STATS 140XP Final Report

Understanding MLB Pitcher Injury Risk

Through Historical Injury Patterns

Joseph Choi, Eric Du, Samson Huynh, Hilary Lin, Neva Williams, Sinan Allahbachayo

Contents

1 Abstract

In this project, our team analyzed Major League Baseball (MLB) pitching data to investigate statistical factors associated with short-term future injury risk. Our analysis centered around a primary predictor widely believed to influence pitcher health which is past injury health, measured using the cumulative number of times a player had previously appeared on the injured list (cum_occur_dl). To conduct this analysis, we merged the MLB performance dataset (xdf_full) with the corresponding injury-history dataset (xfullmx_XY) to create a unified dataset containing both demographic information and historical injury patterns. From this merged dataset, we generated categorical variables for injury history bins, to enable clear comparisons across different pitcher subpopulations and to prepare the data for statistical testing.

With these variables defined, we applied complementary inferential methods to evaluate whether prior injury history influences the likelihood of future injury within 14, 30, and 60 day windows (*y14*, *y30*, *y60*). First, we used chi-square tests of independence to determine whether injury outcomes were distributed differently across the injury history categories. These tests provided an initial assessment of whether the predictors were statistically associated with injury risk. Next, we fitted logistic regression models to quantify the direction and magnitude of these relationships, measuring how strongly increases in past injury counts impacted the probability of a future injury. By comparing the results from both methods, we were able to assess consistency across categorical and model based approaches, strengthening the reliability and interpretability of our findings.

Overall, this combined statistical framework allowed us to form a clear, evidence based understanding of how demographic and historical factors contribute to short-term injury risk among MLB pitchers, offering insight into patterns that are meaningful for both player

evaluation and health management. Taken together, these methods provide a consistent but measured conclusion: while past injury history is statistically linked to short-term injury outcomes, the effect sizes suggest that injury history alone is insufficient for predicting future injuries. This highlights the importance of incorporating additional workload and biomechanical variables when evaluating MLB pitcher health.

2 Introduction

Pitcher injuries are a major issue in Major League Baseball, and they have a huge impact on both players and teams. Since pitching requires explosive, high-velocity movements repeated hundreds of times, pitchers naturally face a higher risk of overuse injuries. Even small changes in workload, fatigue, or recovery can affect whether a pitcher gets hurt. Because of this, teams today rely heavily on data to understand injury patterns and try to prevent them. One question that keeps coming up is whether a pitcher's past injury history can actually help predict if they might get injured again soon.

In this project, we focus on whether a pitcher's previous injury record, specifically how many times they have been on the injured list, is related to their likelihood of being injured again within the next 14, 30, or 60 days. We created a merged dataset that combines MLB performance stats with detailed injury history so we could study how past injuries relate to future ones. This helps us see whether pitchers with longer or more frequent injury histories face noticeably higher short-term injury risk.

To systematically investigate this question, our project is structured around three guiding objectives:

1. To explore the data and understand overall injury patterns across the league.

2. To test whether past injury counts are statistically associated with future injury outcomes.
3. To measure the strength and direction of these relationships using multiple statistical methods, including chi-square tests, logistic regression models, and Tukey post-hoc comparisons.

By combining EDA, hypothesis testing, and model-based methods, our goal is to provide a clear and statistically solid analysis of how historical injury patterns relate to short-term injury risk for MLB pitchers. This helps us better understand whether prior injury history should play a meaningful role in evaluating pitcher health and future risk.

3 Background

Major League baseball (MLB) is one of the most physically demanding professional sports leagues, requiring pitchers to perform high velocity, repetitive throwing motions over extended seasons. Pitching places significant stress on the shoulder and elbow joints making pitchers particularly vulnerable to overuse injuries such as rotator cuff strains and ulnar collateral ligament tears. As a result, pitcher health and workload management have become major concerns for teams seeking to balance performance with long term athlete sustainability.

In recent years, MLB organizations have increasingly relied on data analytics to understand the injury patterns and reduce risk. Metrics such as pitch count, rest days, cumulative workload and historical injury data are now commonly monitored to evaluate pitcher fatigue. Previous research suggested that accumulated stress and insufficient recovery time may increase the probability of injury, emphasizing the importance of monitoring both short term workload and long term injury history.

