**Forecasting Ulnar Collateral Ligament Injuries in Major League Pitchers: PIP and PHI**

Jason P. Hillman, DO, MBA, CPE

**Introduction**

The minimum cost for producing a certain level of output, is equal to the sum of fixed cost and variable cost; the basic definition of a cost function. Major League Baseball teams use various strategies to attempt to control and predict their largest year to year variable cost; player salaries. Player injuries can be devastating to a team’s value, and some of the most devastating injuries occur amongst pitchers.

A tear of the ulnar collateral ligament is perhaps the most serious injury to befall a major league pitcher. The full recovery usually averages 12-16 months, but may be as long as one to two years.[[1]](#endnote-1) Teams plan for the pitchers on their 25-man roster to be healthy for the season, and every year teams lose millions of dollars on pitchers needing Tommy John surgery. A prior UCL tear is a great predictor for a future injury, but how do we predict injuries in pitchers who have never been injured?

**Related Work**

Jang, Jung and Sharma of Stanford University attempted to answer this exact question. They used two primary datasets; player statistics containing 33 features on every MLB pitcher for the past five seasons, and every MLB pitcher who had Tommy John surgery since 1986. Their methods included a series of statistical analyses including k-fold, GDA, linear regression, SVM, Gaussian kernel, majority down-sampling, and defining a cost function.

When evaluating their cost function, one of the most important findings was that **the cost of failing to identify an injury is much higher than the cost of mislabeling a healthy player as injured**. They also found that despite analyzing 33 features for each pitcher, 15 features alone accounted for 99% of the variance in the dataset. A forward search process was used to determine the top 10 highest ranked features (figure 1), which had similar utility versus using all 33.[[2]](#endnote-2)

Figure 1



**Dataset**

I gathered pitcher data for the above 10 features for all 354 MLB pitchers who had Tommy John surgery for the first time between 1993 and 2016.[[3]](#endnote-3) [[4]](#endnote-4) Based on the number of MLB pitchers and the league totals for these 10 features for each year, I calculated an average value for each feature and the probability of UCL injury (TJ%) (figure 2).

Figure 2



**Pitcher Injury Probability**

The PIP is defined by the last column in figure 2, which is a calculation based on the number of individuals whom underwent TJ surgery that year, and the total number of MLB pitchers that year.

**Pitcher Health Index**

The PHI is then defined as

**Results**

A linear regression was run based on the data in figure 2, to determine an equation that would predict a pitcher’s ulnar collateral ligament injury probability (figure 3).

*Figure 3*



The conservative estimate of Adjusted R Square indicates that 62.4% of the variability in a pitcher’s injury probability is due to the variability in these 10 features. Overall, this is an accurate regression. The Significance F also demonstrates the small probability that this regression output occurred by chance. However, our p-values demonstrate that not all of these features are predictive in a meaningful way, indicated by the precipitous drop fromR2 to Adjusted R2. A second regression was run, removing SO, TBF, IP, GP, and R (figure 4).

*Figure 4*



The conservative estimate of Adjusted R Square indicates that 57% of the variability in a pitcher’s injury probability is due to the variability in these five features. Overall, this is a more accurate regression, evidenced by the F score. The Significance F demonstrates the small probability that this regression output occurred by chance. The p-values of the intercept and these five features are all below 0.05, indicating the chance that this predictive relationship occurred by chance to be less than 5%. We can then formulate our equation to predict a player’s PIP based on the intercept and the five independent variables.

The specific values of these five features a pitcher is projected to have for a particular year can be calculated based on a simple, yet very accurate method in three steps[[5]](#endnote-5):

1. A weighted average of the player’s three most recent seasons; if less than three seasons of MLB data exist, an APY (average player year) for that age may be substituted for the missing years
2. A regression of the player’s performance to the league mean, based on the past 24 seasons of player data, particular to age, in order to create an APY for each age
3. An application of an age adjustment based on said regression over the past 24 seasons of player data

Step one is the weighted performance average which can be calculated as 4-2-1, with each successive season having twice the predictive power as the season before.

Step two involves determining the random variation in a particular statistic. By separating all of the statistics used into the even and odd years between 1993 and 2016, we can calculate the correlation (R) for each. The higher the R, the more we can assume that the value of the observed statistic is based on player skill as opposed to random variation or chance. If we subtract the square root of R from one, we can calculate the actual percentage of player’s data which should be regressed to the league mean. So, one minus this number would then be the percentage of a player’s data which would be based on an actual three year projection.

