Determining the Effects of Statcast on Injury Prediction for Major League Baseball Pitchers

Kevin Casey

Haverford College

Department of Economics

Advised by: Anne Preston[1]

April 28th, 2022

Abstract

This paper analyzes the relationship between observable statistics and Major League Baseball pitcher injuries before and after the implementation of Statcast. We determine the effect of the introduction of Statcast data, in 2015, on our regression analysis. We find that the inclusion of Statcast variables in our regression equations increased the R^2 values, however, not by a significant margin. Our findings show that Statcast measures can be used to increase the accuracy of pitcher injury prediction. It is also observed that pitcher workload, as measured through innings pitched, is a poor determinant of pitcher injury. Further research can expand on this study through the inclusion of additional exogenous variables.

[1] I would like to thank Anne Preston as well as my parents, Diana and Sean, for their support and assistance throughout this process.

Table of Contents

I. Introduction

Literature examining injury in Major League Baseball (MLB) pitcher's is ever evolving. Several papers have studied MLB pitcher performance and injury. The most common areas injured by MLB pitchers are the throwing elbow and shoulder (Platt et al., 2021). Certain types of injuries occur as a direct result of throwing a baseball, while other injuries may occur from exogenous factors (Fortenbaugh et al., 2009). The determinant factors of throwing injuries are biomechanical efficiency, mobility, anatomy, strength, velocity, workload, and sleep (Fortenbaugh et al., 2009). There are many different types of injuries that can occur at the shoulder and elbow joint. Tommy John Surgery (TJ) involves reconstructing the Ulnar Collateral Ligament (UCL), and is the main surgical procedure at the elbow joint (Marshall et al., 2019). The UCL is the ligament responsible for accepting valgus force during a throw (Wymore et al., 2016). The most prevalent shoulder injury is a SLAP tear, (Superior Labrum Anterior Posterior). Current literature suggests that pitcher workload, as measured through innings pitched, can predict elbow and shoulder injuries (Marshall et al., 2019). Additionally, the greater a pitcher's velocity, the more force placed on their elbow and shoulder joints (Hurd et al., 2012). Shoulder and elbow injuries occur in other sports, however, with much less prevalence. The literature surrounding arm injuries in baseball has evolved since the development of high speed cameras that allow for acute analysis of biomechanical efficiency (Fortenbaugh et al., 2009). Many MLB teams perform proprietary injury assessments on their pitchers using high speed cameras. One reason for this is because pitcher injuries have been on the rise over the past decade (Platt et al., 2021).

This overview of MLB pitcher injuries serves to inform the reader about a niche topic in baseball. Due to the rise in player injuries over the past decade, many MLB teams are searching for

information that will aid in player acquisition and injury assessment. Players who spend time on the disabled list in the MLB are paid for their time away from the field. This cost is covered by the team, or contract insurance, which is up to the team's discretion to purchase (Svrluga 2021). MLB teams are more likely to insure the contract of players who are injury prone than otherwise. Teams that can accurately predict injuries for pitchers have an advantage in player recruitment over teams without this capability. This research analyzes the effect of Statcast data on the ability to predict pitcher injuries before and after Statcast's implementation.

II. Literature Review

The following literature review encompasses research on hiring under uncertainty, asymmetric information, and predicting injury within a baseball context. This review contains sources from economic perspectives as well as sports medicine perspectives. The sport's medicine literature provides additional information regarding mechanisms of injury which have a wide array of confounding factors. The purpose of this literature review is to inform the reader about the labor market for MLB players, as well as the determining factors that MLB teams consider regarding future player injury when hiring players. MLB teams must consider many factors when determining whether to hire a player. They must observe a player's current value, predicted future value, injury history, future injury risk, and the extent to which team player development resources can maximize value for a given player. The presence of imperfect information in the hiring process makes future projections very tedious. Asymmetric information, on either the player or teams behalf, can influence contract negotiations between MLB teams and players. Papers involving hiring under uncertainty are used to explain this phenomenon.

Hamilton (2017) observes the impact that wage transparency had on MLB salaries after 1985. In 1985, wage transparency in MLB was accomplished through the publishing of all player salaries. The author finds that wage transparency ensured complete information on wages, which resulted in increased wages in the MLB. The author notes that the biggest change in wages occurred in the year immediately following the transparency rule.

Hattery (2017) discusses the implications that predictive injury modeling can have on player informed consent and workplace health risks. The author argues that it is necessary for players to be informed about the conclusions that predictions draw on future injury modeling. In other words,

employees have a right to be informed of health risks uncovered during examination prior to hiring. Disclosing organizational predictive modeling can pose a risk to MLB teams because it reduces asymmetric information that can be leveraged in contract negotiations. Hattery (2017) also discusses differing incentives between organizations and players regarding injury risks. The primary issue is about proprietary risk assessment of MLB players performed with team success in mind as opposed to player health.

As shown above, asymmetric information and uncertain labor conditions greatly affect the ability of MLB teams to accurately predict future player value. For this reason, MLB teams spend vast resources attempting to improve the information present in contract negotiations. One of the most influential pieces of information that teams concern themselves with is a player's future injury risk. The following papers discuss the different ways MLB teams attempt to determine injury potential.

Erickson et al. (2021) presents differences in injury rates amongst players who converted to pitching after previously playing different positions in the MLB. The authors' propose that pitchers who convert later in their career are at lower risk of sustaining a pitching related injury, due to lack of accumulated workload. The results show that pitchers who converted were less likely to be injured, and therefore had smaller stints on the disabled list compared to the control group.

Douglas et al. (2019) look at pitcher performance, and rates of player return following SLAP tear reconstruction. The authors' analyze a panel of 232 players who obtained SLAP repairs between the years 2004-2014. Douglas et al. (2019) find that 83.6% of pitchers who underwent SLAP repairs returned to any level of baseball, while only 52.3% of players returned to prior level of performance.

Conte, Camp, and Dines (2016) observe injury trends in the MLB from 1998 to 2015. The authors' report an average of 464 disabled list appearances and 25,186 days missed while on the

disabled list, annually. The average annual cost that the league loses when assigning players to the disabled list is reported to be around $423,000,000, and the total amount disabled players earned over the past 18 seasons is around $7.6 billion.

Chatha et al. (2019) determine the impact of a disabled list rule change in 2011 on injuries and the labor market in the MLB. The 2011 rule change shortened the 10-day disabled list to 7-days. This paper examines how time spent on the disabled list can greatly impact player costs. The authors' report that the annual player cost while on the disabled list decreased from $1 million to roughly $560,000 following the rule change.

Meldau et al. (2020) research the cost incurred by MLB teams after a player undergoes TJ surgery. The authors observed 194 MLB pitchers that had TJ surgery between the years 2004 to 2014. The average number of missed days for players who underwent surgery is around 180 days. The average cost of recovery per player is roughly $2 million, and almost $400 million across the entire MLB. The authors' reported that only 77% of those who underwent TJ surgery returned to the Major League level.

Marshall et al. (2019) determine MLB pitching performance following TJ surgery. Their data consists of 46 MLB pitchers, who had previously had TJ surgery. Marshall et al. (2019) find that 82% of those who had TJ surgery returned to play in the MLB, and that full tears of the UCL showed better return to performance results than partial tear reconstruction.

