Understanding Metrics and Statistics Aimed at Prediction and Prevention of UCL Injury Risk

Will Rettig

11/23/2025

Denison University

Table of Contents

Abstract....3

Understanding Metrics and Statistics Aimed at Prediction and Prevention of UCL Injury Risk

Introduction....4

Background....4

Literature Review....6

Data & Methods....12

Data:....12

Methods:....16

Results....22

Discussion....25

Conclusion....27

References....29

Appendix....32

Abstract

The prevalence of ulnar collateral ligament (UCL) injuries has increased substantially across all levels of baseball. However, there are limited methods and practices in place for predicting and preventing UCL injuries. Recent discoveries in baseball analytics provide new opportunities to pinpoint mechanical and statistical indicators of injury risk. This research incorporates both large-scale (macro) and smaller-scale (micro) data sets, including Baseball Savant pitching metrics, Sean Lahman's Baseball Database data on individual measurables, a database listing Tommy John Surgeries, and lower-body data collected from a NewtForce Mound. The study attempts to evaluate potential predictors of UCL injury by using correlational analysis, time-series visuals, interaction regression models, and a LASSO logistic regression analysis. The results show relatively weak, yet consistent associations between fastball velocity, body mass index (BMI), and innings pitched as indicators of a need for surgery. The LASSO model was modest in its predictive capabilities (AUC = 0.61), suggesting a multi-faceted approach to predicting UCL injury is most insightful. This research presents one of the first holistic integrations of biomechanical, performance, and statistical data for UCL injury analysis, aimed at predicting and preventing injuries before they happen.

Understanding Metrics and Statistics Aimed at Prediction and Prevention of UCL Injury Risk

Introduction

Background

The history of arm injuries in the game of baseball can best be classified as dynamic. Arm injuries have varied by location, severity, and prevalence as ideologies change within the game each year. Trends in today's game show a growing issue of UCL injuries. The ulnar collateral ligament, known as the UCL, is a sacred ligament in the game of baseball. Located on the inside of the elbow, it holds several tissues together that are important to the throwing motion. Most commonly, baseball players will tear or stretch this ligament due to overuse, or simply greater stress on the ligament (Ulnar Collateral Ligament (UCL) Injury, 2022). The increase in injury prevalence coexists with a new age, commonly referred to as the 'Statcast Era,' in Major League Baseball. Statcast was first introduced in 2015, with expanded statistics and metrics on batted and pitched ball data, including information collected on every single pitch and swing within a game (Major League Baseball, n.d.). As a result, with more numbers available, there is a greater ability to understand what must be done for a pitcher to throw harder. Pitchers hope to gain any advantage they can over a hitter, who hope to find any way to beat the pitcher. For hitters, the stronger they can be, the further they can hit the ball. The better their strength and the higher their launch angle, the more likely they are to hit home runs. Inversely, the way that a pitcher must combat this shift is to make the hitter miss their pitches more often. This is best done by two options, the first being a creation of more spin and being more 'nasty,' and the second being able, simply, to throw harder. The lengths taken to increase spin generation and pitching velocity have adverse effects on arm health. The body and the arm must move faster and exert more force, which is inherently adding more stress to the arm, specifically the UCL.

Further, pitchers are willing to make the sacrifice of putting their arm health at risk at the chance it leads to advantages over hitters.

In this paper, data will be explored regarding arm injuries with the purpose of exposing commonalities of deficiencies that may lead to injury prediction and, ultimately, injury prevention. The research question for this paper holds the definition of, "can arm injuries, specifically injuries to the UCL, be prevented or predicted through analysis of various data points and systems that expose indicators of weakness or potential injuries?" Though injuries in baseball are not limited to the elbow, this remains an ever-important question because baseball players of all ages are suffering UCL injuries (Petersen, 2025). These injuries have been known to consistently hinder an athlete from returning to their full potential, often cutting a player's career short or forcing early retirement. Over time, the development of the surgery process has become more successful. When the surgery was first being performed, the success rate was negligible. Now, the surgery and rehabilitation process has become so normalized that the injury has become nearly obsolete in terms of the athlete's ability to get back to their full potential. Nevertheless, an athlete loses 12-16 months, on average, of baseball progress with rehabilitation and recovery from surgery. This does not take into consideration the mental toll it takes on the injured player. Trends of commonalities between deficiencies would be crucial to reducing the prevalence of this major injury. It would be significant to find alternative methods and approaches to arm care, rehabilitation, and injury prevention, which would outline major implications for the game of baseball. Countless coaches and instructors offer methods and training strategies to 'bullet-proof the UCL,' yet very little science and numbers exist to support these routines. From the current data of this study, it is natural to hypothesize that there is some way to predict arm injuries or areas of weakness in the UCL. The aim of this study is to identify

these areas of weakness and expose them through analysis and interpretation of the methods and data at hand.

The study is not without its limitations. Because each athlete has a different genetic makeup, arsenal of pitches, and mechanical foundation and breakdown, the results will be difficult to generalize to an individual. This does not indicate an insufficient research question. Analysis can still be conducted effectively, though it is important for exposing the trends on a larger scale while considering individual effects. Generalizability will be discussed exclusively to the overall trends for the specific athletes selected by Alpha Baseball and the Major League Baseball pitchers in the Statcast Era (2015-present). Further, these results will not be sufficient to prove causality due to the discussed individual differences between players. There are no methods of which this process can enable causality. This analysis is aimed to see if overlap exists between different data and player sources for variables between platforms and whether these trends can be connected. The exploratory elements along with the foundational analysis of the connection are necessary to future development and research. This analysis is the first of its kind to establish the connection of these various data platforms, further emphasizing the importance of the results of this research.

Literature Review

Despite this analysis being the first of its kind, it remains important to note that significant research and analyses exist on the previously referenced history of arm injuries. This domain knowledge report attempts to provide an overview into the past and present relevant scholarly information on the rich history of arm injuries in baseball players. Much is known regarding the severity of these injuries, yet researchers are skeptical on the true factors that lead

into major injuries, such as tears (partial or full) of the ulnar collateral ligament, tears of the rotator cuff, etc. A commonly held theory is that the extreme wear and tear on the arm that causes an injury to the UCL is exacerbated by deficiencies in other areas. These potential deficiencies will be explored using the micro data obtained from Alpha Baseball. These deficiencies are, but not limited to, insufficient hip mobility, insufficient shoulder range of motion, insufficient internal and/or external rotation of the elbow, etc. Another point of emphasis is to increase weight and body mass, regardless of player height, to decrease arm stress. Each of these factors are commonly treated with equal importance, which furthers the importance of the research behind them to pinpoint and understand potential causes of significant arm injuries. However, the lack of a holistic approach that includes all potential factors into one major study further proves a deficiency in research.

