendar dates for the graded deliverables, a second, individualized timeline will be used for the smaller tasks of this research project. These personal deadlines, which will be built from the timeline of deliverables, will be used to keep the workload consistent and manageable across the entire project timeline. Since the progress report requires full details of the data sources and analysis methods, these tasks will take priority over the next few weeks. The goal is to use the rest of September to conduct the literature review and background section of the project, which will include the necessary biomedical and biomechanical information that will be the catalyst for the methodology section. The first week of October will be used to find the rest of the data sources and data sets that will be used for the

analysis section. The goal is to clean the data and identify key variables of interest for each data set during the second week of October. The third week of October will be spent identifying the methods and software that will be used to analyze the data sets. This may take some time, as the methods will most likely differ between sports because of the literature review findings. The last few weeks of October will be used to conduct the statistical analyses for each of the data sets included in the research project. The first few weeks of November will then be spent on the creation of the full rough draft, which will be submitted on Friday, November 21st, 2025. 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.