

Balancing Performance and Longevity: Modeling Injury Risk in MLB Pitchers

Oscar Su, John Tan, Yuhan Xie, Patrick Xue, Erica Yee

December 9, 2025

Abstract

A range of factors lead to a long and successful career as a pitcher in the MLB. This study investigates the determinants of career longevity and injury risk among Major League Baseball (MLB) pitchers using performance data from 2010 to 2020. The study aims to identify which factors contribute to career duration and whether specific pitch types (fastball, breaking, or offspeed) correlate with injury susceptibility or specific performance outcomes. We use univariate t-tests to distinguish differences in factors between short (< 5 years) and long careers, followed by regression analysis to model the impact of performance variables. Our results indicate that total outs scored, higher average fastball velocity, and lower walk rates within a player's first two seasons are the only statistically significant predictors of increased career longevity. Regarding injury risk, we find that a higher cumulative fastball workload from the last three games is associated with a 3% increase in the odds of an injury per additional pitch, whereas other pitch types and rest days show no significant effect. Performance analysis reveals that while all pitch types contribute equally to generating outs, breaking balls maximize strikeout probability, fastballs correlate with higher walk rates, and offspeed pitches carry the highest risk of yielding home runs. These findings suggest that managing fastball workload is critical for balancing immediate performance with long-term durability.

Keywords: Sports Analytics, Major League Baseball, Pitching, Career Longevity, Injury Risk Modeling, Generalized Linear Models (GLM)

Introduction

Major League Baseball (MLB) teams invest substantial resources in developing young pitchers, but early-career performance, workload, and durability vary dramatically between players. As a result, identifying reliable early indicators of long-term success remains a central challenge for player development, roster planning, and injury prevention.

We aim to analyze pitching performance data from 2010 to 2020 to determine key variables that coaches should look out for when **maintaining and improving** their players. Winning a season is not a sprint, but a **marathon**. It is crucial to keep players in shape and have high performance whilst minimizing the risk of injury. This is particularly important for a pitcher, who often has higher injury rates compared to any other defensive position in baseball (Carr et al., 2022), due to the repetitive and high-force nature of a pitch.

Background

Previous research suggests a strong connection between performance and players' health in MLB pitching; studies have shown that increased pitching velocity and higher workloads tend to result in an increased risk of injury. Increases in velocity put additional stress on the shoulder and elbow. Added stress when applied to underdeveloped tissue in the presence of overuse and poor pitching mechanics can lead to serious injury (Coleman, 2024). Most injuries are to their Ulnar Collateral Ligament (UCL), which often requires Tommy John surgery. Recoveries typically sideline pitchers for about 17 months or later. Missing games for long periods of time can cause young pitchers to lose necessary development time. For veteran pitchers, injuries might potentially end their careers. Our analysis highlights the importance of how performance metrics and workload patterns influence injuries and longevity.

Data

The MLB pitching dataset is taken from MLB's PitchF/x system and contains various pitchers and their games played from the years 2010 to 2020. Each row is one pitcher's performance in one game. The dataset has multiple data frames containing information related to birth country, injury status, pitching performance etc. In addition, the dataset also provides information on the sequence of pitches during a game (sequence of pitch types, speed of each pitch, height release point of each pitch etc.), which can be used to analyze the workload and style of a pitcher across games. We have engineered the data to extract and create the relevant variables that will be used in answering each research question.

Exploratory Data Analysis

To examine the variation in career lengths, we plotted a histogram of the career lengths of all pitchers within the dataset. The variable tracking career length (seniorityYrs) is a continuous variable measuring the number of years since a pitcher's debut game. In Figure 1, we see a right skew in the histogram, since baseball careers are substantially skewed towards early exit and few players retire with exceptionally long careers. To ensure a fair comparison with balanced sample sizes, we defined the threshold for a 'long career' as longer than 5 years. This cut-off divides the dataset into roughly equal cohorts and closely aligns with previous studies estimating the average player's career length (Witnauer et al., 2007). With such a large proportion of pitchers exiting the MLB within their first few seasons, identifying the early indicators helps teams distinguish between pitchers who are likely to remain short-term contributors and those who may develop into long-term assets.