Gaining sufficient knowledge of the relevant discussion in this field is important to gaining a true understanding of how the holistic approach can be implemented. The following domain knowledge studies will be listed in chronological order with emphasis on showing the growth and dynamicity of research in the field. (Brown et al., 1988) examined the range of motion of the upper extremities including the shoulder and elbow. The measurements were taken from MLB baseball players. This study was conducted utilizing measurement with an emphasis on degrees for internal and external rotation and time used paired with the degrees allowed for the calculation of force. There were both position players and pitchers utilized in this study to find the differences between the two groups. It is an important implementation having both position and pitching players included. A small section of the exploratory analysis of this study includes position players, yet there are fundamentally less arm metrics and statistics gathered from position players due to the lack of UCL injury prevalence in position players. Internal and

external elbow rotation deficiencies were not found to be significant as they pertain to UCL injuries, nor is data readily available on MLB pitchers and position players on these specific elbow metrics. (Bartlett et al., 1989) examined the force production of the upper extremities and its parts along with the correlation to the throwing speed of overhead athletes. This study seems inherently flawed, as it seems logical to conclude without any evidence that higher force would naturally lead to greater throwing velocities, as measured by a radar gun. Another less commonly considered flaw is the radar gun itself, as even in 2025 there is only one pitch speed measuring system known to be the most accurate. Depending on the angle, distance from the pitcher, and other factors, the speeds can vary by three to five miles per hour. Pitch speeds, then, are subjective depending on the radar gun utilized. Regardless of the flaws, this study adds the following notion to the domain knowledge; more force production does equal higher throwing velocities. Pitch speed metrics for the data of this study were collected with stadium Trackman, accessible by all 30 MLB teams, and the top fastball velocities from the Alpha Baseball study were collected using the same Trackman system. These data points were examined in the exploratory analysis phase, testing if higher force could result in higher wear and tear on the upper extremities, which is enticing to look at the relationship between higher velocities and injury prevalence in high performance athletes. (Wilk et al., 2002) examined the rehabilitation techniques for overhead throwing athletes at the time. Included are significant analyses, paired with an overview of potential rehabilitation techniques. There are also areas of potential deficiencies that could lead to a need for rehabilitation. This scholarly discussion is important, as it lists and identifies all the factors that should be considered for a holistic analysis. It provides a sound basis for a holistic approach to potential research and further emphasizes the need to examine factors on a macro-level. (Scher et al., 2010) emphasized the importance for

examination of upper half and lower half measurements. It attempts to examine relationships between hip range of motion and shoulder range of motion and the potential impacts of deficits on injury history. This is an important discussion as it studies the relationship of both upper half and lower half measurements, and how the relationship can impact future injuries. Such an inclusion of factors supports having both upper and lower half data in a holistic analysis. (Garrison et al., 2012) examined only the shoulder range of motion (excluded hip range of motion like the previous study) and the impact it has on baseball players with injury to the ulnar collateral ligament. Simply, this study showed that range of motion deficits in the shoulder were found to be associated with a tear (grade of tear not explicitly specified) in the ulnar collateral ligament in the throwing elbow of the examined athlete. This distinction and finding holds important because it emphasizes a need to include more factors and metrics impacting the arm, as opposed to the elbow itself. (Garrison et al., 2013) examines the imbalance of the lower extremities as they pertain to UCL health. This furthers the importance of the NewtForce mound data, which is readily available within the micro data of this analysis. It provides sound research for the imbalances of the lower half of the body and the prevalence of UCL injuries in pitchers. (Conte et al., 2015) examines the prevalence of UCL reconstruction in professional baseball players at the Major and Minor league levels. While a much more simplistic study than the rest, it is important to note for background knowledge. 25% of Major League pitchers and 15% of minor league pitchers have had the surgery. It is important that we understand the prevalence of this injury to understand the severity of the pandemic of arm injuries. This will be important data to note for the changes over time, to see how the statistics of prevalence have changed from 2015 to the currently available data in 2025. (Peters et al., 2018) is a more subjective study as it examines the success of return to throwing in baseball following UCL reconstruction. More

evidence would be necessary for a greater study and to give more options pertaining to generalizability. Following UCL injuries, returning to throwing is a long, difficult rehabilitation process. The return to the previous level is less prevalent in professional sports as compared to high school and collegiate athletes. This shows that age likely plays a much bigger role than originally anticipated and can be included as a potential variable of interest for analysis. This furthers the importance to include workload (measured in innings pitched) to make more of a sound, holistic analysis. (Jang et al., 2019) examines the management of UCL reconstruction processes for baseball players returning to throwing at the time. The similarity of the previous study, (Peters et al., 2018), is an important note, as this study examines the return to throwing rates following reconstructive surgery pertaining to the UCL. It is important for understanding the odds for success following surgery while also understanding risks involved with reconstructive surgery. This adds to the knowledge regarding the prevalence of return to play following the UCL injury process. (Mine et al., 2021) provides an overview of the potential risk factors that are involved with both elbow (UCL) and shoulder injuries in baseball players. This article does not provide sound conclusions within the results for findings other than limited shoulder range of motion. It is important to provide an overview for potential risk factors, but not to make any generalizations on the specific results for these risk factors. Further, it shows the importance to include more statistical factors along with deficiency metrics when determining injury risk. This proves the need for Baseball Savant metrics, the macro data, along with the NewtForce metrics, the micro data. This scholarly discussion's intended utilization is for the importance of understanding potential risk factors. Author and MLB sportswriter Jeff Passan takes a deep dive into the history of arm injuries in baseball players. He explores the advances in surgery technology, advances in rehab, history of the injury, and common misconceptions. He

tells individual stories of medical miracles and mysteries, providing in depth knowledge of the prior history and current state of arm injuries in Major League Baseball. His inclusion of Doctor Neil ElAttrache provides a parallel to one of the scholarly sources listed above (Passan 2016). The mention of major surgeons is shown through one of the variables of the Tommy John data set utilized in analysis.

There is significant research regarding individual deficiencies that have been known to be causes of significant arm injuries in baseball players. From the previous scholarly discussions examined, there were findings that, because of their individual age, remain wholly relevant, somewhat relevant, and not relevant at all. The scholarly discussion is imperative to provide background knowledge within the domain as it establishes a basis for this holistic analytical approach. While individual research of deficiencies and exploration is important, it is the holistic inclusion of numerous data points and platforms that will be incorporated into this research. In this way, the emphasis remains on depth, complexity, and offering future considerations based on the findings.