With the advancement of sports analytics, teams now use data to identify patterns that may predict future injuries.This study explores whether a pitcher's accumulated injury history, measured by the number of times they have appeared on the Injured List, can help predict short-term future injury risk. The goal is to determine whether pitchers with greater history of injury are more likely to experience another injury within a short time frame, contributing to a deeper understanding of how fatigue impacts player health.

4 Methods

4.1 Data

The datasets used were:

- xdf_full – main MLB pitcher performance dataset
- xfullmx-XY – injury dataset containing prior injury history and future injury indicators

4.2 Hypothesis

1. Effect of Past Injury History

   H_0: Past injury history does not influence future injury risk.

   H_a: Past injury history does have influence on future injury risk.

4.3 Variables of Interest

1. The variable that we chose to account for injury history is : *cum_occur_dl* - the total number of times a player was injured throughout their career. We binned the values into three categories to convert the variable to categorical and called this variable *injury_category*. The corresponding levels are:
   - No Injury (0 injuries)
   - One Prior Injury (1 injury)

- Multiple Prior Injuries (1+ injuries)

2. The variables that we chose to account for future injury risk are:
    a. *y14* – Does the pitcher end up on the injured list within the next 14 days? 0 for no, 1 for yes.
    b. *y30* – Does the pitcher end up on the injured list within the next 30 days? 0 for no, 1 for yes.
    c. *y60* – Does the pitcher end up on the injured list within the next 60 days? 0=no, 0 0 for no, 1 for yes.

To evaluate our hypothesis from multiple angles, we applied a combination of categorical tests, predictive modeling, and post-hoc group comparisons.

4.4 Statistical Tests

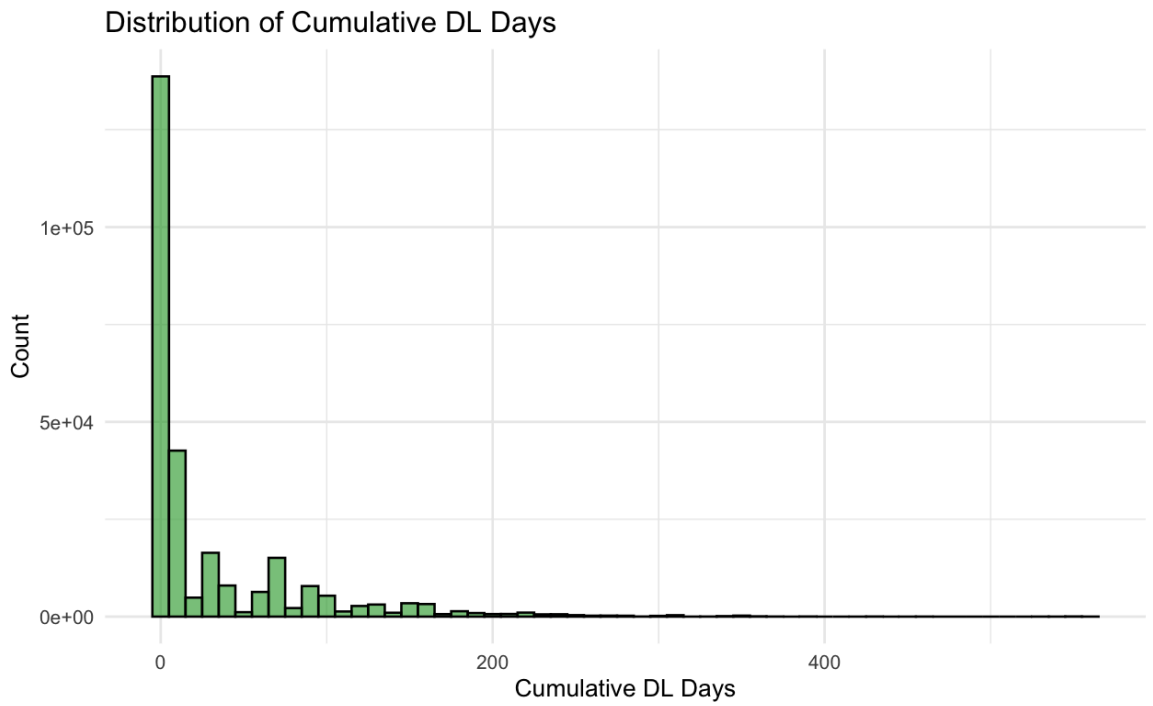
1. *Chi-Square Tests of Independence*
    - Used to test whether categorical predictor (*injury_category)* are associated with categorical outcomes (*y14*, *y30*, *y60*).
2. *Logistic Regression*
    - Used to model the probability of future injury (*y14, y30, y60*) as the response and cum_occur_dl (numerical) as the predictor, providing effect sizes and significance tests to determine how strongly each variable influences injury risk.
3. *Tukey's Post-Hoc Pairwise Comparison*
    - By using logistic regression models with future injury (*y14, y30, y60*) as the outcome variable and *injury_category* as the basis, we will conduct

Tukey-adjusted post-hoc comparison to see whether differences in categories for *injury_category* are significant.