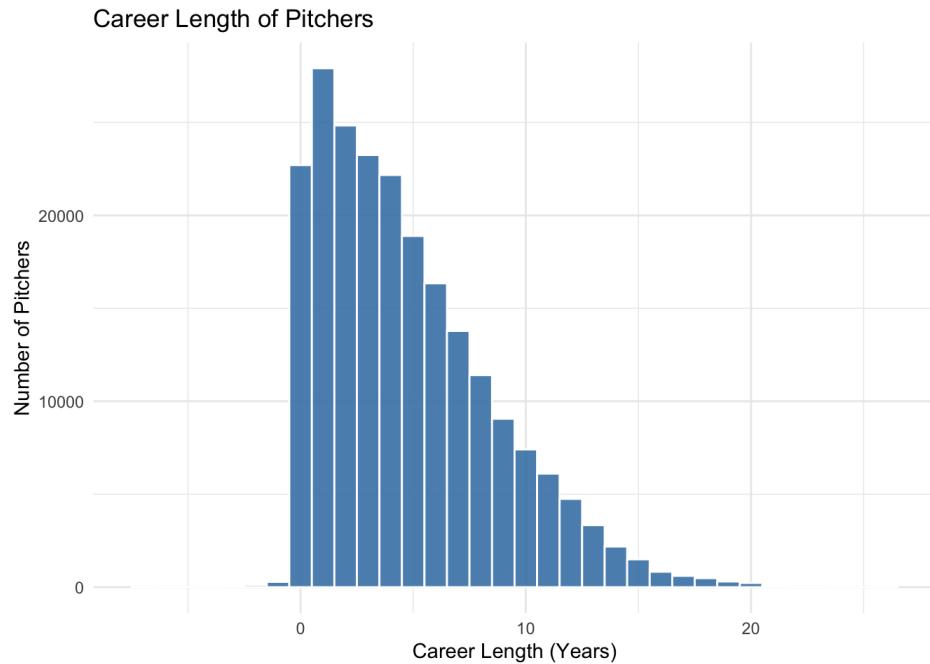


Figure 1: Distribution of total MLB career length (in years) for pitchers

Figure 2 shows the distribution of performance metrics: home runs (HR), strikeouts (K), outs, and walks (W) from a pitcher's first two seasons. Since the variables are all count variables, the distributions are in discrete form. The distributions somewhat vary in shape, but all have a right skew. The limited variance in variables such as home runs (HR) presents a potential limitation in our statistical analysis and should be considered when interpreting key findings. However, the observed variation in other performance metrics raises an important question: do specific pitch types drive these outcomes and can they be leveraged to improve performance without increasing injury risk?