Ethical Considerations

The data of this study was collected previously and uploaded to various databases. However, ethical considerations are still important to note. Alpha Baseball has granted access of NewtForce Mound data for usage in this analysis exclusively. There are nine players included in the NewtForce Mound dataset, and each of these athletes have given consent for their data and names to be utilized within this analysis. The NewtForce data of these Alpha Baseball athletes was uploaded to the cloud system at the Alpha facility in Mason, Ohio. It was shared with the

express consent of player development directors and coaches, Mitchell Bault and Gregg Williams.

Baseball Savant data is publicly available for personal use of exploration and analysis. Even though there are human subjects involved, there is no blinding involved in this analytical process, as the project does not constitute an experiment. There is no specific treatment group, even though there are two groups of players, in theory (MLB players and Alpha athletes). The last ethical consideration pertains to the usage of a publicly available Tommy John list. Credit belongs to Jon Roegele (MLBPlayerAnalys on X) for creating, updating, and outsourcing the sheet. Created on Google Sheets, the Tommy John Surgery List includes specific information relating to every surgery that has been documented and publicized by all professional baseball players, dating back to the very first surgery performed on the namesake of the surgery, Tommy John. A specific note is made that some players may be missing, as some organizations tend to be more private with their release of specific medical information pertaining to Tommy John surgeries and other procedures. Further explanation of variables and data sets will be explored in the Data section.

Data & Methods

Data

As previously mentioned, there were several platforms from which data was collected and utilized within this analysis. There are two groups in which the data can be segmented into. These two groups will consistently be referred to throughout the paper as ‘macro’ and ‘micro’ data. Macro data includes the Major League Baseball information, as it is more generalizable. It also includes a significant number of observations, meaning the sample size is large and includes nearly the entire population of Major League Baseball data as it pertains to pitchers from 2015-

2024. Micro data refers specifically to the Alpha Baseball information, as it is less generalizable. This is due to the lack of players, meaning the small N of the sample size.

Because the research was performed through RStudio, Sean Lahman's Baseball Database also plays a key role. While the Lahman database contains various datasets and sub datasets, only the 'People' dataset was utilized. From the People data set, the full name of the player, weight, height, batting side, and throwing side were used. Some cleaning was needed, as it was concluded that only position players and pitchers that played from 2015 on would need to be included, because the only players for analysis are those in the Statcast Era. A variable labeled as the full name had to be created to make merging of other datasets possible.

The Tommy John Surgery list was next to be implemented into the analysis. The variables of usage were the player's name, the surgery date; team/level/position at the time of surgery, throwing side, return date to the same level of competition, and the recovery time in months. This data set also requires significant cleaning for analysis. Because some players and pitchers have had more than one Tommy John surgery, they belong to more than one row. The data had to be cleaned to combine rows for players, so that one player had one row instead of one player having more than one row with their name. This would skew clarity of analysis and cause duplicates within the results. The Lahman data set was then merged with the Tommy John Surgery list, so that all players had their specific surgery data in a new data set. Those that did not have a surgery have no data included for those variables, which does not hinder analysis.

Next, the implementation of the Baseball Savant dataset was constructed. This dataset involved the most complex cleaning process, as it was the largest original dataset to be included for analysis. The Baseball Savant website allows for direct downloading of a CSV file for personal exploration and analyses. Overall, 37 variables were selected for analysis. The playerID

was automatically included as part of the downloading process. This variable was removed, because the player's name is an important model considering individual injury and surgery data. In this way, the player's name overtakes the necessity to have a player ID. The variables include: player name, year, innings pitched, throwing hand, number of pitchers, arm angle (available after 2021, related to shoulder health), and then individual pitch metrics and statistics. Those metrics include the average velocity (in MPH) and usage (%'s) of: four-seam fastballs, sliders, changeups, curveballs, sinkers, cutters, splitters, knuckleballs, sweepers, slurves, forkballs, and screwballs. These individual pitches are then broken down into three categories: fastballs, breaking balls, and off-speed. Further, these average velocities and usage (MPH and %) were calculated for these categories. These categories have been established because different pitchers throw different pitches; some throw multiple, some throw only one or two. Including every pitch that is thrown by pitchers does allow for analysis, and due to the large N of players (8,264), allows for significance for each individual pitch, as well as the three major categories. This concludes the section pertaining to macro data.

The focus shifts now to the micro data, obtained from Alpha Baseball. The data includes some inconsistent time between the different inputs for some individuals. This is not unexpected, as throwing schedules may be misaligned with tournament schedules, lack of logged sessions or visits to the facility, or injuries. The original factors considered pertained only to the NewtForce data collected in each throwing session, however these metrics alone were not enough to provide a significant analysis of the micro data. Mound data deals primarily with the force output from the lower body and how it corresponds with different periods of the pitching motion (Newtforce 2025). ArmCare App measures the force and strength from the upper half, specifically the throwing arm (ArmCare 2025). NewtForce data includes the force put into the ground towards

home plate, straight down, from back leg to front leg, stride ratio, and player velocity. For the ArmCare App data, following deletion and mishandling of data, there is only the average strength weight included along with the player's name. There was further exposure of new factors to consider following this initial exploration. Height, weight, and handedness (right/left) are relevant inclusions. In addition, BMI is a variable that can be calculated through height and weight and will be an important predictor for significance of velocity and injury prediction, with potential for parallel trends with the macro data. While some of the NewtForce variables will not be utilized, due to their overlap and correlation with other variables, they are all necessary to include for potential analysis. The ArmCare app data includes metrics related to strength as they pertain to recovery methods and how well the arm can produce force independent of the lower half force generation.