Wymore et al. (2016) look at performance of players drafted into the MLB, who had already undergone TJ surgery. The authors use panel data with player level observations on 38 pitchers and observe performance metrics and reinjury risk following surgery. The author's report that TJ surgery did not affect the chances of a professional advancement within the minor leagues, and had no significant effect on performance metrics. Additionally, the presence of TJ surgery did predict future

injury risk, with 86.8% of those having previously undergone surgery being reinjured, compared to 64% injury within the previously uninjured control group.

Hamano et al. (2021) researched the relation between hip range of motion, and strength, as it relates to injuries of the shoulder and elbow. The authors use panel data with player level observations on 135 high school pitchers. They collected strength and range of motion data for each pitcher, and found that a lack of adequate hip external rotation on the throwing side of a pitcher is correlated with shoulder and elbow injury.

Hurd et al. (2012) determine the relation between pitch velocity and elbow injuries. The authors' obtained cross sectional data with player level observations from 26 healthy high school pitchers. Each pitcher threw 10 fastballs with biomarkers to track pitching mechanics, and elbow stress. Hurd et al. (2012) utilize a linear regression of peak pitcher velocity on peak elbow adduction forces, and find that greater pitch velocities resulted in greater forces on the elbow ligaments.

Erickson et al. (2016) describe significant variables in predicting MLB pitcher injuries. The authors' find that workload can be predictive of injury in youth pitchers, however, not as effectively in MLB pitchers. Erickson et al. (2016) discuss selection effects that MLB pitchers undergo as they move up the competitive baseball ladder as a potential reason for greater workload capacity. The authors' also discuss the prevalence of anatomical differences, as well as biomechanical inefficiencies related to injury potential in pitchers. For example, differences greater than 5 degrees of total shoulder rotation compared to the non-throwing shoulder resulted in roughly triple the injury risk.

James and Bennison (2020) use Statcast and PitchFX data to create a model that predicts pitcher injury within 2 weeks of a game. The model takes a vector of player data and weighs this data differently based on the model's prediction of injury potential. According to the authors', the variables pitch count (season), release velocity, release position, and pitch count (game) were the most

determinant when predicting injury potential. James and Bennison (2020) speculate that differences in a player's game velocity compared to their previous 5 game average velocity is a strong determinant of injury, as well as differences in release position of their pitches.

Fortenbaugh et al. (2009) observe the influence of biomechanics on performance metrics and injury potential. The authors observe specific parameters within the throwing motion through an observational study. Fortenbaugh et al. (2009) find that mechanical patterns such as low lead leg bracing force, timing of shoulder external rotation, and high levels of shoulder abduction lead to greater injury risk.

This study contributes to the existing literature by observing data on all active MLB pitchers over the past 20 years, and quantifying the improvement in injury predictions that result from new statistics on pitchers available through Statcast.

III. Data

Multiple data sources were collected for analysis of injury potential in MLB Pitchers. The data being analyzed represents panel data from 4 unique sources (Wooldridge, 2013, pp. 9). The 4 Databases used to create the dataset were BaseballSavant.mlb.com provided through MLB, prosportstransactions.com provided through Frank Marousek, injury data provided through Jon Roegele, Player Salary Database provided through Jeff Euston, as well as additional salary data provided though Doug Pappas (Pappas 2005, Roegele 2021, Marousek 2021, Euston 2021, BaseballSavant 2021). The data consists of every pitcher who pitched more than 9 innings in the MLB between 2000 and 2021. Dropping observations who pitched less than 9 innings in a season is performed to exclude unused pitchers, and any potential position players who may have had pitching appearances. Baseball Savant contains all MLB pitcher data used. The Statcast data only became available beginning in 2015. Conventional pitching statistics used are innings pitched, strikeout percentage, walk percentage, total pitches (career), age, and lagged innings pitched. There are 9376 unique pitcher-year observations. The Statcast database is important for my analysis because it gives more precise information about specific physical characteristics of MLB pitchers such as their velocity, pitch types, pitch movement profiles, and spin rates. The sample size for our Statcast data is 2,953. Changes in these advanced metrics may provide insight into potential future injury.

Figure 1 shows the frequency of pitcher injuries by year. Inj_Year is a variable that is equal to 0 if there is no injury in a given year, and 1 if there is an injury. Figure 1 shows a clear rising trend in injuries since the year 2000. There is a dip in injuries for the year 2020, which can be explained by the shortened 60-game season. 2021 marked the year with the highest number of pitcher injuries in baseball to date, with nearly 250. 3,216 pitchers experienced at least one stint on the disabled list over

our 20 year sample. Of those 3,216 pitchers, 1,755 were injured due to an arm related issue. Therefore, 54.5% of all pitcher injuries are related to the arm.

Figure 1. Frequency of Injury by Year

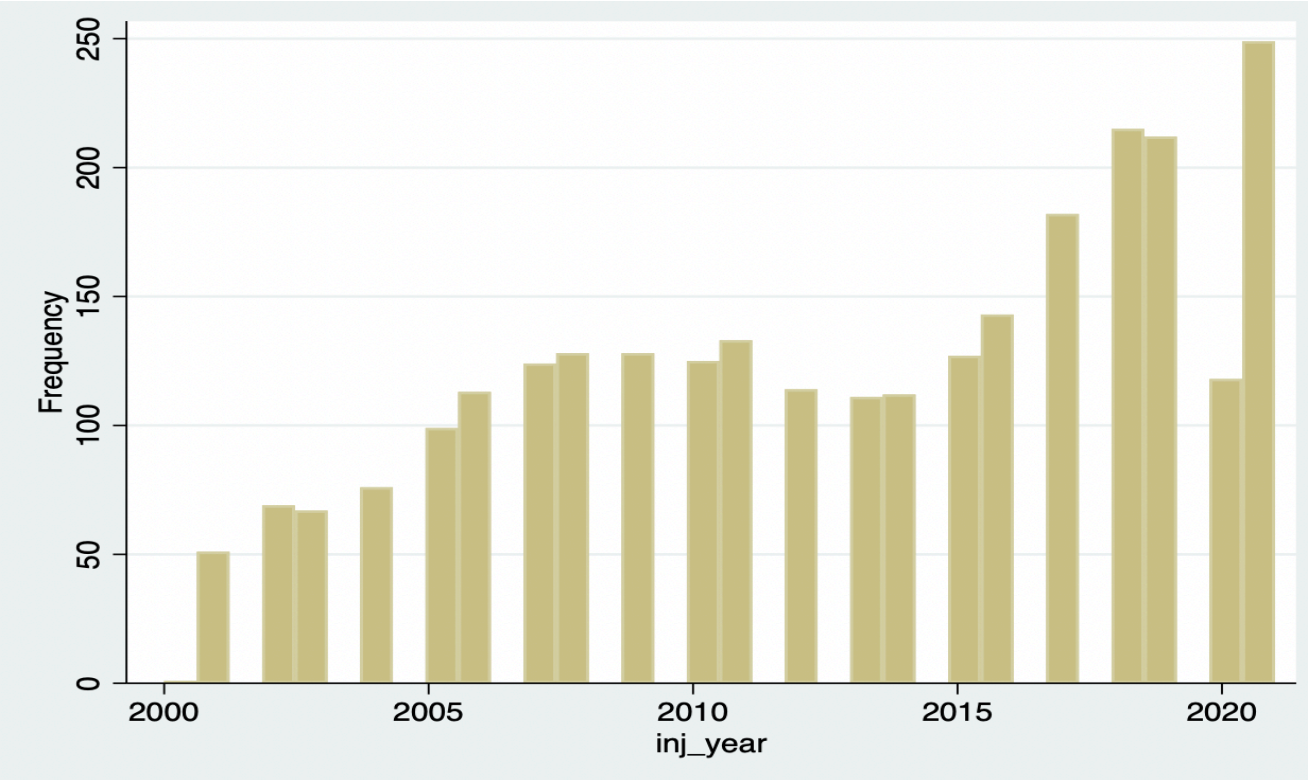


Table 1 displays summary statistics for the data. The summary statistics are split between conventional pitching data, Statcast data, and injury data. The conventional and injury data range from 2000 to 2021, while the Statcast data ranges from 2015 to 2021. It is also important to note the large variability as demonstrated by the standard deviations for variables such as Salary, Total Innings Pitched in Career, Total Pitches thrown in Career, and Innings Pitched. The reason for the high standard deviations for these variables is due to the difference between relievers and starting pitchers in terms of career innings, pitches, and seasonal values for innings pitched. The variation seen in

salary is due to many players earning the league minimum salary, while a select few “superstar” players earn enormously high salaries. The highest annual salary in the dataset is $39,300,000.

