DSC680

Chris Kellogg

**MVP or TJS?  
*Using pitch data to predict elbow injuries in MLB pitchers***

**Business Problem**

Major League Baseball (MLB) teams lose significant amounts of money and face wasted post-season opportunities when pitchers miss time with elbow injuries. I’ll attempt to answer the question: *Are there patterns in readily available data that indicate that a pitcher might be at a higher risk for elbow injuries?*

**Background**

MLB teams spend a tremendous amount of money paying players in general. Over the last 3 years, the total yearly commitment to the top 10 players has increased from $363 million to $418 million, and even if no other new high-priced free agent contracts are signed, the salary commitment to the top 10 players total will remain above $420 million (Spotrac, 2025). Starting pitchers account for the bulk of the largest annual contract commitments, ranging from 32-42 out of the top 100 contracts in the years from 2022-2027.

There is a ligament in the elbow, the ulnar collateral ligament (UCL), which can become damaged due to the stress of repeated overhand throws, especially high effort throws. A procedure to repair a UCL was attempted on Los Angeles Dodgers pitcher Tommy John in 1974 by orthopedic surgeon Frank Jobe, Dodgers team doctor (Ulnar collateral ligament reconstruction, 2025). Because Tommy John was the first pitcher to return to pitching at his previously high level, UCL reconstructive surgery is colloquially called Tommy John surgery (TJS).

The AMA indicates that more than 1,000 professional pitchers have had UCL reconstruction since 1974, and that over 35% of active major-league pitchers had undergone TJS in 2023 (Berg, 2024). A throwing athlete with surgically reconstructed UCL can typically require up to nine months before returning to throwing, and often a year or more before building up enough strength and stamina to return to high-level competition (Johns Hopkins Medicine, 2025).

Because the salary of a player who is unavailable to his MLB team due to injury still counts as part of the team’s payroll, and because there are financial penalties to an MLB team to discourage too high a payroll (MLB, 2025), it’s a serious problem for a team when a pitcher suffers an injury that keeps him out for such an extended period. There would be a significant advantage to any MLB team/organization that could identify patterns in innings load, pitch selection, or pitch measurements that might indicate a higher risk of elbow injury for pitcher.

The number of pitcher injuries and days missed have both more than doubled over the last 20 years (NBC Sports, 2024). Various studies suggest that pitch count (Wikipedia, 2025), velocity (Berg), max effort pitches (NBC Sports), and spin rates (NBC Sports) might all be factors in the increased injury numbers.

I’ll be investigating several observable and widely available statistics to determine their predictive ability on whether a pitcher will have an elbow injury requiring surgery in the next year. This information could help rule makers to legislate environments that protect pitchers, pitchers to better advocate for themselves in injury prevention, and teams to use their pitchers more safely and maximize their considerable investments.

**Datasets**

*Pitch data*<https://baseballsavant.mlb.com/statcast_search> or pybaseball API for python

I used the pitch data from 2000-pitch seasons over the last 4 seasons. About 100 pitchers per year throw 2000 pitches in MLB regular-season games. I collected the data for each “season,” meaning the total number of pitches thrown, number of fastballs thrown, maximum & average fastball velocity, how close the pitcher was to “max effort” on his fastballs (as average fastball velocity divided by maximum fastball velocity), number of breaking balls thrown (which have been claimed to put extra stress on an elbow), maximum & average spin rate for the breaking balls, and how close the pitcher was to “max spin effort” on his breaking balls (as average breaking ball spin divided by maximum breaking ball spin).

*Surgery information: Tommy John Surgery List (@MLBPlayerAnalys)*<https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/>

I used the names and surgery dates to classify each “season” according to whether the pitcher had TJS in the 12 months after the season ended. This approach will classify a particular season as “preceded TJS” not only if the pitcher has surgery in the ensuing offseason, but also if the pitcher’s next season is cut short by surgery. This classification approach isn’t perfect; for example, it wouldn’t account for a pitcher who threw 2000 pitches in a season, then struggled on and off with elbow pain for most of the next season before having surgery. Those cases were extremely rare and would have required significant case-by-case research, so they were out of scope for this project.

*Data Preparation (See appendix A)*

My basic datasets were available through a public API (pybaseball, leveraged from MLB’s Statcast data) and a public Google sheet from @MLBPlayerAnalys. While the pitch data required some transformation to be in a usable form for my purposes, the spreadsheet data was quite straightforward after I downloaded it as a CSV. I used python in a Jupyter notebook to get the data and then build and combine dataframes.

Both the API and the spreadsheet provided the MLB Advanced Media identifier (mlbam\_id) for each pitcher, making it easy to combine the datasets and ensuring accuracy. Both datasets were consistent in the data they provided, with really the only data preparation issue arising from a few pitches that had missing velocities or spin rates. In order to stay conservative, I dropped thos pitches from the datasets.

I merged the two separate datasets (pitches and surgery info) on mlbam\_id and year to create the master dataset for my analysis. For each 2000 pitch season, the master dataset shows the aggregated pitch data and the label indicating whether the pitcher underwent surgery to repair a UCL in the 12 months after the season was complete.