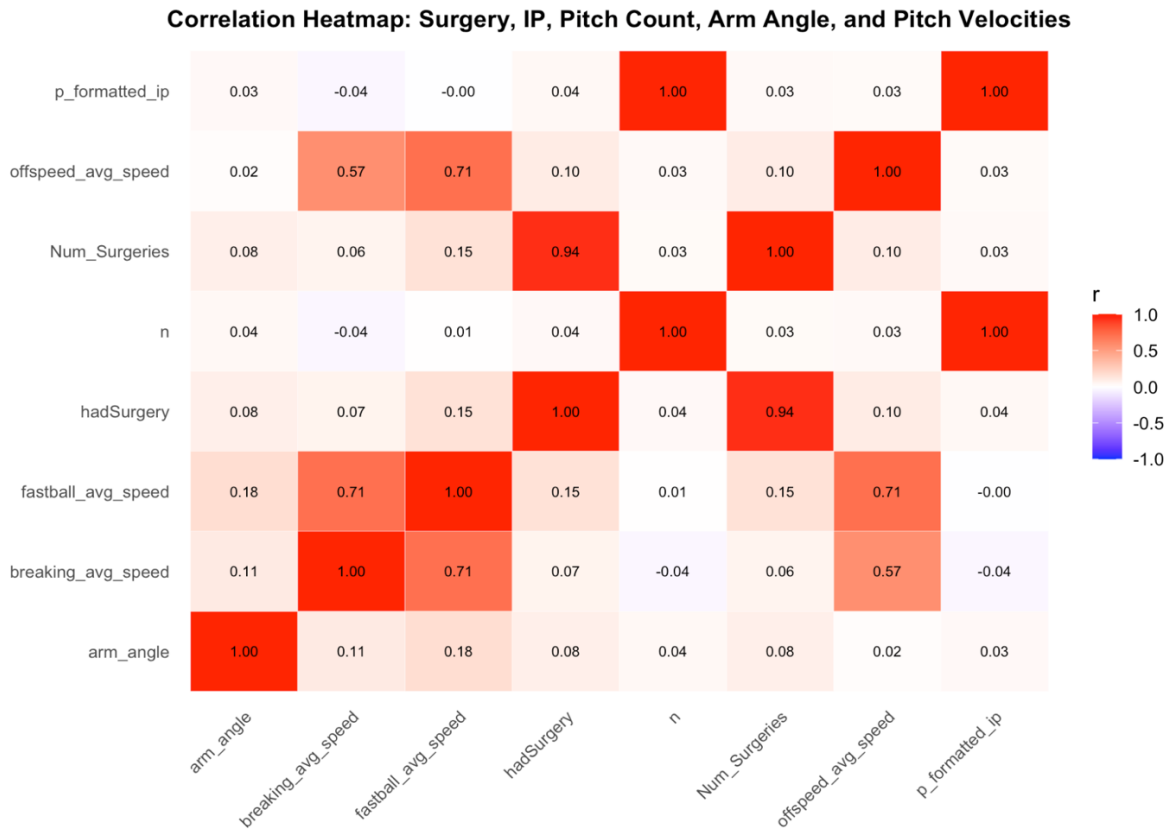
Each of these data sources is important when considered independently. When utilized with one another for a holistic analysis, it becomes increasingly important to analyze the correct variables and understand how they complement one another. For this reason, the merging of the datasets was done methodically and carefully to ensure that there were no inconsistencies or errors when merging. Due to the size of the datasets and the repeated names, it was crucial to establish effective naming techniques and coding comments, so the correct datasets were utilized within the individual analyses and data visual creations. Further, the micro data should not be used as a precursor in understanding the macro data. To conclude, the macro data is included to further the importance of trends in the micro data.

Methods

As has been discussed throughout the paper, a holistic methodical approach is a top priority of this research. There is no one specific method will be classified as the best. Rather, multiple statistical tests and methods will aim to complement and validate each other. These include, but are not limited to, multiple regression capturing both logistic and linear methods, time-series analysis, and interaction models to validate statistical significance and provide further evidence of potential relationships. Each of these methods, independently and holistically, will be imperative to assess the mechanical and kinesiological variables and indicators helpful in preventing arm injuries. The regression models intend to show which specific metrics and variables are going to be likely to increase the odds of injury, which are not, and whether these metrics and statistics will be effective in predicting whether a player will become injured or not. The time-series will aim to uncover trends and patterns with the timing of the injury, with the goal of predicting when the injuries may happen again based on metric trends (Bullock et al., 2022).

As mentioned within the data and variables sections, the key inputs will include the fatigue/strength/recovery variables, throwing workload, pitch arsenal, height and weight, and then mechanical deficiencies and outliers, finishing up with overall mechanical strengths. The first figure for analysis within methods is a correlational heat map, showing variable relationships.

Figure 1



The correlation heat map shows the relationships between different variables in the Baseball Savant data set. The darker the coloring (red/blue), the higher the correlational value. Most of this heat map shows little coloration, showing small correlational relationships, to bright red, showing strong positive correlational relationships. Very few indicate negative correlational relationships, with the lack of blue coloring. From the results of the correlation heat map, there are few variables of which need to be excluded. Some variables need to be removed because the higher their correlation, the more likely they will create a model that is not representative of the true result. For example, if a predictive model is created for whether a player had surgery (hadSurgery), it would not be sufficient to use the number of surgeries (Num_Surgeries) to predict that, because those variables are highly correlated and would thus wrongly influence the

model. On the other hand, because most of the variables are not highly correlated, examining individual relationships beyond correlation will be important to establish alternate interpretations. The results of this heatmap support the decision to include the LASSO model, examined in the results section.

Scatterplots were important exploratory visuals to show the relationship between BMI (height and weight), on pitching velocity. While scatterplots are important and provide sound visual representation of data, they do not include specific statistical outputs. The scatterplots are confirmed and strengthened through the interaction model discussed later in this section (Table 1). There are three total scatterplots for the three different groups explored.

