**Machine Learning and Single-Season Sabermetrics**

**to Identify UCL Injury in MLB Pitchers**

Timothy I Morales

Department of Statistical Science, Southern Methodist University

STAT 6341: Sports Analytics

Dr. Charles South

December 16, 2020

**Abstract**

**Introduction**

There is money in America’s favorite pastime. Forbes reports that Major League Baseball, MLB, saw a record $10.7 billion in revenue in 2019. In that $10.7 billion includes a $5.1 billion dollar television deal with Fox. The MLB also signed a $1 billion uniform deal with Nike that started in the 2020 season (Brown, 2019). With that huge revenue comes great costs, a $4.7 billion payroll to be exact (Brown, 2019). Fans are paying to watch the MLB, brands are paying to host the MLB, fashion companies want to dress the MLB, and teams want to make the most effective use of their payroll to maximize profits. Since all contracts in the MLB are guaranteed, player injury is at the forefront of mitigating payroll risks, especially injuries associated with long-term performance decreases.

Ulnar collateral ligament, UCL, reconstruction, more commonly known as Tommy John surgery, is notorious in the game of baseball. The sport involving over-hand throwing motions has also been associated with high risks of elbow injury. Over-hand throwing motions can produce extremely high levels of valgus stress in the elbow which eventually leads to instability and increased risk for UCL injury (Chen et al., 2001). Recent research has shown, MLB pitchers are statistically the most at risk group for this UCL injury (Conte et al., 2015). In 2017, Fangraphs a commonly used baseball database, reported that 86.7% of MLB games in the 2017 season included at least one pitcher who had undergone UCL surgery in the past (FANGRAPHS).

Recovery times from UCL reconstruction vary between 12 and 15 months and can span across multiple seasons. Not only does the surgery require long recovery times, recent research has found the surgery linked to performance decreases (Selley et al., 2019). These long absences for recovery as well as long-term performances decreases are concerning for the league, the team, the individual players and the fans. With the afore mentioned billions of dollars going into player payroll each year, understanding and preventing injuries like UCL injuries has huge monetary value.

Since 2010, there have been multiple studies centered around identifying key risk factors for UCL injury. There have also been studies designed around classifying at risk players relative to a control group. In Erickson et. al. (2014), researchers found MLB players who attended high schools in warmer climates to be at a significantly higher risk. Keller et. al. (2015) found that individuals who throw a higher percent of fastballs were more likely to undergo the surgery relative to a control but found in difference in the effect of velocity.

Most similar to the research in this paper is Whiteside et al. (2016) where researchers used support vector machines and naïve bayes models including variables like average spin rates and average velocity to classify individuals against a control group of equal size. Although Whiteside et. al. (2016) was able to classify individuals correctly 75% of the time, this number is an extreme overestimate of the model’s actual predictive power. This 75% is the maximum average five-fold cross validation accuracy from testing over 4000 different models. Each fold was 41 or 42 individuals and had perfect class balance. Therefore, with no hold-out set to assess the optimal model’s true performance, and test folds that are only 42 individuals with perfect class balance, these findings do not stem from a research design that mimics the practical use of a classification model for UCL injury. Roughly 1000 player pitch in the MLB each season and only 2% of those pitchers will undergo UCL injury within the following year. (I hope this paragraph makes Frank Harrell proud.)

**References**

Brown, M. (2019). MLB Sees Record $10.7 Billion In Revenues For 2019. *Forbes. https://www. forbes. com/sites/maurybrown/2019/12/21/mlb-sees-record-107-billion-inrevenues-for- 2019/(accessed May 08, 2020)*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321- 357.

Chen, F. S., Rokito, A. S., & Jobe, F. W. (2001). Medial elbow problems in the overhead-throwing athlete. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, *9*(2), 99-113.

Conte, S. A., Fleisig, G. S., Dines, J. S., Wilk, K. E., Aune, K. T., Patterson-Flynn, N., & ElAttrache, N. (2015). Prevalence of Ulnar Collateral Ligament Surgery in Professional Baseball Players. The American Journal of Sports Medicine, 43(7), 1764–1769. <https://doi.org/10.1177/0363546515580792>

Erickson, B. J., Harris, J. D., Tetreault, M., Bush-Joseph, C., Cohen, M., & Romeo, A. A. (2014). Is Tommy John surgery performed more frequently in Major League Baseball pitchers from warm weather areas?. *Orthopaedic journal of sports medicine*, *2*(10), 2325967114553916.

Keller, R. A., Marshall, N. E., Guest, J. M., Okoroha, K. R., Jung, E. K., & Moutzouros, V. (2016). Major League Baseball pitch velocity and pitch type associated with risk of ulnar collateral ligament injury. *Journal of Shoulder and Elbow Surgery*, *25*(4), 671-675.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a random forest?. In *International workshop on machine learning and data mining in pattern recognition* (pp. 154-168). Springer, Berlin, Heidelberg.

Rhinehart, E. (2020). Tommy John Surgery: Potential Risk Factors and Causes in Major League Pitchers.

Selley, R. S., Portney, D. A., Lawton, C. D., Shockley, M. D., Christian, R. A., Saltzman, M. D., & Hsu, W. K. (2019). Advanced baseball metrics indicate significant decline in MLB pitcher value after Tommy John surgery. *Orthopedics*, *42*(6), 349-354.

Whiteside, D., Martini, D. N., Lepley, A. S., Zernicke, R. F., & Goulet, G. C. (2016). Predictors of ulnar collateral ligament reconstruction in Major League Baseball pitchers. *The American journal of sports medicine*, *44*(9), 2202-2209.