5 Exploratory Data Analysis (EDA)

Injury Category	Mean_y14	Mean_y30	Mean_y60
Never Injured	0.02326168	0.04705696	0.08215138
One Prior Injury	0.03583805	0.06935170	0.11674837
Multiple Prior Injury	0.04682084	0.09017562	0.14897119

In order to visually see if there was an actual difference between the three categories we made for past injury history, we grouped by the categories and took a look at their means at y14, y30, and y60. As we can see from the table, we can see an increase in mean as we go up in levels for *injury_category*. As we saw this ascending pattern, we figured that this variable was worth investigating. Whether or not these differences in group mean are significant will later be tested. This monotonic pattern suggests a potential dose–response relationship, where pitchers with more prior injuries exhibit slightly higher future injury rates. Although this does not imply causality, it provides motivation for deeper statistical testing



The distribution of the cumulative DL days is highly right-skewed with most pitchers having very few prior days on the disabled list and a small number of pitchers accumulating extremely high totals. Such skewness is expected in professional sport datasets where severe or chronic injuries affect only a limited proportion of players. This supports the idea that injury risk is not uniformly distributed and that prior injury burden may meaningfully separate high risk players from the rest. These patterns underscore why categorizing pitchers by prior injury frequency may help identify whether a small subgroup of heavily injured players faces disproportionately higher short-term risk.

6 Results

future_injury	chi_square	df	p_value	cramers_v
y14	942.8272	2	1.852e-205	0.0588
y30	1657.9678	2	0.000e+00	0.078
y60	2470.7549	2	0.000e+00	0.0953

We conducted three chi-square tests of independence to examine whether past injury history (categorized as no injury, one injury, or multiple injuries) is associated with future injury risk at 14, 30, and 60 days. The chi-square tests of independence show that there is a statistically significant association between future injury and past injury history. However, the corresponding Cramér's V values suggest that the strength of this association is weak. To further evaluate these relationships and model them on a continuous scale, we fitted a series of logistic regression models using each future injury outcome (y14, y30, y60) as the dependent variable and the original continuous cum_occur_dl variable as the predictor.

To complement the chi-square analysis, we next modeled injury history as a continuous predictor using a series of logistic regression models.

Logistic Regression Models: y14, y30, y60 ~ cum_occur_dl

model	term	estimate	std.error	statistic	p.value	OR
y14	(Intercept)	-3.61	0.0135	-268	<1e-16	0.027
y14	cum_occur_dl	0.141	0.0041	34.4	<1e-16	1.15
y30	(Intercept)	-2.89	0.0098	-296	<1e-16	0.055
y30	cum_occur_dl	0.139	0.0031	44.9	<1e-16	1.15
y60	(Intercept)	-2.31	0.0077	-301	<1e-16	0.099
y60	cum_occur_dl	0.139	0.0026	54.2	<1e-16	1.15

As shown in the regression table, the effect estimate for cum_occur_dl is quite small, ranging only from approximately 0.139 to 0.141 across the three models. This indicates that

although the predictor is statistically significant, its practical effect on future injury risk is minimal. This aligns with the low Cramér's V values from the chi-square tests: the relationship between past injury history and future injury risk exists, but the strength of the association is weak. Statistically significant does not always imply a strong or meaningful effect size, and in this case the coefficient and Cramér's V both point toward a weak association. In baseball terms, this means that although pitchers with more prior IL stints are statistically more likely to be injured again, the increase in risk per additional past injury is small and unlikely to meaningfully influence roster or workload decisions on its own.