Figure 2A

Figure 2B

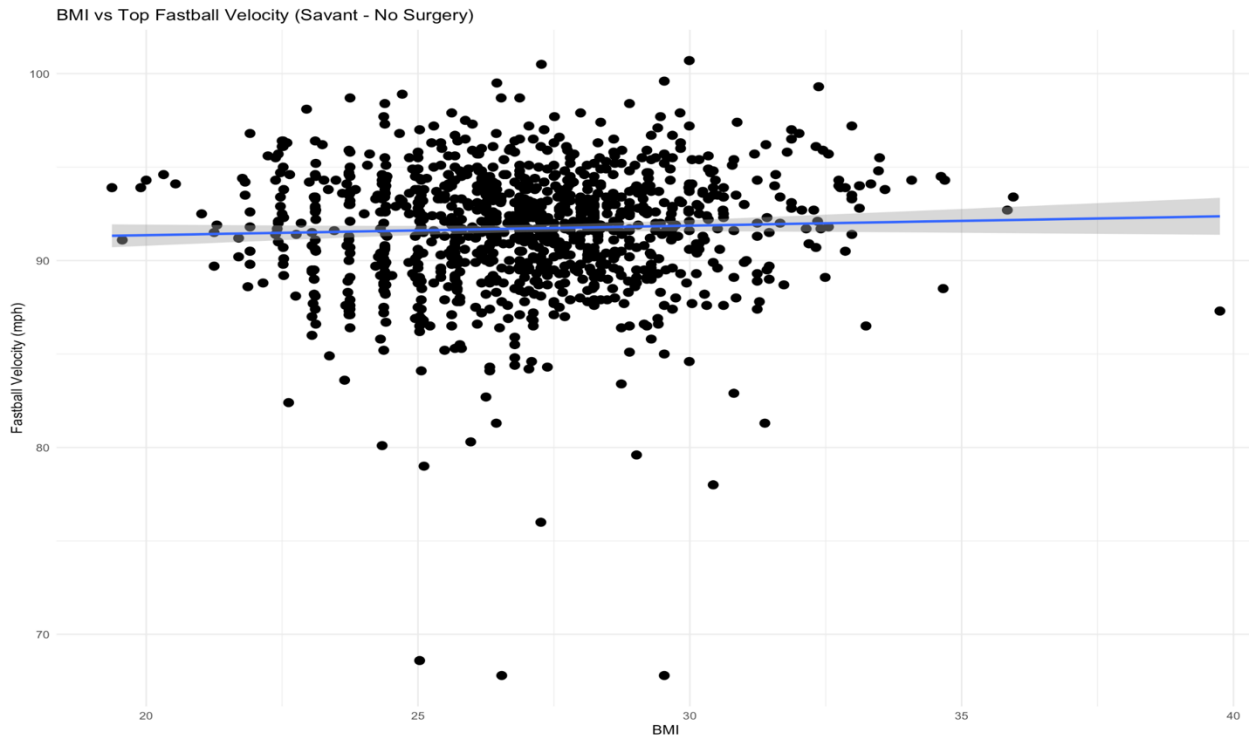
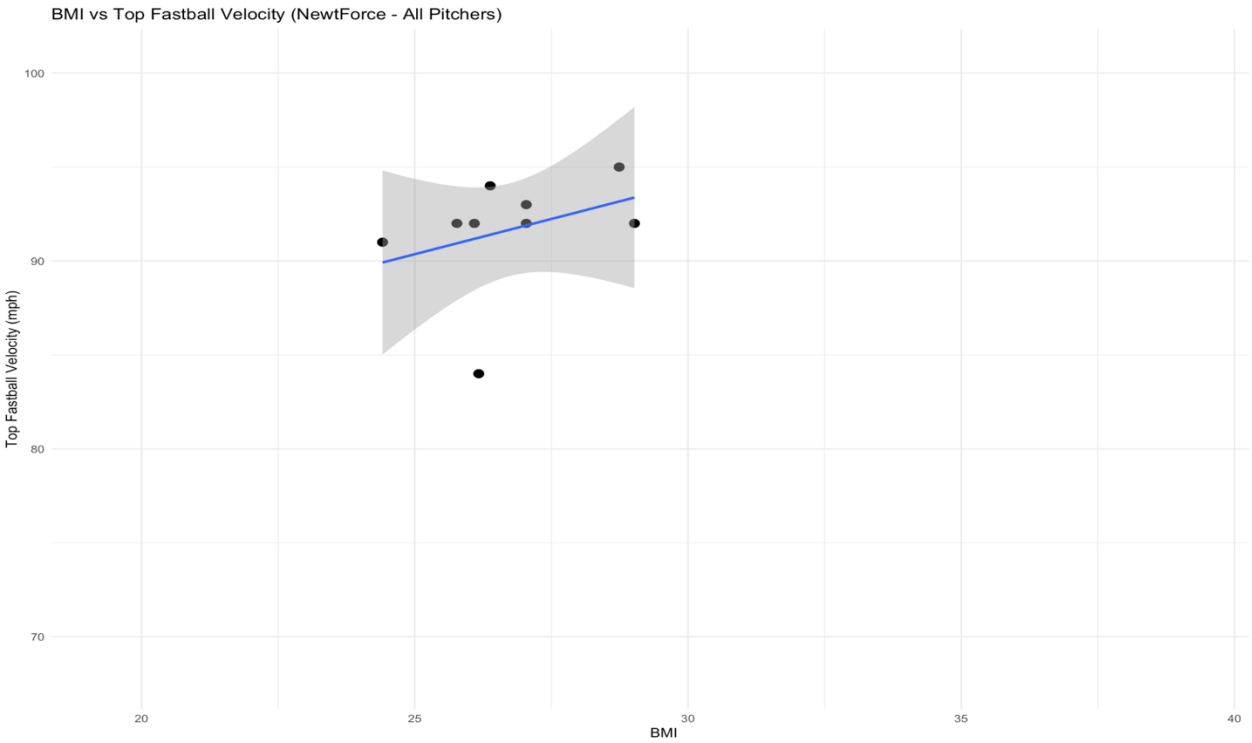


Figure 2C



Figures 2A, 2B, and 2C were included specifically to explore the visual significance of the commonly held theory that BMI is positively associated with an increase in pitching velocity. Figures 2A and 2B explore the trends. Figure 2C does not explore a trend, because there cannot be a trend in the micro data due to the small N. In future analysis, if a greater N is obtained for micro data, this would be an important visual to explore.

An interaction model is like that of a regression model, though in this case it provides statistical basis for the visual scatterplots shown above, specifically Figure 2A and Figure 2B.