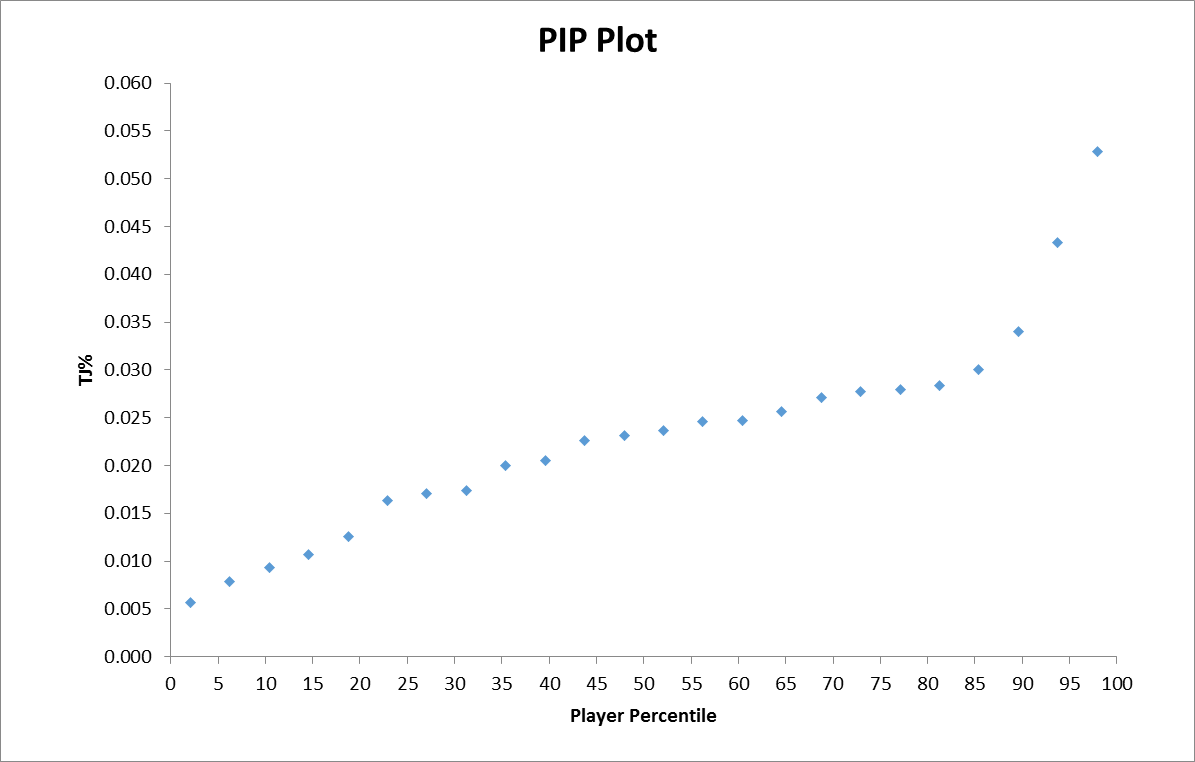
Step three reminds us that all players have a natural age curve of development and decline. If you take the mean value for each statistic over the past 24 years, grouped by each age from 20 to 40 years old, you can calculate the average delta (season to season variability) for a chosen statistic.

Let’s bring it all together in an example. In 2014, 2015, and 2016, Clayton Kershaw’s ERA was 1.77, 2.13 and 1.69 for ages 26, 27, and 28 of his career. The correlation, or R, in the ER and IP statistic for all players is 0.9790 and 0.9831. One minus the square root of R for ER is 0.0106, and one minus this number is 0.9894. So, another way of saying this is that 98.94% of the variation in ER is attributable to skill and 1.06% is attributable to chance, and should be regressed to the league mean. Do this for IP as well and we find that 99.15% of the variation in IP is attributable to skill and 0.85% is attributable to chance, and should be regressed to the league mean. When examining the age curve for season ages 28 and 29, the delta in ERA is an increase of 0.0265. Putting together the three year weighted average, the appropriate regression to the mean, and the age adjusted curve, we would predict in 2017 that Clayton Kershaw’s ERA (ER\*9/IP) would be 2.00.

**Discussion**

Because the average PIP for any MLB pitcher is 0.023 over the past 24 years, one must look at a percentile instead of an absolute probability determination in order to define the true utility of this statistic; the approximate PHI of all MLB pitchers each year is 0.977. When we examine the probability plot of the last regression, we can get an idea of when a PIP may be concerning (figure 5).

Figure 5



The pitcher’s PIP (TJ%) ascends rather steadily and predictably, until it assumes a value of around the 85th player percentile. As the PIP increases over 0.03, it does so precipitously. So, we could generally say that any pitcher with a PIP of below 0.02, is at **below average** risk for UCL injury, 0.02-0.03 about **average** risk, and above 0.03 at **above average** risk.

There will always be variation amongst the anatomy and physiology of each pitcher, some naturally more resilient to injury than others. However, one of the main categories of information to explain the other 43% of variability in PIP is mechanics. Possible future research should include an exhaustive mechanical analysis of pitcher careers without injuries versus those who’ve been injured. How does the pitcher stand at the beginning of his motion? What does the pitcher do in his first motion? How would you describe his leg kick? At what angle does the pitcher deliver the ball? Is the pitcher’s arm fully extended during the throw? Is the pitcher looking at home plate throughout his motion? How would you describe what the pitcher does with his trail leg after release? Is the pitcher balanced when he completes his motion? Is the pitcher’s motion quick or slow, jerky or smooth?[[6]](#endnote-6) These questions can all contribute to how stress is being applied to the ulnar collateral ligament in the throwing arm.

**Notes**

1. <http://m.mlb.com/pitchsmart/tommy-john-faq/>. [↑](#endnote-ref-1)
2. Jae Jang, Raejoon Jung, Prafull Sharma, “Identifying Injury-Inducing Factors in Baseball Pitchers”, 2015. <http://cs229.stanford.edu/proj2015/109_report.pdf>. [↑](#endnote-ref-2)
3. <https://mlbreports.com/tj-surgery/>. [↑](#endnote-ref-3)
4. <http://www.fangraphs.com/>. [↑](#endnote-ref-4)
5. Vollman, Rob. 2016. *Stat Shot.* Toronto: ECW Press. [↑](#endnote-ref-5)
6. James, Bill. 1982. *The Bill James Baseball Abstract.* New York: Ballantine Books.

   [↑](#endnote-ref-6)