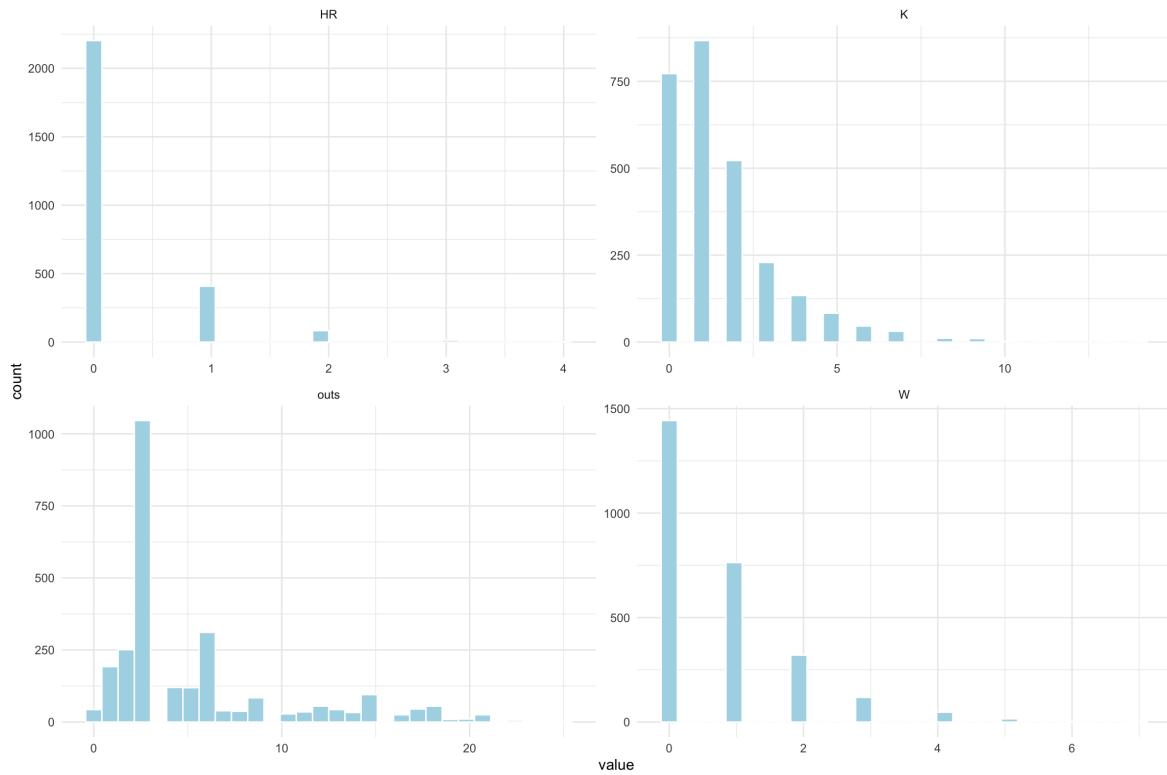


Figure 2: Distribution of performance metrics

Research Questions

For this study, we are investigating two research questions that together provide valuable insight to MLB baseball teams on how they can better support their pitchers.

- 1. Factors contributing to Career Length:** Which variables in a pitcher's early years contribute the most to a long career length?

In particular, we would like to know the qualities that pitchers who end up having a long career possess during their rookie years, allowing coaches and team managers to recognize these potential players early on and support their growth.

- 2. Pitch types on injury & performance:** Among new pitchers, are there specific pitches that lead to a higher chance of injury or better performance?

Young pitchers in their early career are malleable in terms of pitching habits and style. Identifying which pitch types will increase risk of injury and which will lead to better performance is critical in preventing premature injuries while helping rookies perform well in their games.

RQ1: Factors contributing to Career Length

To address potential survivorship bias, we noted that the dataset includes veteran pitchers whose debuts occurred prior to the 2010 start date. Thus, their first appearance in our data does not reflect their true rookie season. To account for this, we subset the data to include only pitchers who debuted between 2010 and 2015. The 2015 upper bound ensures that every pitcher has the possibility of achieving a ‘long career’ before the dataset concludes in 2020.

Furthermore, we restricted our analysis to performance data from a pitcher’s first two seasons to predict whether they will have a short or long career. Including the second season helps mitigate the noise of a single outlier rookie year, providing a more stable baseline for predicting long-term career outcomes. The final sample used for analysis consisted of 131 pitchers, down from the original 2,911. We have also grouped the pitch types into general groupings (fastball, breaking, and offspeed) based on official MLB definitions (Major League Baseball, 2025).

Using this subset of data, we created several variables for analysis. A description of the variables can be found in Table 1 below:

Table 1: Variables used for RQ1

| Type | Metric | Description |
|------------------|----------------------|--|
| Physicality | Height | Height of player in inches (in). Taken as the first recorded height when entering dataset. |
| | Weight | Weight of player in pounds (lbs). Averaged across the two seasons. |
| | Total_Days_Injured | Total cumulative number of days injured. |
| Core Performance | K_per_9 | Number of strikeouts per 9 innings. |
| | BB_per_9 | Number of walks given per 9 innings. |
| | HR_per_9 | Number of home runs given per 9 innings. |
| | Avg_Fastball_Overall | Average speed of a fastball over their first 2 seasons. |
| Workload | Avg_Breaking_Overall | Average speed of a breaking ball over their first 2 seasons. |
| | Avg_Ofspeed_Overall | Average speed of an offspeed ball over their first 2 seasons. |
| | Avg_Pitches_Per_Game | Average number of pitches per game over their first 2 seasons. |
| Style of Pitch | Prop_Fastball | Average proportion of fastballs thrown over first 2 seasons. |
| | Prop_Breaking | Average proportion of breaking balls thrown over first 2 seasons. |
| | Prop_Ofspeed | Average proportion of offspeed balls thrown over first 2 seasons. |

Analysis

Univariate T-test

We first performed a univariate t-test to see if there is a statistically significant difference of means between the group of players with a short career (< 5 years) and the group of players with a long career (≥ 5 years).