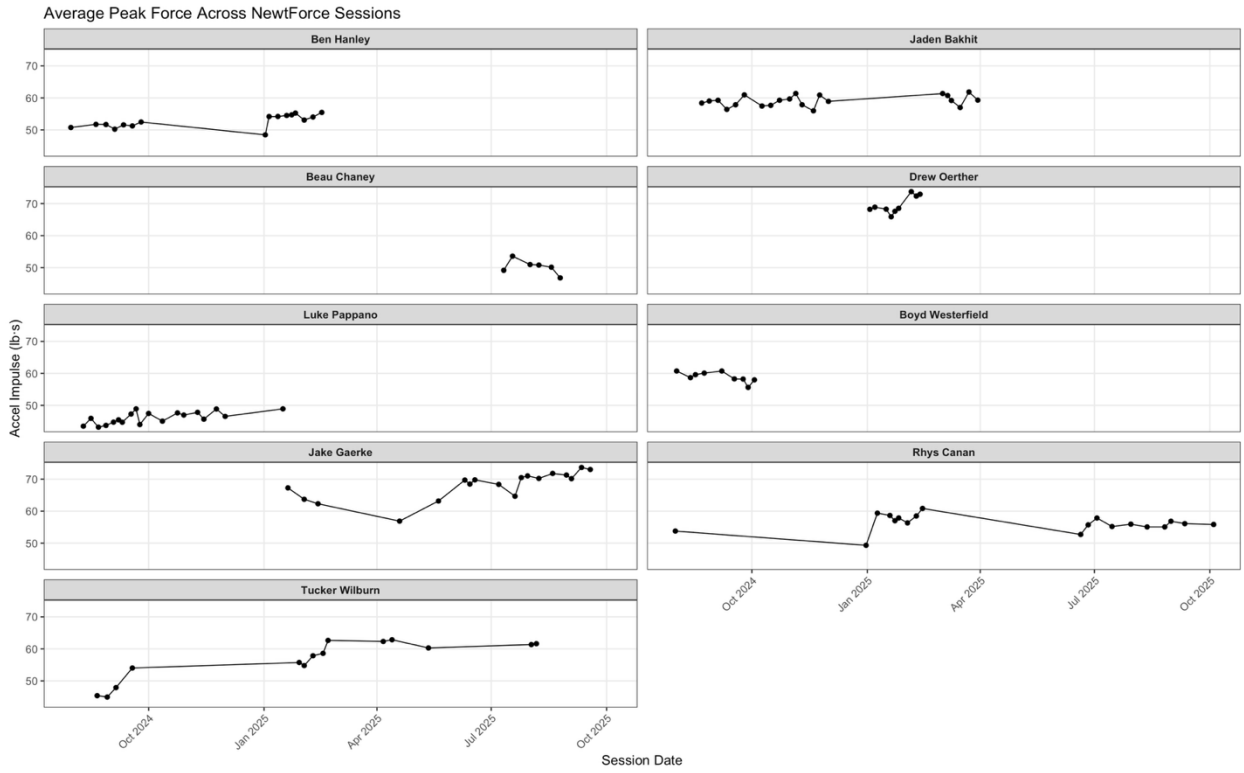
Table 1

Interaction Model Results: BMI, Velocity, and Surgery Group

term	estimate	std_error	t_value	p_value
(Intercept)	90.336	0.984	91.770	0.000
BMI	0.051	0.036	1.407	0.159
groupSavant_Surgery	0.424	1.708	0.248	0.804
BMI:groupSavant_Surgery	0.018	0.063	0.288	0.773

Table 1 shows the results of the interaction model. The specific interaction model in place for this analysis involved examination of BMI and its relationship on predicting fastball velocity for Major League pitchers who have had surgery, versus those who have not. This table further shows the lack of visual difference from Figures 2A and 2B by using statistics to establish further significance of results.

Figure 3



The time series in Figure 3 shows the difference in the average peak force across the different NewtForce bullpen sessions for each individual athlete at Alpha Baseball. The average peak force measures the lower half efficiency across players coming back from Tommy John surgery. These measures show growth, yet they also show areas for potential weakness. The peaks show signs of increased strength and positive lower half efficiency, indicating less stress on the arm. The valleys, however, raise questions of whether too much stress is being placed on the arm. These time series do exemplify the changes shown from session to session and attempt to give greater insight to the day and instance of injury during the return to throwing process. Though this visualization does not directly support the inclusion of the LASSO model, it is an important note and thus is a method.

The most significant portion of the results will examine a LASSO logistic regression model. This will take the most qualified variables and put them into a model to see how likely a

prediction can be made on whether a pitcher will need surgery based on the selected variables. The LASSO model will likely be able to perform better than regular logistic regression due to the presence of more inconsistent trends/results from athletes who do not have consistent data inputs (Karnuta et al., 2020). This methodological, holistic approach seeks to provide a sound conclusion to the research question and will show trends and relationships between various datasets, metrics, statistics, and results.

Results

This section includes three major results of the most important model: Figure 4A, Figure 4B, and Table 2. The inclusion is limited to these three figures and tables to emphasize their importance. The LASSO model aims to predict whether a pitcher will need surgery based on metrics and statistics variables from the fully merged dataset.

Figure 4A

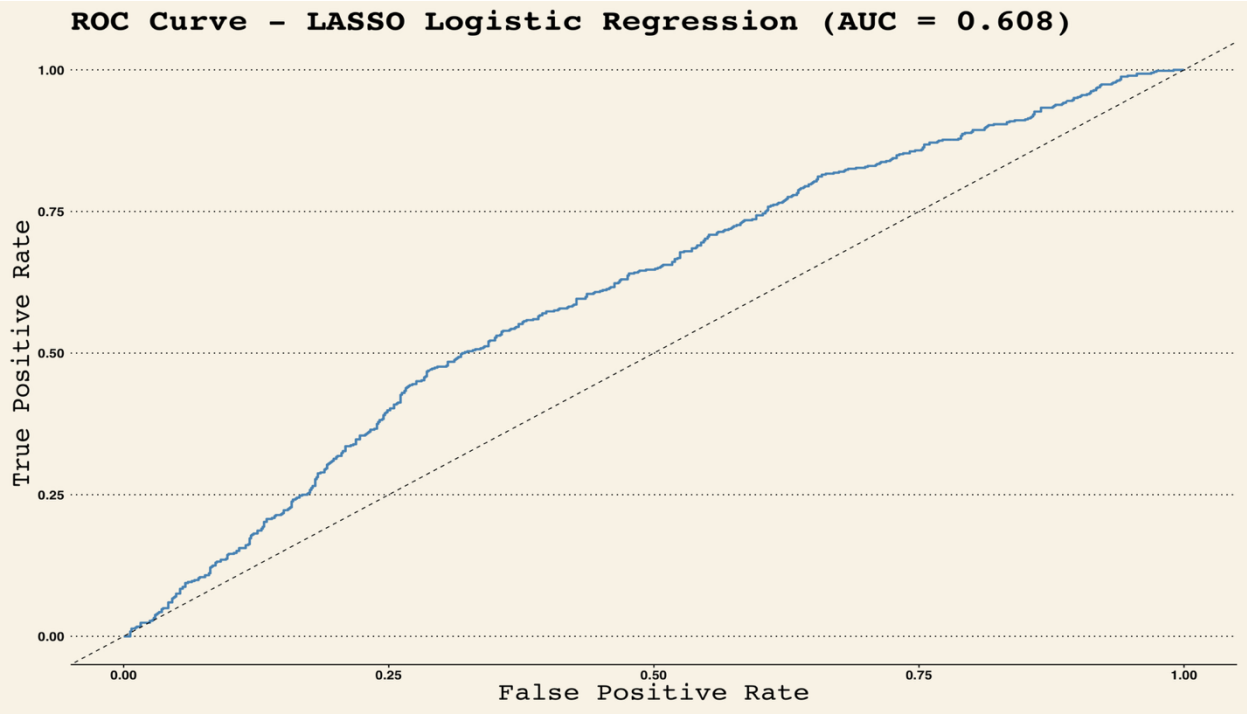


Figure 4A shows a visual representation of the ROC curve of the LASSO logistic regression model. It shows the relationship of obtaining a false positive, meaning falsely predicting a pitcher needing surgery when he will not need one, and obtaining a true positive, which is predicting a pitcher needing surgery when he will need one. The AUC value, which is 0.61, indicates the likelihood of correct prediction. An AUC of 0 indicates the model is never correct in its prediction, AUC = 0.5 indicates the model is synonymous with a coin flip, and an AUC = 1 indicates a perfect model. In this model, the AUC = .61, which indicates a modest model of correctly predicting whether a pitcher will need surgery.