The Statcast data contains different numbers of observations for different sets of variables. The reason for this is because the Statcast data gives pitch specific data on pitches such as Sliders, Curveballs, and Fastballs. The Delta (Δ) Fastball Velocity and Spin variables give the change in velocity and spin for a given pitcher from the previous year. This is important because the highest gain in average Fastball velocity is 5.5mph and the biggest loss is 6.1mph. The highest gain for average Fastball spin is 360rpm’s, while the biggest loss is 428 rpm’s. The variable Bauer units take the average fastball spin rate in revolutions per minute (RPM’s) and divide it by the average fastball velocity. This is done to normalize the spin factor given that spin increases with velocity.

The Injury Data has some important notes about it. Primarily, the ranges of the variables. For example, the variables regarding the disabled list range from 0 to 140 days. Another important observation within the Injury Data is the range of the variables Elbow and Shoulder Reinjury. These variables range from 0-4, with 0 representing no reinjury, and 1-4 representing how many reinjuries occurred within the given season. The average time spent on the disabled list over the 21 year period is 6.89 days, and 3.85 days due to arm injury. Lastly, the average salary earned while on the disabled list is $127,361, with the maximum salary earned on the DL being $15,700,000. This is very important because the MLB leaves contractual insurance up to the owner's discretion, meaning that not all injured player salaries are protected by insurance (Svrluga 2021).

Table 1. Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Max
Data (2000-2021):					
Year	9,376	2012.15	5.99	2000	2021
Age	9,376	29.06	4.15	19	50
IP	9,376	81.15	61.45	9	266
Strikeouts	9,376	67.62	52.22	2	372
Walks	9,376	28.4	20.1	0	119
Strikeout Percentage	9,376	19.99	6.31	3.2	53
Walk Percentage	9,376	8.93	3.16	0	30.6
Batting Average Against	9,376	0.256	0.044	0.065	0.458
Total Balls Thrown (Season)	9,376	467.58	347.23	25	1560
Total Pitches Thrown (Season)	9,376	1310.8	968.93	48	4078
Salary	9,376	2,701,907	4,336,421	10,600	39,300,000
Total IP (Career)	9,376	384.21	449.6	9	3393.5
Total Pitches (Career)	9,376	5297.1	6981.75	48	53,380
Delta Innings Pitched	9,375	7.82	52.62	-220.1	204.2
Statcast Data (2015-2021):					
Total Strikes Thrown (Season)	3,869	691.46	540	80	2455
Slider Usage Percentage	2,953	23.98	13.63	0	90.9
Slider Average Speed	2,953	84.44	3.1	70	96.5
Slider Average Spin	2,953	2324	267.84	867	3295
Curveball Usage Percentage	2,529	14.82	10.66	0	67.5
Curveball Average Speed	2,529	78.47	3.51	50	92.5
Curveball Average Spin	2,527	2419.57	304	784	3349
Fastball Usage Percentage	3,869	60.45	12.57	7.5	100
Fastball Average Speed	3,869	92.7	2.71	80.3	101.1
Fastball Average Spin	3,869	2235.1	159.9	1593	2889
Bauer Units	3,869	24.1	1.66	16.65	30.96
Delta Fastball Velocity	2,517	-0.17	1.081	-6.1	5.5
Delta Fastball Spin	2,517	3.86	69.97	-428	360
Injury Data (2000-2021):					
Injury Dummy	9,376	0.325	0.468	0	1
Total DL Time (Season)	9,376	6.89	14.43	0	140
Days on DL due to Arm Inj	9,376	3.85	11.65	0	140
DL Salary Earned	9,376	127,361	537,228	0	15,700,000
Salary Earned on DL for Arm Inj	9,376	67,401	420,755	0	14,600,000
Elbow Reinjury	9,376	0.108	0.386	0	4
Shoulder Reinjury	9,376	0.11	0.37	0	4

Sources: Baseball Savant. (2021). MLB Player Pitching Statistics, 2015-2021. Accessed 2021-12-01.
Baseball Savant. (2014). MLB Player Pitching Statistics, 2000-2014. Accessed 2021-11-15.
Marousek, Frank. (2021). Baseball Transactions Injury Database, 2000-2021. Accessed 2021-12-01.
Pappas, Doug. (2005). SABR Business of Baseball Committee. Player Salary Database, 1985-2004. Accessed 2021-11-20.
Roegele, Jon. (2021). Tommy John Surgery List, 1974-2021. Accessed 2021-11-25.

IV. Methodology

I am asking the question: does the more sophisticated Statcast data improve predictions about arm injuries? The process for analyzing my data will be to run OLS regression equations aimed at providing comparative analysis between estimates before Statcast implementation and after. The first and second regression equations utilize data from after the 2014 season. The third regression equation uses the whole sample. The main dependent variable for these 3 equations is days on disabled list due to arm injury.

- $\mathbf{Y_i = \beta_0 + \beta_1 LaggedInningsPitched_i + \beta_2 TotalCareerPitches_i + \beta_3 Age_i + \beta_4 StrikeoutPercentage_i + B_5 WalkPercentage_i + \epsilon_i}$ **(1)**

- $\mathbf{Y_i = \beta_0 + \beta_1 LaggedInningsPitched_i + \beta_2 TotalCareerPitches_i + \beta_3 Age_i + \beta_4 StrikeoutPercentage_i + B_5 WalkPercentage_i + B_6 \Delta FastballVelocity_i + B_7 FastballAverageBreak_i + B_8 BauerUnits_i + B_9 \Delta FastballSpin_i + B_{10} FastballUsagePercent_i + \epsilon_i}$ **(2)**

- $\mathbf{Y_i = \beta_0 + \beta_1 LaggedInningsPitched_i + \beta_2 TotalCareerPitches_i + \beta_3 Age_i + \beta_4 StrikeoutPercentage_i + B_5 WalkPercentage_i + \epsilon_i}$ **(3)**

Furthermore, these equations are re-estimated using the same independent variables and different dependent variables: probability of an arm injury and probability of arm reinjuries. For each independent variable the first and second regression contain the same sample size. The second regression equation includes Statcast variables, and the third regression equation does not include Statcast variables but utilizes the full sample for the data. The independent variables in both equations are included because they are believed to influence likelihood of injury. ΔFastball Velocity and

ΔFastball Spin represent the difference in the parameters from the previous year. Lagged Innings Pitched, and Total Career Pitches serve to control for workload differences between years for each pitcher. BauerUnits refer to the spin of a fastball divided by its velocity. Age refers to the player's age in the given season. Fastball Average Break measures the break of a pitcher's fastball. In estimating equation 1 and 2 for the 2015-2020 period, I am determining whether inclusion of the advanced statistics increases the amount of variation in days on the disables list or probability of injury I can explain with the regression.