H_0 : Early-career indicator is the same for short and long career pitchers

H_a : Early-career indicator differs between groups

Table 2: Univariate T-test Results

| Variable | p-value | Mean (Short) | Mean (Long) |
|------------------------------|--------------|--------------|-------------|
| Height | 0.586 | 74.28 | 74.48 |
| Weight | 0.734 | 213.16 | 214.40 |
| Total Days Injured | 0.563 | 4.19 | 5.95 |
| K/9 | 0.287 | 7.00 | 7.57 |
| BB/9 | 0.017 | 4.79 | 3.62 |
| HR/9 | 0.095 | 1.38 | 0.99 |
| Avg Fastball Velocity | 0.001 | 90.98 | 92.43 |
| Avg Breaking Velocity | 0.134 | 80.49 | 81.41 |
| Avg Offspeed Velocity | 0.017 | 83.00 | 84.33 |
| Avg Pitches per Game | 0.054 | 33.76 | 42.90 |
| Prop Fastball | 0.255 | 0.57 | 0.59 |
| Prop Breaking | 0.814 | 0.21 | 0.20 |
| Prop Offspeed | 0.633 | 0.09 | 0.08 |

The results from Table 2 show three statistically significant predictors of long career length ($p < 0.05$). The following were:

- **BB/9**: Pitchers who walk fewer batters have a longer career. Long career pitchers give an average of 1.17 fewer walks overall. This suggests that good pitch command is a key indicator of sustained MLB success.
- **Avg Fastball Velocity**: Higher fastball velocity strongly correlates with career longevity. Long career pitchers' average fastball velocity is 1.45 mph faster than their short career peers. A high speed fastball keeps pitchers competitive.
- **Avg Offspeed Velocity**: Higher offspeed velocity also correlates with longer careers, suggesting both speed and technique are important tools to keep in a pitcher's arsenal.

Multiple Linear Regression

We then regress all of these variables against career length (seniorityYrs) to determine which predictors contribute the most to career longevity. A backward stepwise selection procedure identifies three key predictors of career length:

$$\text{Career Length} \sim \text{Total Outs} + \text{BB/9} + \text{Avg Fastball Overall}$$

Table 3: MLR Summary

| Variable | Estimate | Std. Error | p-value | VIF |
|-----------------------|----------|------------|-----------------------|------|
| Intercept | -44.20 | 9.10 | 3.79×10^{-6} | - |
| Total Outs | 0.00684 | 0.00139 | 2.82×10^{-6} | 1.15 |
| BB/9 | -0.244 | 0.116 | 0.0367 | 1.15 |
| Avg Fastball Velocity | 0.533 | 0.0991 | 3.95×10^{-7} | 1.00 |

Model Fit: $R^2 = 0.383$, Adjusted $R^2 = 0.367$, Residual SE = 2.704, F-statistic $p < 10^{-11}$

Table 3 shows the summary of the MLR after backwards stepwise selection. The final model has an $R^2 = 0.383$, implying that the model captures around 38% of the variation in career length. The R^2 value is quite strong considering only three variables have been selected, no interaction variables have been used, and the response variable is highly complex. In addition, the VIF values suggest no multi-collinearity present within the model.

- **Avg Fastball Velocity** is the strongest positive predictor of career length. This suggests that pitchers with a strong fastball tend to last significantly longer in the league. An increase of 1 mph of average fastball velocity results in a 0.533 increase in career length.
- **BB/9** is negatively associated with career length, albeit has less of an effect than average fastball velocity. This suggests that poor command of the pitch early in a career is a red flag for longevity. An increase of 1 walk given per 9 innings results in a -0.244 decrease in career length.
- **Total Outs** reflects a pitcher's workload and skill; the effect is positive and statistically significant, though the magnitude of the effect is small. This suggests that total early-career outs has a slight positive effect on the length of a career.

Taken together, the regression suggests that **command of the ball**, a **higher average fastball velocity**, and **demonstrated workload** are the most important early indicators of whether a pitcher will have a long MLB career.

The residual diagnostics plots in Figure 3 indicate that the OLS assumptions are largely met. In the Residuals vs Fitted plot, residuals are scattered around 0; however the curvature in the red trend line suggests mild heteroskedasticity, as the variance appears slightly higher for fitted values near the center of the distribution compared to the extremes. The Q-Q plot suggests the residuals are normally distributed, and there are no points beyond the Cook's distance line.