**Methods**

*Logistic regression*

I split the 400-ish seasons into training and testing subsets, scaled the data, and fitted a logistic regression to the labeled training dataset. I then attempted to predict the correct classification for the testing dataset and compared the results to the actual labels for the testing set.

*Confusion Matrix*

I used a confusion matrix to evaluate how effectively the logistic regression predicted the test classifications and was tremendously disappointed to see the results.

*Figure 1. How many wins does a team need to reach the playoffs?*A screenshot of a computer code

Description automatically generated

I used 25% of the data set for testing (100 records) and expected to classify 8 of them as seasons that preceded a UCL reconstruction, while the other 92 records should have been classified as healthy seasons (requiring no surgery). Instead, all 100 test records were classified as healthy, with the model predicting NO elbow injuries at all. In fact, the model seems entirely satisfied with predicting every season to be a healthy one and accepting the 92% accuracy of that approach.

*Looking for patterns in the variables*

I decided I would just inspect the data for each of the 9 features to see if any of them looked capable of providing any predictive on an individual basis.

*Figures 2-4. Distribution of individual features with seasons preceding TJS shown in red  
A graph of numbers and a line of dots

Description automatically generated with medium confidence*

*A diagram of a number of red dots

Description automatically generated*

*A graph of a number of objects

Description automatically generated with medium confidence*

*Figures 5-10. Distribution of individual features (cont’d) with seasons preceding TJS shown in red*

*A diagram of a number of dots

Description automatically generated*

*A graph with red dots and black text

Description automatically generated*

*A graph of numbers and a line of dots

Description automatically generated A graph of a number of red dots

Description automatically generated*

*A diagram of a number of red dots

Description automatically generated*

*A graph with numbers and a red dot

Description automatically generated*

As shown in the strip plots, the distribution of the red dots (reflecting seasons that preceded TJS) for each feature is very similar to the overall distribution of all seasons.

**Analysis**

The logistic regression was simply not predictive. It didn’t predict a single season to precede a UCL surgery, even though there were several such seasons in each year of the dataset. I ran a correlation matrix to look for the possibility that there was some predictive value in at least one of the features.

*Figure 11. Correlation of each potential feature with the target classification  
A close-up of a person's face

Description automatically generated  
A screenshot of a computer code

Description automatically generated*

I was really surprised to learn that none of the features provided even a hint of predictive value. As often as I’ve read that torque on the elbow damages the ligaments, or that the quest for velocity has led to more injured pitchers, I was confident that at least one of the features would correlate with the target class.

**Conclusion**

Over the limited data that I used, none of the nine features that I collected were useful for predicting an elbow injury that required surgery.

**Assumptions/Limitations/Challenges**

I think one limitations in the project might have been that I didn’t have enough members in the target class. I’m not entirely sure that would have mattered, though, simply because the features alone didn’t even demonstrate a hint of correlative value.

**Future Uses**

In my exploration and analysis, I found research indicating that other projects have found significant evidence to indicate that higher velocity and a higher volume of breaking pitches do in fact lead to more pitching injuries. There may be other data that I need to consider, or I may need a higher volume of useful data in order to create a better model.

**Recommendations**

I think the most reasonable recommendation I could make is that teams, organizations, coaches, and pitchers continue to educate themselves on the risk of injury and ways to avoid increased risk.

**Implementation Plan**

As more ballparks (and especially more minor league ballparks) are fitted with data collection systems like Statcast, I’m hopeful that more of that data becomes publicly available. More data can only lead to improvement in research like this.

**Ethical Assessment**

There’s little reason for concern about the ethics of using the data, since it’s all publicly available from multiple sources, including through MLB itself. There’s also no likely ethical impact from the results, since all 30 MLB organizations and many independent sites are already doing similar modeling.

**References**

Spotrac. (viewed 2025, January 4). Spotrac. MLB Salary Rankings.  
https://www.spotrac.com/mlb/rankings/player/\_/year/2025/sort/cap\_total

Wikipedia. (viewed 2025, January 4). Ulnar collateral ligament reconstruction. <https://en.wikipedia.org/wiki/Ulnar_collateral_ligament_reconstruction>

Berg, S. (2024, July 12). AMA. What doctors wish patients knew about Tommy John surgery.  
<https://www.ama-assn.org/delivering-care/public-health/what-doctors-wish-patients-knew-about-tommy-john-surgery>

Johns Hopkins Medicine. (viewed 2025, January 4). Tommy John Surgery (Ulnar Collateral Ligament Reconstruction).  
<https://en.wikipedia.org/wiki/Major_League_Baseball_postseason>

MLB. (viewed 2025, January 4). Competitive Balance Tax.  
https://www.mlb.com/glossary/transactions/competitive-balance-tax

NBC Sports. (2024, December 17). MLB study: Velocity, max efforts likely causing pitching injuries; rule changes should be considered.  
<https://www.nbcsports.com/mlb/news/mlb-study-velocity-max-efforts-likely-causing-pitching-injuries-rule-changes-should-be-considered>