V. Empirical Results

In Table 2, Equation 1 regresses basic pitching statistics on Post-2014 MLB Pitchers, and equation 2 also regresses on Post-2014 MLB Pitchers, but includes the addition of Statcast variables. Equation 3 regresses basic pitching statistics on the complete sample of pitchers. For Table 2, the R^2 for our initial regression equation 1 is 0.0017, and following inclusion of Statcast variables in equation 2, the R^2 increases to 0.0092, given equal sample sizes. Regression results in Table 2 show significant effects in equation 2 for independent variables ΔFastballVelocity and Fastball Average Break at the 5% and 10% levels, respectively. Equation 3 shows significant effects for Total Career Pitches and Walk Percentage at the 10% level. The R^2 value increased from 0.0017 to 0.0092 following introduction of the Statcast independent variables in regression equation 2. The negative coefficients for ΔFastball Velocity mean that positive increases in these variables from the previous year represent less time spent on the disbaled list due to an arm injury. In other words, If a pitcher throws more innings in the current year than the previous year, or has experienced an increase in average fastball velocity, then they are less likely to spend time on the disbaled list. We must be cautious that these results do not display reverse causality. It might be the case that these variables do not precede injury, and that injury itself is what determines these variables. It is important to mention the small R^2 values seen in our regressions. The reason for this, I believe, is because there are many exogenous factors that may influence a players likelihood of sustaining an injury. The small reported R^2 values mean that statistical inference in predicting pitcher injuries does not provide a strong basis from which to make substantial claims.

Table 2: Determinants of Days on the Disabled List

	Days on DL		Days on DL		Days on DL	
Lagged Innings Pitched	0.00049	(0.00805)	0.00156	(0.0081)	-0.00262	(0.00245)
Total Pitches (Career)	-0.0000144	(0.0000662)	-0.0000237	(0.0000662)	0.0000734**	(0.0000237)
Age	-0.124	(0.119)	-0.180	(0.120)	0.012	(0.0335)
Strikeout Percentage	-0.0618	(-0.054)	-0.041	(0.057)	0.031	(0.0194)
Walk Percentage	0.0804	(-0.112)	0.614	(0.112)	0.119**	(0.04)
Delta Fastball Velo			-0.683*	(0.339)		
Fastball Average Break			-0.374**	(0.133)		
Bauer Units (Fastball Spin/Velocity)			0.309	(0.22)		
Delta Fastball Spin			-0.0058	-0.0053		
Fastball Usage %			-0.0174	-0.0268		
Constant	10.35**	(3.88)	11.84	(6.88)	1.931	(1.18)
N	2517		2517		9376	
R-sq	0.0017		0.0092		0.0024	

Standard errors in parentheses

Key: * p < 0.05, ** p < 0.01, *** p < 0.001

Sources: Baseball Savant.(2021). MLB Player Pitching Statistics, 2015-2021. Accessed 2021-12-01.

Marousek, Frank. (2021). Baseball Transactions Injury Database, 2000-2021. Accessed 2021-12-01.

Pappas, Doug. (2005). SABR Business of Baseball Committee. Player Salary Database, 1985-2004. Accessed 2021-11-20.

Roegele, Jon. (2021). Tommy John Surgery List, 1974-2021. Accessed 2021-11-25.

Table 3 uses the dependent variable Inj_Dummy, which signifies 1 if a pitcher has been injured in a given year, and 0 if a pitcher has not sustained a disabled list designation. Table 4 uses a Post-2014 pitcher sample, and dependent variables Elbow Reinjury and Shoulder Reinjury. Elbow Reinjury and Shoulder Reinjury are variables that represent 0 if a player does not experience reinjury in a given season, and 1 if there is reinjury. In Table 3, the R^2 in equation 1 is 0.0089, and increases to 0.0184 in equation 2. Table 3 shows significant effects in equation 1 for Total Career pitches at the 10% level. Additionally, equation 2 shows significant effects at the 10% level for ΔFastballspin

Average Break, ΔFastballVelocity, and Total Career pitches. Additionally, Total Career Pitches and Walk percentage were significant determinants in the full sample regression at the 1% level, as well as Age at the 5% level.

Table 3: Determinants of Injury Dummy

	Inj_Dummy		Inj_Dummy		Inj_Dummy	
Lagged Innings Pitched	0.00028	(0.000234)	0.000328	(0.000235)	-0.000141	(0.0000975)
Total Pitches (Career)	0.0000049*	(0.00000192)	0.00000456*	(0.00000192)	0.00000807***	(0.00000094)
Age	-0.00074	(0.00347)	-0.00255	(0.00349)	0.004**	(0.00133)
Strikeout Percentage	-0.00071	(0.0016)	-0.00024	(0.0016)	0.0013	(0.00077)
Walk Percentage	0.003	(0.0033)	0.0021	(0.0033)	0.008***	(0.0016)
Delta Fastball Velo			-0.02*	(0.0098)		
Fastball Average Break			-0.00808*	(0.00386)		
Bauer Units (Fastball Spin/Velocity)			0.0116	(0.00638)		
Delta Fastball Spin			-0.000394*	(0.000153)		
Fastball Usage %			-0.00118	(0.000779)		
Constant	0.408***	(0.113)	0.387	(0.199)	0.076	(0.0469)
N	2517		2517		9376	
R-sq	0.0089		0.0184		0.0176	

Standard errors in parentheses

Key: * p < 0.05, ** p < 0.01, *** p < 0.001

Sources: Baseball Savant.(2021). MLB Player Pitching Statistics, 2015-2021. Accessed 2021-12-01.

Marousek, Frank. (2021). Baseball Transactions Injury Database, 2000-2021. Accessed 2021-12-01.

Pappas, Doug. (2005). SABR Business of Baseball Committee. Player Salary Database, 1985-2004. Accessed 2021-11-20.

Roegele, Jon. (2021). Tommy John Surgery List, 1974-2021. Accessed 2021-11-25.

Table 4 only contains significant results for ΔFastballVelocity at the 10% level in equations 2 and 4. Table 4 shows similar increases in the R^2 following the inclusion of Statcast variables.

Table 4: Determinants of Elbow and Shoulder Reinjury

	Elbow Reinjury		Elbow Reinjury		Shoulder Reinjury		Shoulder Reinjury	
Lagged Innings Pitched	0.00022	(0.000234)	0.00021	(0.000236)	0.000168	(0.000214)	0.000152	(0.00021)
Total Pitches (Career)	0.000000514	(0.0000019)	0.000000595	(0.00000193)	0.00000074	(0.00000176)	0.00000071	(0.0000018)
Age	-0.00234	(0.00347)	-0.0035	(0.0035)	-0.002	(0.00317)	-0.0027	(0.0032)
Strikeout Percentage	0.002	(0.00156)	0.0025	(0.0016)	-0.0013	(0.0014)	-0.00073	(0.0015)
Walk Percentage	-0.0009	(0.0033)	-0.00154	(0.0033)	0.0044	(0.003)	0.004	(0.003)
Delta Fastball Velo			-0.0238*	(0.0099)			-0.019*	(0.009)
Fastball Average Break			-0.0039	(0.00388)			-0.0035	(0.0035)
Bauer Units (Fastball Spin/Velocity)			0.0052	(0.0064)			0.00065	(0.0058)
Delta Fastball Spin			-0.000035	0.000153			0.0000305	(0.00014)
Fastball Usage %			-0.000459	0.00078			-0.00048	(0.000714)
Constant	0.176	(0.113)	0.175	(0.2)	0.18	(0.103)	0.265	(0.183)
N	2517		2517		2517		2517	
R-sq	0.0014		0.0052		0.0018		0.0043	

Standard errors in parentheses

Key: * p < 0.05, ** p < 0.01, *** p < 0.001

Sources: Baseball Savant.(2021). MLB Player Pitching Statistics, 2015-2021. Accessed 2021-12-01.