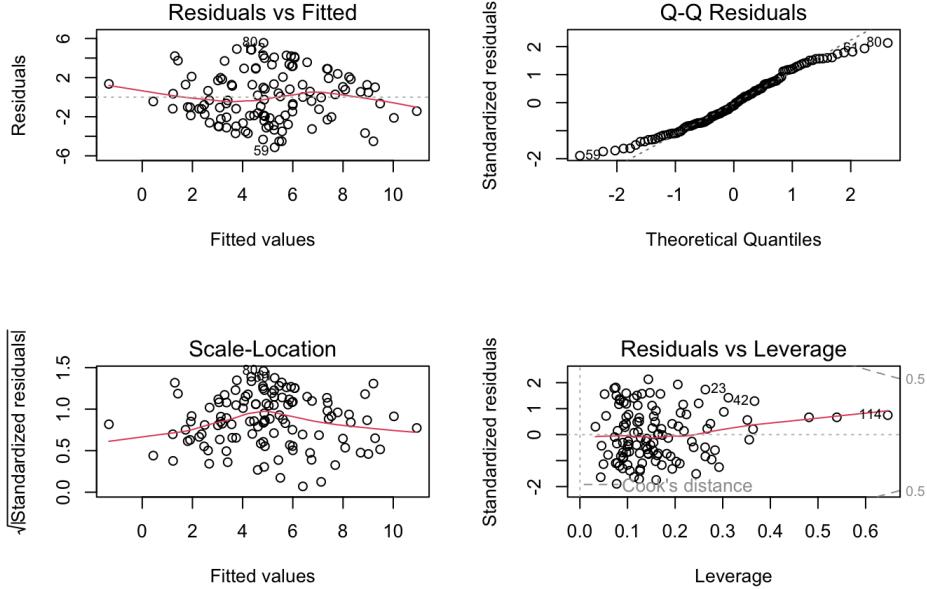


Figure 3: Residual Diagnostics after MLR

RQ2: Pitch types on injury & performance

To evaluate whether pitch selection contributes to injury and performance, we examine pitchers in their **debut season**. Only using the season from new players helps avoid confounding effects from older, established pitchers who naturally have higher injury risk. The final sample used for analysis consisted of 1,423 pitchers, down from the original 2,911.

As physical strain from previous games carries over into subsequent appearances, so does the associated risk of injury. To account for cumulative workload when pitchers appear in multiple games, we constructed a 3-game rolling pitch count; for each observation we summed up the frequency of each pitch type thrown in the current game and the two preceding games, resulting in a cumulative pitch count for every pitch type.

In addition, we computed the number of days between a pitcher's consecutive appearances and treated this value as a pitcher's "rest days". We hypothesized that greater rest days mitigate cumulative physical strain across games. A description of the variables used for RQ2 can be found in Table 4 below.

Table 4: Variables used for RQ2

| Metric | Description |
|-----------------------|--|
| 3 Game Fastball Count | Cumulative number of fastballs thrown in the last 3 games |
| 3 Game Breaking Count | Cumulative number of breaking balls thrown in the last 3 games |
| 3 Game Offspeed Count | Cumulative number of offspeed balls thrown in the last 3 games |
| Rest Days | Number of days in between recorded games |
| y14 | Binary variable; whether the pitcher got injured in the next 14 days after this game |

RQ2a: Pitch types and injury

Model fitting

We first examined whether the number of specific pitch types is more likely to cause injuries.

H_0 : The 3-game rolling pitch count of each pitch type and number of rest days do not affect the chance of a pitcher getting injured.

H_a : There is at least one variable that affects the chance of a pitcher getting injured.

We fit a logistic regression model with 'y14' (whether the pitcher gets injured in the next 14 days after this game) as the response and cumulative pitch types and rest days as the predictors.

Table 5: Logistic regression model summary

| Variable | Estimate | Std. Error | p-value |
|-----------------------|----------|------------|-----------------------|
| Intercept | -5.765 | 0.535 | $< 2 \times 10^{-16}$ |
| 3 Game Fastball Count | 0.0297 | 0.00979 | 0.0024 |
| 3 Game Breaking Count | -0.0210 | 0.0237 | 0.377 |
| 3 Game Offspeed Count | -0.0315 | 0.0332 | 0.343 |
| Rest Days | 0.0188 | 0.0184 | 0.307 |

Model Fit: residual deviance = 254.30, AIC = 264.3.