In order to examine the group differences between "Never Injured," "One Prior Injury," and "Multiple Prior Injuries," and to test whether these differences were statistically significant, we performed a Tukey-adjusted post-hoc comparison using estimated marginal means from a logistic regression model. To further examine how these injury-history groups differ from one another, we performed Tukey-adjusted pairwise comparisons. For this analysis, we replaced the continuous predictor cum_occur_dl with the categorical injury_category variable so that we could directly compare injury probabilities between the three groups and formally test the significance of each pairwise difference.

y_14 ~ injury_category

contrast	odds.ratio	SE	df	null	z.ratio	p.value
Multiple Prior Injuries / Never Injured	2.06	0.0495	Inf	1	30.1	0
Multiple Prior Injuries / One Prior Injury	1.32	0.039	Inf	1	9.44	2.7e-14
Never Injured / One Prior Injury	0.641	0.0195	Inf	1	-14.6	0

contrast	odds.ratio	SE	df	null	z.ratio	p.value
Multiple Prior Injuries / Never Injured	2.01	0.0349	Inf	1	40.1	0
Multiple Prior Injuries / One Prior Injury	1.33	0.0288	Inf	1	13.2	0
Never Injured / One Prior Injury	0.663	0.0146	Inf	1	-18.6	0

Table 5

contrast	odds.ratio	SE	df	null	z.ratio	p.value
Multiple Prior Injuries / Never Injured	1.96	0.0267	Inf	1	49	0
Multiple Prior Injuries / One Prior Injury	1.32	0.0228	Inf	1	16.3	0
Never Injured / One Prior Injury	0.677	0.0117	Inf	1	-22.5	0

The previous three tables are the contrast tables obtained from the Tukey-adjusted post-hoc comparisons based on the logistic regression models for *y14, y30*, and *y60*, using *injury_category* as the predictor. The odds ratio column shows us that across all three time windows, players with Multiple Prior Injuries consistently had the highest risk with odds of future injury that were approximately 2 times higher than those who were never injured and approximately 1.3 times higher than those with one prior injury. The p-values in these contrast tables indicate whether the differences between each pair of injury-history groups are statistically significant. As shown in the tables, all p-values are effectively zero ($p < 0.0001$), demonstrating that every pairwise comparison is statistically significant. This confirms that the three groups differ meaningfully from one another in terms of future injury risk. Although these differences are significant, the absolute probabilities remain modest and align with the conclusion that past injury history alone is not a strong standalone predictor.

7 Conclusions and Limitations

7.1 Conclusions

Our analysis produced a nuanced finding regarding MLB pitcher health. The Chi-square tests and Logistic Regression models both rejected the null hypothesis, confirming that a pitcher's cumulative injury history (cum_occur_dl) is a statistically significant predictor of future injury risk across 14, 30, and 60 day windows.

However, this statistical significance does not equate to strong predictive power. While the p-values were extremely low, demonstrating significance, the Cramer's V Values (0.05-0.09) and Odds Ratios (~1.15) reveal that the magnitude of this effect is minimal. In practical terms, while a pitcher with a history of injuries is technically more likely to be injured again than one without, the difference is not large enough to be the sole basis for player evaluation. This suggests that while past health is a factor, it is not the dominant driver of short-term injury risk. Future injury risk is likely stochastic or dependent on other dynamic variables not modeled. Altogether, these findings suggest that pitcher injury risk is multifactorial, and while past injury counts provide some signal, a more accurate risk assessment would require incorporating biomechanics, real-time workload metrics, and recovery patterns.

In addition, we tested other variables such as age that from a logical standpoint would seem to affect the likelihood of future injury. Although significance was shown, even lower Cramer's V Values demonstrated a minimal effect.

7.2 Limitations

To contextualize these findings, several limitations of our study must be considered:

1. Omitted Variable Bias: Our background research highlighted that pitch count, velocity, and rest days are critical factors in determining pitcher fatigue. However our model only accounted for injury history. By excluding workload metrics (e.g number of pitches thrown in the previous game), our model likely missed the immediate mechanical causes of injury.

2. Definition of Injury: The variable cum_occur_dl measures the number of times a player appeared on the Injured List, but does not account for the severity of these injuries. This can obscure the true physical toll an injury takes on a pitcher.
3. Binary Outcome Limitations: We modeled future injury as a binary outcome (Yes/No within X days). This does not capture players who may be playing through minor injuries or those whose performance declines due to fatigue without being formally placed on the Injured List.
4. Sample Bias: Pitchers with frequent injuries may retire early, pitch fewer innings, or be placed in less demanding roles, which can influence how often they appear in the dataset during high-risk periods.