Marousek, Frank. (2021). Baseball Transactions Injury Database, 2000-2021. Accessed 2021-12-01.

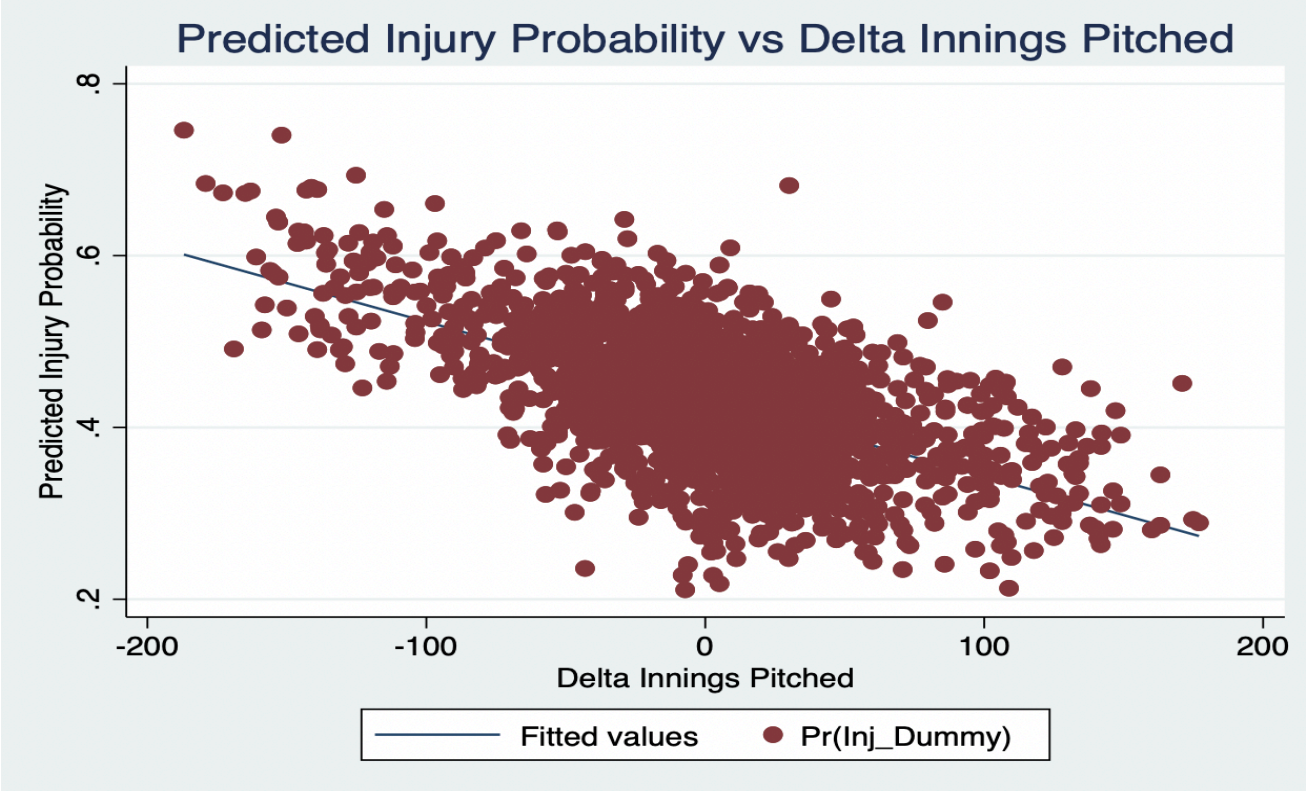
Pappas, Doug. (2005). SABR Business of Baseball Committee. Player Salary Database, 1985-2004. Accessed 2021-11-20.

Roegele, Jon. (2021). Tommy John Surgery List, 1974-2021. Accessed 2021-11-25.

Figure 2 shows predicted injury probability against Δinnings pitched. There is an observed negative correlation between Δinnings pitched and predicted injury probability. Intuitively, pitchers who experience an increase in innings pitched from the previous year are less likely to experience injury in that given year. This result could be due to pitchers who are not healthy enough to

experience a gain in innings pitched and would likely be on the disbaled list to begin with. Alternatively, if a pitcher experiences a gain in innings pitched from the previous year it is likely that they are more healthy and thus enabled to throw more innings. Figure 2 shows a negative correlation between the model's predicted value for injury against Δinnings pitched. This intuitively makes sense because if a pitcher has a positive value for Δinnings pitched then they threw more innings in a given year than the previous year. If a pitcher spent significant time on the disabled list in the previous year, and returned healthy the following year then they will likely have a high value for Δinnings pitched, although pitcher type (starter, closer, reliever) plays a role in determining this value.

Figure 2. Predicted Injury Probability vs ΔInnings Pitched



The same intuition can be used to interpret Figure 3 which shows predicted injury probability against Δfastball velocity. Pitchers who gain velocity on their average fastball from the previous year

are shown to have lower predicted injury probabilities. Selection effects are likely at play, in that pitchers who experience gains in velocity are likely experiencing gains in velocity due to increased arm health, or conversely a gain in average fastball velocity could be the result of more efficient biomechanics which result in a healthier pitcher. Figure 3 plots Δfastball velo against predicted injury probability and also shows a negative correlation. Therefore, if a pitcher gained velocity from the prior year it is more likely that they will not end up on the disabled list. This appears to go against the research from Hurt et al. which showed higher values for fastball velocity represent higher force placed on elbow and shoulder joints. An intuitive explanation for this negative correlation is that in order to increase fastball velocity from one year to the next, a pitcher usually must either become physically stronger, or produce more efficient throwing mechanics which can likely decrease injury potential. It should also be noted that decreases in pitcher velocity from year to year can represent a pitcher who is likely to be injured and is showing signs of arm fatigue via drops in velocity.