The logistic model achieves an **AUC of 0.6895** (Figure 4), indicating moderate discriminatory ability.

The model performs noticeably better than random guessing ($AUC = 0.5$), but does not provide strong predictive accuracy, which is expected given the unpredictability and noise of short-term injuries.

Observations and interpretations

From the model estimates in Table 5, fastball frequency is the only significant predictor of short-term injury. In particular, each additional fastball thrown over the past three games increases the log odds of injury by approximately 3%. Conversely, breaking and offspeed pitches show no significant association with injury risk.

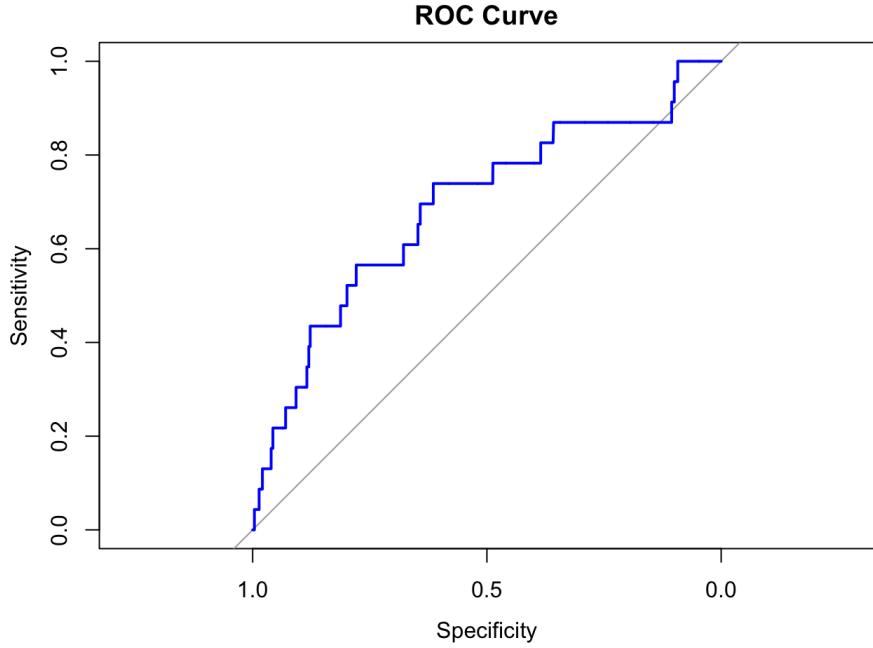


Figure 4: ROC Curve for Logistic Regression

This is expected, as fastballs impose the largest biomechanical load on the arm due to their high force and velocity demands. This explains their strong statistical relationship with injury; even small increases accumulate quickly for players with high fastball usage. In contrast, breaking and offspeed pitches may not register in early-career injuries as they generate lower peak stress or are thrown less frequently by new players.

Surprisingly, the number of rest days did not significantly affect the odds of getting injured. One possible explanation is that younger pitchers may recover quickly, reducing the benefit of longer breaks between appearances. Additionally, rest days are not necessarily periods of inactivity, as players may continue to train and throw, so the recovery gained during these intervals may be insufficient to meaningfully reduce cumulative strain.

In summary, a heavy fastball workload is a clear early warning signal for short-term injury among new MLB pitchers. Coaches should carefully monitor a pitcher's fastball usage to reduce the chance of unwanted injury during the pitcher's early career.

RQ2b: Pitch types and performance

We then investigated how the frequency of each pitch type affects overall performance.

Model fitting

Recall (Figure 2) that the key performance metrics (home runs, strikeouts, outs, and walks) are all discrete count variables. Ordinary linear regression cannot capture the relationship effectively if the response is discrete, therefore we use a Poisson regression instead.

H_0 : The counts of each pitch type do not affect a given key performance metric.

H_a : There is at least one pitch type that affects the key performance metric.