Figure 4B

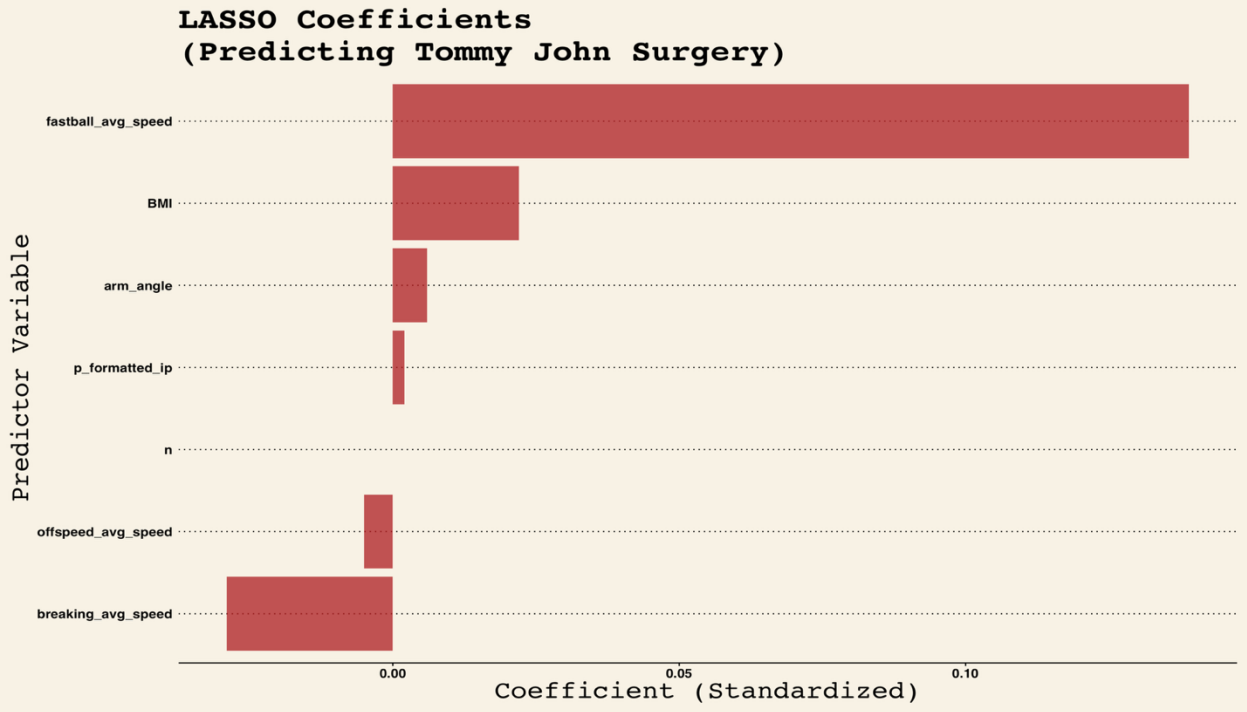


Figure 4B shows the individual breakdown of the variables and their influence on the LASSO model. These seven variables were selected with the purpose of creating the best model, based on previous correlation tests and training results. The variables are fastball average velocity (MPH), BMI, arm angle, innings pitched, number of pitches in a season, off-speed

average velocity (MPH), and breaking-ball average velocity (MPH). This figure is included to show how much more significant variables are compared to others. For example, fastball average velocity is significantly more important at predicting injury according to the model than number of pitches, or innings pitched.

Table 2

LASSO Coefficients and Model AUC

variable	coefficient	AUC
fastball_avg_speed	0.139	0.608
breaking_avg_speed	−0.029	0.608
BMI	0.022	0.608
arm_angle	0.006	0.608
offspeed_avg_speed	−0.005	0.608
p_formatted_ip	0.002	0.608
n	0.000	0.608

Table 2 is included to show the statistical results of Figure 4B. Instead of a visual representation of the variables, it includes the statistical results to show numerically how each of the variables impacts the LASSO model. The AUC is included next to the coefficient of each variable for consistency. The variable names are not changed because they are explained for their individual meanings under Figure 4B.

Discussion

The discussion section will focus mostly on the implications of the figures of the results section and will also examine the results of the visuals shown in the methods section. Even though the ultimate model was LASSO logistic regression, the other exploratory analysis elements are useful to include within this discussion section.

The correlation heat map does not create significant cause for concern. The heat map was included to test whether variables are highly correlated with one another. If unexpected variables are highly correlated, this clouds potential predictive analyses. Having two variables that are highly correlated will then cause those variables to be better at prediction when analysis is conducted. This inherently makes the models insignificant. The correlation model does not raise any concerns with high correlation values. Rather, the variables all seem to be weak to moderate positive correlations. This indicates that there does not need to be concern with highly correlated variables and predictive analysis can be run.

Figures 2A, 2B, and 2C show a visual representation of the relationship between BMI and fastball velocities. Figures 2A and 2B utilize MLB pitcher data. Figure 2C uses the Alpha Baseball athlete data. Figures 2A and 2B are important because they appear to show no visual difference in the discussed relationship between pitchers that have had surgery and pitchers that have not had surgery. On the other hand, Figure 2C appears to show a positive relationship between BMI and fastball velocity. Figure 2C cannot show a trend, the number of observations is not nearly large enough to prove any statistical significance. It does offer an interesting interpretation, in that age may play a difference. These athletes are 15-18 years old, whereas the players in the MLB data set are 20-45 years old. It proves that sequencing of biomechanics paired with maturity likely plays a larger role than weight and size. There are, of course,

statistical outliers, players with very low and very high BMI's that may skew the results, however N is large enough for Figures 2A and 2B to negate the impact of these outliers. More data and a greater N would be necessary for the micro data in this case to show a trend, as is shown in Figures 2A and 2B.

From the results of the interaction model, Table 1, we see that BMI is responsible for roughly 2% of all variables in predicting the average fastball velocity. This shows that BMI is not a significant predictor of fastball velocity. Further, it shows there is not a statistically significant difference in the average fastball velocity for MLB pitchers who have had Tommy John surgery as compared to MLB pitchers who have not had the surgery. The p-value meets the criteria of falling below the 0.05 level of significance, indicating these results are statistically significant.

The LASSO logistic regression includes three parts. The visualization of the ROC curve (Figure 4A), the coefficient of variables visual (Figure 4B), and the coefficients table (Table 2). Each of these sections is important to understanding what the model is doing. The ROC curve captures the relationship of the true positives and false positives. A true positive is whether the model correctly predicts a player to have surgery, based on the variables and coefficients of those variables included within the model. A false positive is just the opposite. False positives, in this case, are when the model wrongly predicts a player to have a surgery when in fact they did will not need a surgery. The area under the curve, or AUC, has a value around 0.61. This indicates that the model is modest at predicting whether a pitcher will need surgery. The key variables in predicting these injuries are fastball average velocity in MPH and BMI. Breaking ball average velocity in MPH interestingly has a negative coefficient, indicating that higher breaking ball velocity indicates that a player may be less likely to have a surgery.