Figure 3. Predicted Injury Probability vs ΔFastball Velocity

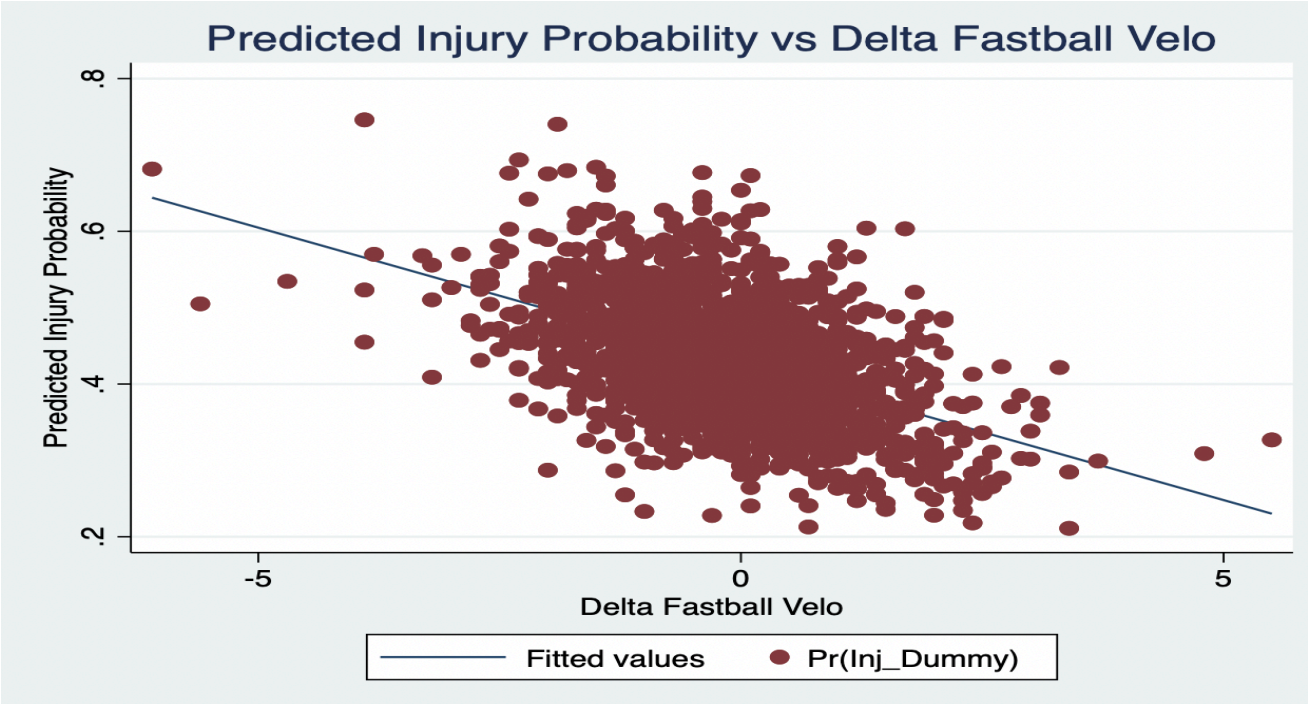


Figure 4 shows the frequency of predicted injury probability based on regression 1 in table 3. Figure 4 shows that the average predicted injury probability for a pitcher given our data is just about 0.3. In other words, there is roughly a 30% chance of any random pitcher in the dataset sustaining any injury that results in disbaled list participation.

Figure 4. Frequency of Predicted Injury (All Data)

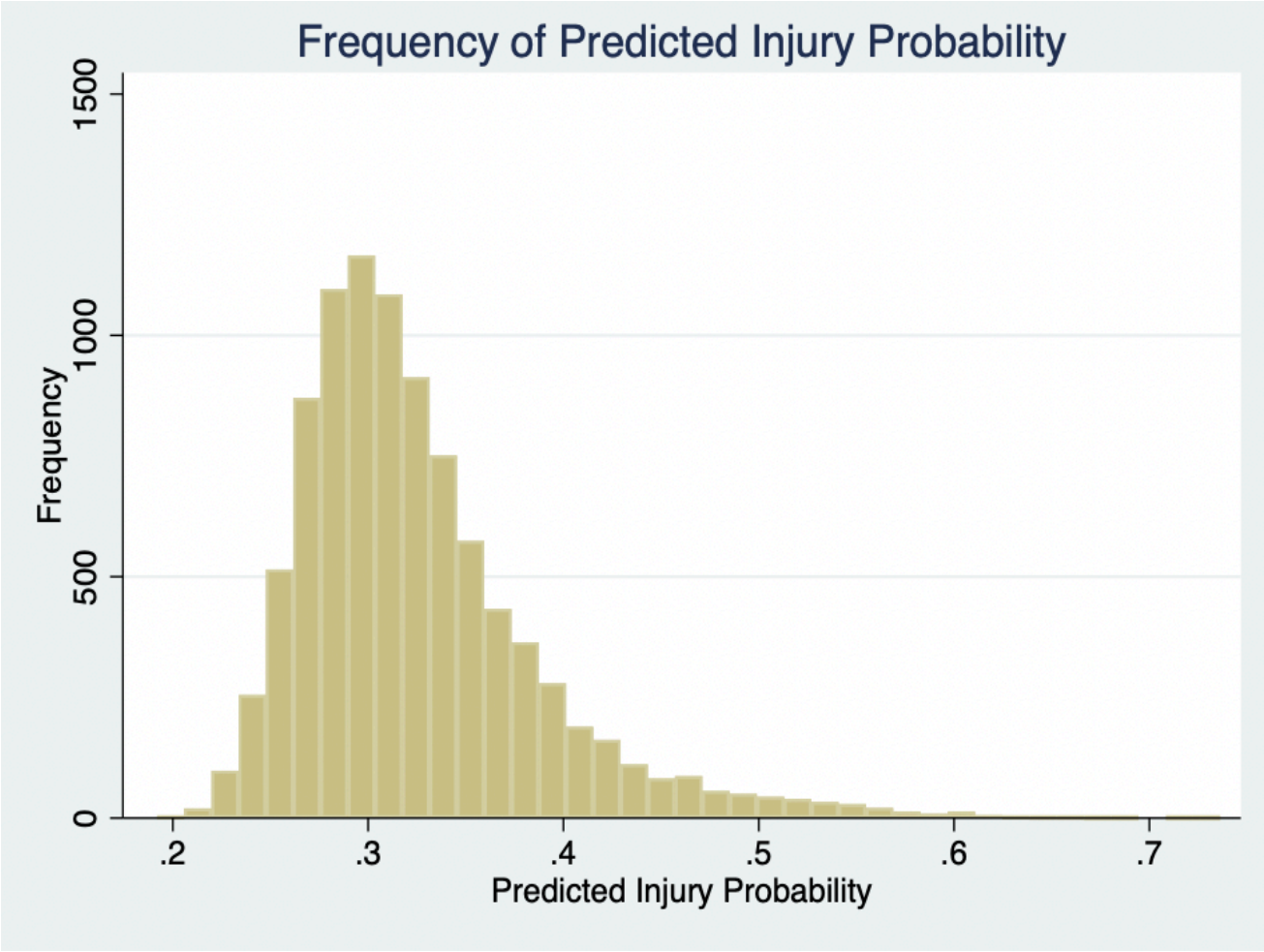
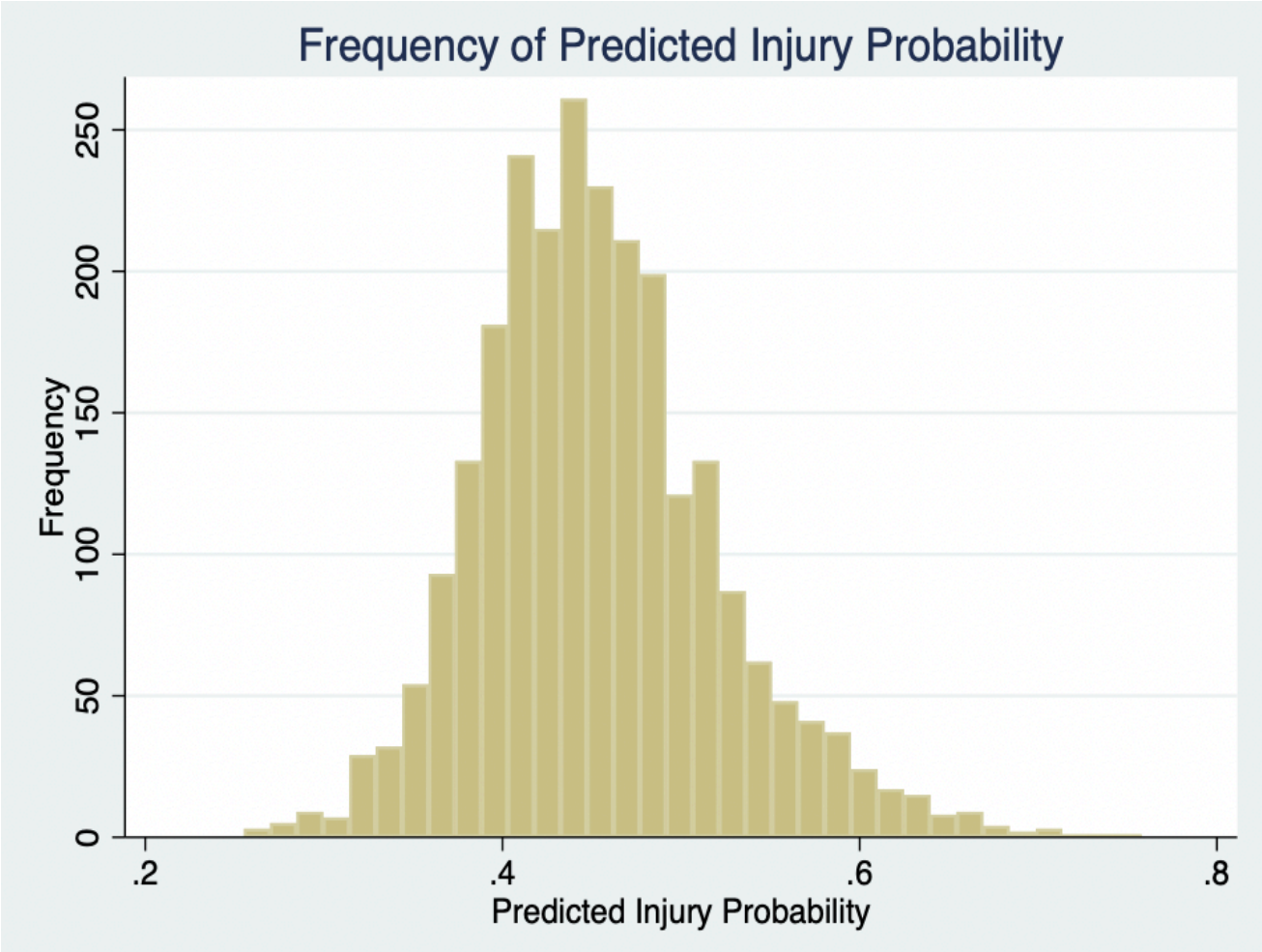