We fit a model in the following form for each key performance metric as the response,

$$\log(E[\text{outcome}]) = \beta_0 + \beta_1 \cdot \text{fastball} + \beta_2 \cdot \text{breaking} + \beta_3 \cdot \text{offspeed}.$$

where the predictors are the number of pitches of a particular type thrown in a game. Each β coefficient represents the change in the log expected count of the outcome for each additional pitch of that type.

Table 6: Summary of Poisson Regression Models for Pitch Count Effects

| Effect of Pitch Counts on K | | | |
|---------------------------------|----------|------------|------------------------|
| Variable | Estimate | Std. Error | p-value |
| Intercept | -0.4185 | 0.0282 | $< 2 \times 10^{-16}$ |
| Fastball Count | 0.0189 | 0.0009346 | $< 2 \times 10^{-16}$ |
| Breaking Count | 0.0284 | 0.0017070 | $< 2 \times 10^{-16}$ |
| Offspeed Count | 0.0195 | 0.0024030 | 5.61×10^{-16} |
| Effect of Pitch Counts on Walks | | | |
| Variable | Estimate | Std. Error | p-value |
| Intercept | -1.2290 | 0.0415 | $< 2 \times 10^{-16}$ |
| Fastball Count | 0.0251 | 0.001324 | $< 2 \times 10^{-16}$ |
| Breaking Count | 0.0210 | 0.002574 | 3.23×10^{-16} |
| Offspeed Count | 0.0137 | 0.003470 | 7.78×10^{-5} |
| Effect of Pitch Counts on HRs | | | |
| Variable | Estimate | Std. Error | p-value |
| Intercept | -2.5381 | 0.0782 | $< 2 \times 10^{-16}$ |
| Fastball Count | 0.0190 | 0.002372 | 1.11×10^{-15} |
| Breaking Count | 0.0325 | 0.004235 | 1.59×10^{-14} |
| Offspeed Count | 0.0386 | 0.005669 | 9.26×10^{-12} |
| Effect of Pitch Counts on Outs | | | |
| Variable | Estimate | Std. Error | p-value |
| Intercept | 0.7911 | 0.0152 | $< 2 \times 10^{-16}$ |
| Fastball Count | 0.0217 | 0.000493 | $< 2 \times 10^{-16}$ |
| Breaking Count | 0.0229 | 0.000936 | $< 2 \times 10^{-16}$ |
| Offspeed Count | 0.0243 | 0.001232 | $< 2 \times 10^{-16}$ |

Observations and Interpretations

All coefficients in all four models from Table 6 are positive and statistically significant ($p < 0.001$), but their magnitudes vary by outcome.

For strikeouts (K), breaking balls have the strongest effect. The coefficient of breaking ball counts is the largest among the three ($\beta = 0.0284$), indicating that breaking pitches are the most effective at generating strikeouts.

For walks (W), fastballs increase walk probability the most. Fastballs have the highest coefficient ($\beta = 0.0251$) in the model, suggesting that fastballs are harder to accurately throw consistently. This is especially the case for rookies.

For home runs (HR), offspeed pitches have the strongest association with giving a home run. Offspeed pitches have the largest coefficient ($\beta = 0.0386$), which aligns with pitching theory; offspeed pitches are designed to disrupt timing, but when poorly executed or anticipated by the batter, they tend to generate slower ball speeds and are easier to drive the ball for home runs.

For outs, all three pitch types have similar effects. The coefficients for outs are nearly identical across pitch types, indicating that no single pitch type dominates in generating outs.

Table 7 summarizes the pitch type that is most strongly associated with each performance metric as well as the reasons for the association.

Table 7: Summary of Poisson Coefficient Interpretations for Each Outcome

| Outcome | Most Associated Pitch Type | Reason / Interpretation |
|---------------|----------------------------|--|
| Strikeout (K) | Breaking Ball | Highest coefficient; hardest pitch type to hit |
| Walk (W) | Fastball | Harder to control at high velocity |
| Home Run (HR) | Offspeed | Easiest to hit if anticipated or poorly executed |
| Out | All Similar | Coefficients nearly equal across pitch types |

Taken together, different pitch types are specialized in different outcomes. Breaking balls are best for strikeouts, fastballs carry a higher walk risk, offspeed pitches pose home-run danger, and outs appear largely pitch-agnostic.

Limitations of the Study

The findings from our study must be comprehended with a few considerations in mind. We will discuss each of them below.