The implications of this LASSO model and the other methods mainly show what can be classified as a null result. Though a null result seems to have a negative connotation, a null result in this case offers important interpretations. There must be more analyses conducted, with different platforms and bases of data to provide more sound conclusions. It is not incorrect, then, to interpret and conclude that injuries are subjective and pertain to more individuality than they are originally given credit for. Genetic makeup, individual metrics such as strength and recovery, and even individual workload might be important inclusions. Overall, holistic analyses are not sufficient for group research, rather holistic analyses should be implemented on individual data. These results are, unfortunately, heavily exploratory with exception to the LASSO regression model. They do give reasoning to support the inclusion of the LASSO model and give reasoning to negate group-based analysis in the case of injuries. Pitcher injuries, specifically injuries to the UCL, have too many individualistic factors to be generalized to a larger group of individuals, such as MLB pitchers and high school athletes.

Conclusion

The research question for this analysis is, can arm injuries, specifically injuries to the UCL, be prevented or predicted through analysis of various data points and systems that expose indicators of weakness or potential injuries? This question was introduced to gain insights into why UCL injuries are happening and what, if anything, can be done to prevent them. While the results were not definitive enough to provide injury conclusions, the ‘null’ results do express an answer. Arm injuries are subjective and pertain to the individual. Because each pitcher has individual differences in genetic makeup, mechanics, and pitch arsenals, it is difficult to generalize results and findings to groups of pitchers. It does appear that there are weak

associations between faster pitching velocities of fastballs, innings pitched, arm angle, and BMI leading to a greater chance of requiring surgery from a UCL injury. Further analysis of biomechanical data must be observed. It would be helpful to obtain greater NewtForce results for lower-half metric generalizations to be concrete. It would also be helpful for MLB pitchers to have lower-half metric data readily available, but MLB pitcher access to NewtForce mound across the league is impractical. From here, further methods and results with statistics must be established for further bases and conclusions. These weak relationships should not be discredited because they are weak, rather understood for their statistical meaning. Pitchers should proceed with caution and take extra recovery and rehabilitation methods when throwing consistently at higher velocities in greater volumes.

References

Bartlett, L. R., Storey, M. D., & Simons, B. D. (1989). Measurement of upper extremity torque production and its relationship to throwing speed in the competitive athlete. The American Journal of Sports Medicine, 17(1), 89–91. https://doi.org/10.1177/036354658901700115

Brown, L. P., Niehues, S. L., Harrah, A., Yavorsky, P., & Hirshman, H. P. (1988). Upper Extremity Range of Motion and Isokinetic Strength of the Internal and External Shoulder Rotators in Major League Baseball Players. The American Journal of Sports Medicine, 16(6), 577–585. https://doi.org/10.1177/036354658801600604

Conte, S. A., Fleisig, G. S., Dines, J. S., Wilk, K. E., Aune, K. T., Patterson-Flynn, N., & ElAttrache, N. (2015). Prevalence of Ulnar Collateral Ligament Surgery in Professional Baseball Players. The American Journal of Sports Medicine, 43(7), 1764–1769. https://doi.org/10.1177/0363546515580792

Garrison, J. C., Arnold, A., Macko, M. J., & Conway, J. E. (2013). Baseball Players Diagnosed With Ulnar Collateral Ligament Tears Demonstrate Decreased Balance Compared to Healthy Controls. Journal of Orthopaedic & Sports Physical Therapy, 43(10), 752–758. https://doi.org/10.2519/jospt.2013.4680

Garrison, J. C., Cole, M. A., Conway, J. E., Macko, M. J., Thigpen, C., & Shanley, E. (2012). Shoulder Range of Motion Deficits in Baseball Players With an Ulnar Collateral Ligament Tear. The American Journal of Sports Medicine, 40(11), 2597–2603. https://doi.org/10.1177/0363546512459175

Jang, S.-H. (2019). Management of Ulnar Collateral Ligament Injuries in Overhead Athletes.

Clinics in Shoulder and Elbow, 22(4), 235–240. https://doi.org/10.5397/cise.2019.22.4.235

Major League Baseball. (n.d.). *Statcast | Glossary*. MLB.com. https://www.mlb.com/glossary/statcast

Mine, K., Milanese, S., Jones, M. A., Saunders, S., & Onofrio, B. (2021). Risk Factors of Shoulder and Elbow Injuries in Baseball: A Scoping Review of 3 Types of Evidence. Orthopaedic Journal of Sports Medicine, 9(12), 232596712110646. https://doi.org/10.1177/23259671211064645

Passan, J. (2016). The Arm: Inside the Billion-Dollar Mystery of the Most Valuable Commodity in Sports. HarperCollins.

Peters, S. D., Bullock, G. S., Goode, A. P., Garrigues, G. E., Ruch, D. S., & Reiman, M. P. (2018). The success of return to sport after ulnar collateral ligament injury in baseball: a systematic review and meta-analysis. Journal of Shoulder and Elbow Surgery, 27(3), 561–571. https://doi.org/10.1016/j.jse.2017.12.003

Petersen, G. (2025, November 6). *UCL injuries rising among younger baseball players, doctors say*. Https://Www.wmbfnews.com; WMBF. https://www.wmbfnews.com/2025/11/06/ucl-injuries-rising-among-younger-baseball-players-doctors-say/

Roegele, J. (Accessed November 2025). *Tommy John Surgery List*. Google Sheets. https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/edit?gid=0#gid=0

Scher, S., Anderson, K., Weber, N., Bajorek, J., Rand, K., & Bey, M. J. (2010). Associations Among Hip and Shoulder Range of Motion and Shoulder Injury in Professional Baseball

Players. Journal of Athletic Training, 45(2), 191–197. https://doi.org/10.4085/1062-6050-45.2.191

Ulnar Collateral Ligament (UCL) Injury. (2022). www.nationwidechildrens.org. https://www.nationwidechildrens.org/conditions/ulnar-collateral-ligament-injury

Wilk, K. E., Meister, K., & Andrews, J. R. (2002). Current Concepts in the Rehabilitation of the Overhead Throwing Athlete. The American Journal of Sports Medicine, 30(1), 136–151. https://doi.org/10.1177/03635465020300011201

Appendix

https://github.com/willrettig36/Will-Rettig-DA-401-Final-Project