Figure 5 displays the frequency of predicted injury probability based on regression 2 in table 3. There is a noticeable shift in the predicted injury probabilities when Statcast variables are included. It is important to note that the values shown are what our model predicts a players injury probability

is based on our independent variables included. It should also be considered that there is no distinction between minor and major injuries, although only injuries affecting the arm are included. Furthermore, predicted injury probability may be high based on our models predictions, however, they may not sustain any major injuries that jeopardize their career.

Figure 5. Frequency of Predicted Injury (Statcast Data)



VI. Conclusion

This paper utilized a conglomeration of data from multiple sources to analyze the effect of the implementation of Statcast data on predicting pitcher injuries. As seen from figures 4 and 5, it is clear that the Statcast data has better predictive ability for pitcher injury potential than data before 2015. This analysis can be used by MLB teams for a multitude of reasons. Teams can use this to monitor current player injury potential, and cross reference with how a player describes his pain and fatigue levels. This data can also be used for player acquisition analysis, which can allow teams to use proprietary data collected from their minor league ballparks to aid in player trades and signing. Additionally, teams can use this information to influence contract negotiations if players show high likelihoods for injury.

The analytics revolution in the MLB has only recently begun, however is shaping the landscape of the game. An important consideration when analyzing MLB injuries is that the MLB does not provide mandatory insurance policies on player contracts, and therefore teams must seek out contract insurance from outside sources (Svrluga 2021). If a team decides to seek insurance on a high-paying contract, then the company that insures the contract will have to negotiate prices and assess player injury risk (Svrluga 2021). The insurance premium on fragile pitcher contracts can be upwards of 7% of the annual contract value. Most MLB insurance contracts require injured players to remain on the 40-man roster, and miss a predetermined number of days due to injury, in order for the policy to pay out. If an insurance policy does pay out, the insurer usually only covers up to 80% of the total contract value (Svrluga 2021). Therefore, accurate injury assessment of MLB pitchers can be valuable to both insurance agencies and MLB teams when engaging in negotiations.

The main findings of this paper contrast with the sports medicine literature that suggests pitcher workload as a significant determinant of pitcher injury. The workload findings were presented by Fortenbaugh et al., (2009), Marshall et al., (2019), and Erickson et al., (2016). An anecdote that contrasts the pitcher workload hypothesis is Japan's premier high school baseball tournament "Koshien." Koshien is a two week long single elimination tournament (Allen 2019). It is not uncommon, during the tournament, for Japanese high school pitchers to throw upwards of 500 pitchers in a week (Allen 2019). In 2018, Kosei Yoshida threw 881 pitchers over the span of 6 games. In Major League Baseball, it is uncommon for Starting Pitchers to throw more than 130 pitches in a single game, and also uncommon for them to take the ball again without a full 5 days of rest. In Table 1, the maximum value for pitches thrown in an MLB season is at just over 4,000 pitches. Koshien represents a situation where there is a workload demand that exceeds MLB standards. In this culture, it is viewed as an honor to take the baseball and represent your high school, however there have been concerns voiced over the potential damage done to these young players as a result of lack of pitch count oversight. Pitchers who are not able to safely make this amount of pitches are weeded out of contention before the tournament even begins. In other words, selection effects are likely the reason that there is not a significant increase in pitcher injuries as a result of the tournament (Allen 2019.)

The ability of Statcast to predict injury grows with each season, as more data is collected, and more analysis can be done. Our results show that ΔFastballVelocity, and Total career pitches are the highest correlated variables with days missed due to arm injury. Therefore, we can conclude that the implementation of Statcast does increase our ability to predict pitcher injuries, however, not by a significant margin. It is also entirely possible that there are exogenous variables not mentioned within this study that significantly contribute to injury probability. It is more likely that exogenous variables have more influence on potential injury than observable statistical outcomes. This leads me to reject

the notion that pitcher workload is solely responsible for the increase in pitcher injuries shown in Figure 1.

Another important consideration in this study is the existence of large sample sizes. The presence of large sample sizes produces relatively small R^2 values within our regressions. The reason for this is because there are many pitchers within our data who do not experience injury. For example, of the 9,376 pitcher-years in our sample, 3,216 pitchers experienced any form of injury and 1,755 were injured due to an arm related issue. Therefore roughly 54.5% of pitcher injuries affected the arm. The sample size is important when comparing the results of this study to the results presented in the related literature. Many of the related literature results were performed by sports medicine professionals who utilized small sample sizes and thus found more robust results. Future research should aim to include exogenous data points such as caloric intake, sleep, height, weight, degrees of motion in affected joints, as well as strength metrics such as grip strength and shoulder strength. It is entirely possible that one, or some, of these variables can better predict injuries than Statcast data.