Seniority inconsistencies

In the dataset, use players who have negative ‘seniorityYrs’. These are players who have not officially debuted at the time of the observation, but were playing practice matches in the off-season. Unofficial Major League games could influence pitching statistics because the competitive stakes are lower, potentially reducing pressure and altering pitcher behavior. Despite this, we chose to include them to avoid further subsetting our pool of pitchers used for analysis.

Length of data collected

The last entry of the dataset ends in 2020, so pitchers labeled as “short-career” (fewer than 5 years) may still be active or returning from injury, leading to an underestimation of true longevity. Given the large number of pitchers analyzed in our study, the likelihood of this having a severe impact on our statistical analysis is small.

Injury-list ambiguity

Injury List (IL) placement is not always strictly medical, as baseball teams may use it for strategic rest, roster flexibility, or minor, undocumented issues. In addition, the injury list placement does not reflect the severity of an injury. These may distort the true injury status of a pitcher.

Despite this, even if the team places a pitcher on their injury list for strategic reasons, the pitcher must have gone through a certain degree of physical toil that the team is willing to let the pitcher rest. While our analysis might not capture the severity of actual physical injuries, it estimates the chance of the team withholding a pitcher from future games.

Unobserved confounding physical conditions

The dataset does not record the actual physical conditions of the pitchers, such as training load, health history, and conditioning, all of which contribute to injury risk and long-term development.

In spite of missing these critical measures, our models still show significant and interpretable results, allowing teams to draw insight and make informative decisions in the absence of more complex data.

Conclusion and Recommendations

Based on our analysis and modeling results, we now present the key findings and practical recommendations for teams and coaching staff to support player health and performance.

Injury prevention

As we have found in RQ2a, recent cumulative fastball counts are predictive of pitcher injury; the more a fastball is pitched across recent games, the more likely a pitcher is to get injured.

With that in mind, baseball teams should closely monitor their pitchers’ rolling fastball workload, as sustained stretches of high fastball usage meaningfully increase short-term injury risk. Real-time pitch-mix tracking can help identify emerging fatigue before it becomes harmful. Recovery plans should also be tailored to recent pitch volume rather than relying solely on rest-day counts, which often fail to reflect true physiological stress.

Player development

We have found in RQ1 that a pitcher’s career longevity is associated with lower rates of walks, higher fastball velocity, and total outs achieved in their first two years of their career.

From RQ2b, we found that breaking balls are most associated with strikeouts, fast balls with walks, and offspeed with home runs.

Taking the findings from both research questions together, early-career development should emphasize sharpening command, since lower walk rates consistently predict longer MLB careers. Improving breaking-ball quality offers the greatest upside for generating strikeouts, while refining offspeed control can directly reduce susceptibility to home runs. Together, these adjustments help pitchers balance effectiveness with long-term durability.

Scouting and Evaluation

Our findings apply not only to maintaining and developing existing players, but also to scouting new potential pitchers. Scouts should prioritize identifying prospects with strong fastball quality, as this pitch is the foundation of pitching success and is the most important pitch in the game. A well-executed fastball tends to drive better performance outcomes across strikeouts, weak contact, and overall run prevention, which in turn is strongly associated with longer career longevity. Because pitchers with effective fastballs are more likely to remain competitive at higher levels, scouts should focus on evaluating fastball strength, command, and sustainability.

References

- Carr, J. B., McElheny, K. D., Corrigan, A., Rowe, D., Ma, K., & Curriero, F. C. (2022). The most common type, severity, and expected frequency of injuries vary by defensive position in professional baseball players. *American Journal of Sports Medicine*, 50(9), 2534–2541.
- Coleman, G. (2024, January 5). *The effects of high velocity on players' health, safety and performance*. Professional Baseball Strength & Conditioning Coaches Society. Retrieved December 8, 2025, from <https://pbsccs.org/the-effects-of-high-velocity-on-players-health-safety-and-performance/>
- Major League Baseball. (2025). *Pitch types — glossary*. MLB.com. Retrieved December 8, 2025, from <https://www.mlb.com/glossary/pitch-types>
- Witnauer, W. D., Rogers, R. G., & Saint Onge, J. M. (2007). Major league baseball career length in the twentieth century. *Population Research and Policy Review*, 26(4), 371–386. <https://doi.org/10.1007/s11113-007-9038-5>