VII. Reference List

Allen, J. (2019, December 3). *Baseball: High school pitch limits tip of Japan's injury iceberg*. Kyodo News+. Retrieved April 25, 2022, from https://english.kyodonews.net/news/2019/12/62bc917294f4-focus-baseball-high-school-pitch-limits-tip-of-japans-injury-iceberg.html

Chatha, K., Al-Mansoori, A., Guo, E., Whaley, J. D., & Sabesan, V. J. (2019). Impact of the 7-Day Disabled List Rule Change on Economics and Performance After Reported Concussion Injuries in Major League Baseball. *Orthopaedic journal of sports medicine*, *7*(2), 2325967119825502.

Conte, S., Camp, C. L., & Dines, J. S. (2016). Injury Trends in Major League Baseball Over 18 Seasons: 1998-2015. *American journal of orthopedics (Belle Mead, N.J.)*, *45*(3), 116–123.

Cunningham, S. (2021). *Causal inference: the mixtape*. Yale University Press. Pp. 38-185

Douglas, L., Whitaker, J., Nyland, J., Smith, P., Chillemi, F., Ostrander, R., & Andrews, J. (2019) "Return to Play and Performance Perceptions of Baseball Players After Isolated SLAP Tear Repair," *Orthopaedic journal of sports medicine*, *7*(3).

Erickson, B. J., Chalmers, P. N., Bush-Jospeh, C. A., & Romeo, A. A. (2016). Predicting and Preventing Injury in Major League Baseball. *The American Journal of Orthopedics*, *45*(3), 152–156.

Erickson, B. J., Chalmers, P. N., D'Angelo, J., Ma, K., Rowe, D., & Ahmad, C. S. (2021). Do injury rates in position players who convert to pitchers in professional baseball differ from players who have always been pitchers? *Orthopaedic Journal of Sports Medicine*, *9*(10), 232596712110509. https://doi.org/10.1177/23259671211050963

Fortenbaugh, D., Fleisig, G. S., and Andrews, J. R. (2009) "Baseball Pitching Biomechanics in Relation to Injury Risk and Performance," *Sports Health*, 1(4), pp. 314–320.

Hamano, N., Shitara, H., Tajika, T., Ichinose, T., Sasaki, T., Kuboi, T., Shimoyama, D., Kamiyama, M., Miyamoto, R., Endo, F., Nakase, K., Kobayashi, T., Yamamoto, A., Takagishi, K., & Chikuda, H. (2021) "Relationship between upper limb injuries and hip range of motion and strength in high school baseball pitchers," *Journal of Orthopaedic Surgery*, 29(1), pp. 1-6

Hamilton, A. G. (2017). *Reducing Asymmetric Information: Empirical Evidence From Major League Baseball* (thesis). University of North Carolina, Charlotte.

Hattery, M. (2017). Major League Baseball Players, Big Data, and the Right to Know: The Duty of Major League Baseball Teams to Disclose Health Modeling Analysis to Their Players. *Marquette Sports Law Review*, *28*(1), 1–28.

Hurd, W. J., Jazayeri, R., Mohr, K., Limpisvasti, O., Elattrache, N. S., & Kaufman, K. R., (2012) "Pitch velocity is a predictor of medial elbow distraction forces in the uninjured high school-aged baseball pitcher," *Sports health*, *4*(5), pp. 415–418.

James, M., & Bennison, D. (2020, December 16). *MLB Pitcher Injury Prediction using Statcast Data*. Mallet James. Retrieved February 28, 2022, from https://www.maljames.com/post/mlb-pitcher-injury-prediction/ .

Krautmann, Anthony and John L. Solow (2015) "(Ma)lingering on the Disabled List". *Contemporary Economic Policy, 33*(4), pp. 689-697.

Lehn, Kenneth. (1982). "Property Rights, Risk Sharing, and Player Disability in Major League Baseball," *The Journal of Law & Economics*, *25*(2), pp. 343–366.

Marshall, N. E., Keller, R., Limpisvasti, O., Schulz, B., & Elattrache, N. (2019) "Major League Baseball Pitching Performance After Tommy John Surgery and the Effect of Tear Characteristics, Technique, and Graft Type," *The American Journal of Sports Medicine*, 47(3), pp. 713–720.

Maxcy, J. G., Fort, R. D., and Krautmann, A. C. (2002) "The Effectiveness of Incentive Mechanisms in Major League Baseball," *Journal of Sports Economics, 3*(3), pp. 246-255.

Meldau, J. E., Srivastava, K., Okoroha, K. R., Ahmad, C. S., Moutzouros, V., & Makhni, E. C. (2020). Cost analysis of Tommy John surgery for Major League Baseball teams. *Journal of shoulder and elbow surgery*, *29*(1), 121–125. https://doi.org/10.1016/j.jse.2019.07.019

Platt, B. N., Uhl, T. L., Sciascia, A. D., Zacharias, A. J., Lemaster, N. G., & Stone, A. V. (2021) "Injury Rates in Major League Baseball During the 2020 COVID-19 Season," *Orthopaedic Journal of Sports Medicine.*

Poitras, Mark A. and Daniel Sutter (2020) "Property Rights and Resource Use: Evidence from MLB Starting Pitchers," *Applied Economics, 52*(60), pp. 6514-6524.

Svrluga, B. (2021, October 23). *Worried about your ace's $100 million arm? There is insurance for that.* The Washington Post. Retrieved April 27, 2022, from https://www.washingtonpost.com/news/sports/wp/2016/09/09/worried-about-your-aces-100-million-arm-there-is-insurance-for-that/

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). South-Western Cengage Learning.

Woolway, Mark D., (1997) "Using an Empirically Estimated Production Function for Major League Baseball to Examine Worker Disincentives Associated with Multi-Year Contracts," *The American Economist*, *41*(2), pp. 77–83.

Wymore, L., Chin, P., Geary, C., Carolan, G., Keefe, D., Hoenecke, H., and Fronek, J. (2016) "Performance and Injury Characteristics of Pitchers Entering the Major League Baseball Draft After Ulnar Collateral Ligament Reconstruction," *The American Journal of Sports Medicine*, 44(12), pp. 3165–3170.

VIII. Data Sources

Baseball Savant. (2021). MLB Player Pitching Statistics, 2015-2021. Accessed 2021-12-01. https://baseballsavant.mlb.com/

Baseball Savant. (2014). MLB Player Pitching Statistics, 2000-2014. Accessed 2021-11-15. https://baseballsavant.mlb.com/

Euston, Jeff. (2021) Cot's Baseball Contracts. Opening Day Salaries, 2000-2021. Accessed 2021-11-24. https://legacy.baseballprospectus.com/compensation/cots/

Marousek, Frank. (2021). Baseball Transactions Injury Database, 2000-2021. Accessed 2021-12-01. https://prosportstransactions.com/ [2]

Pappas, Doug. (2005). SABR Business of Baseball Committee. Player Salary Database, 1985-2004. Accessed 2021-11-20. http://roadsidephotos.sabr.org/baseball/salaries.zip

Roegele, Jon. (2021). Tommy John Surgery List, 1974-2021. Accessed 2021-11-25. https://tinyurl.com/2p95ra6u [3]

IX. Appendix

All Stata Code can be found at:
https://drive.google.com/file/d/15JZ0Z2rCLBliWqHs4b7am1P64-HvKl9j/view?usp=sharing ***(Updated link 06/03/2025)***

[2] Used Program "Data Miner" to scrape data from this website. Also used Github Repository @robotallie (https://github.com/robotallie/baseball-injuries) with partially scraped data.

[3] Jon Roegele's Database was found on his website Cot's Baseball Contracts (https://legacy.baseballprospectus.com/compensation/cots/)