NURSG 741 Project

Lacey Gleason

3/28/2018

# Milestone 2

## Basic Information

**Project Title:** Predictors of Major League Baseball Player Injury in 2017  
**Author Name:** Lacey Gleason  
**Email Address:** [lpgleas@emory.edu](mailto:lpgleas@emory.edu)

### Load Libraries

library(readxl)  
library(plyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(wordcloud)

## Loading required package: RColorBrewer

library(RColorBrewer)  
library(tm)

## Loading required package: NLP

library(NLP)

### Import MLB Transaction Data

# Import transaction data from MLB.com  
# Downloaded as Excel file - includes all transactions for all months in 2017  
  
transactions <- read\_excel("Transactions.xlsx",   
 sheet = "Sheet1")  
  
head(transactions)

## # A tibble: 6 x 2  
## Date Transaction   
## <dttm> <chr>   
## 1 2017-01-02 00:00:00 Miami Marlins signed free agent OF Mark Traylor to …  
## 2 2017-01-02 00:00:00 Cincinnati Reds signed free agent RHP Geoff Broussa…  
## 3 2017-01-02 00:00:00 Cincinnati Reds signed free agent RHP Deunte Heath …  
## 4 2017-01-02 00:00:00 Cincinnati Reds signed free agent C Adrian Nieto to…  
## 5 2017-01-02 00:00:00 Washington Nationals signed free agent LHP Stone Sp…  
## 6 2017-01-03 00:00:00 Miami Marlins signed free agent 1B Tyler Moore to a…

There were 11,727 MLB transactions during the 12 months of 2017. The imported dataset contains a column of transaction dates and a column of text that describes the transactions. Within this text is the information about team, player name, player position, type of transaction, and type of injury (where applicable for disabled list transactions) that we want to pull out.

### Filter Transaction Data for Disabled List Transactions

First, we need to filter just for disabled list transactions.

# Filter just for those transactions that contain the term 'disabled list' - call dataframe new  
  
new <- dplyr::filter(transactions, grepl('disabled list', Transaction))  
head(new)

## # A tibble: 6 x 2  
## Date Transaction   
## <dttm> <chr>   
## 1 2017-02-14 00:00:00 Cincinnati Reds placed RHP Homer Bailey on the 60-d…  
## 2 2017-02-14 00:00:00 Texas Rangers placed DH Prince Fielder on the 60-da…  
## 3 2017-02-14 00:00:00 Texas Rangers placed LHP Jake Diekman on the 60-day…  
## 4 2017-02-14 00:00:00 Atlanta Braves placed LHP Jacob Lindgren on the 60-…  
## 5 2017-02-15 00:00:00 Kansas City Royals placed LHP Brian Flynn on the 60…  
## 6 2017-02-15 00:00:00 Los Angeles Dodgers placed RHP Yimi Garcia on the 6…

There were 1,386 MLB transactions in 2017 that involved the disabled list. We would like to know what types of transactions those were.

### Create Variables in Data Frame to Describe the Transactions

# create a new variable for action to reflect what kind of DL action it is  
 ## Placed means first instance of player being assigned to DL  
 ## Transferred means player was moved from shorter DL list to longer DL list  
 ## Activated means player was put back on active roster from DL  
  
new$action <- ifelse(grepl("transfer", new$Transaction, ignore.case = T), "Transferred",   
 ifelse(grepl("place", new$Transaction, ignore.case = T), "Placed", ifelse(grepl("activate", new$Transaction, ignore.case =T), "Activated", "Other")))  
  
#make a frequency table of action   
act\_count <- plyr::count(new, 'action')  
act\_count

## action freq  
## 1 Activated 543  
## 2 Placed 710  
## 3 Transferred 133

The proportion of disalbed list actions that were activations was 39.8%. The proportion of disabled list actions that were placements was 52.0%. The proportion of disabled list actions that were transfers was 9.7%. The percentages do not add up to 100% due to rounding. Since there were no “Other” actions in frequency table, we know that we’re not missing a DL action other than those listed.

Next, we’d like to pull out player characteristics from the text.

# Create variable for position of player  
new$position <- ifelse(grepl(" C ", new$Transaction, ignore.case = F), "Catcher",   
 ifelse(grepl("1B", new$Transaction, ignore.case = F), "First Baseman", ifelse(grepl("2B", new$Transaction, ignore.case =F), "Second Baseman", ifelse(grepl("3B", new$Transaction, ignore.case =F), "Third Baseman", ifelse(grepl("SS", new$Transaction, ignore.case =F), "Shortstop", ifelse(grepl("RF", new$Transaction, ignore.case =F), "Right Fielder", ifelse(grepl("CF", new$Transaction, ignore.case =F), "Center Fielder", ifelse(grepl("LF", new$Transaction, ignore.case =F), "Left Fielder", ifelse(grepl("RHP", new$Transaction, ignore.case =F), "Pitcher", ifelse(grepl("LHP", new$Transaction, ignore.case =F), "Pitcher", ifelse(grepl(" P ", new$Transaction, ignore.case =F), "Pitcher", ifelse(grepl("DH", new$Transaction, ignore.case =F), "Designated Hitter", ifelse(grepl("OF", new$Transaction, ignore.case =F), "Outfielder", "Other")))))))))))))  
  
# Create frequency table of position  
pos\_count <- plyr::count(new, 'position')  
pos\_count

## position freq  
## 1 Center Fielder 91  
## 2 Designated Hitter 15  
## 3 First Baseman 50  
## 4 Left Fielder 105  
## 5 Other 95  
## 6 Outfielder 6  
## 7 Pitcher 763  
## 8 Right Fielder 51  
## 9 Second Baseman 66  
## 10 Shortstop 66  
## 11 Third Baseman 78

We see that 55% of the players involved in DL transactions were pitchers. Later, after we de-duplicate the placements, transfers, and activations, we can compare the player type distribution to that of the active players in the league overall in 2017.

Next, we will look at which type of disabled list is utilized in most disabled list transactions. There is a 10-day disabled list, a 60-day disabled list, and a special 7-day disabled list for concussions. There have been some interesting changes to the length of the disabled list over the years (e.g., elimination of 15-day disabled list and re-introduction of 10-day disabled list) to try to nudge teams and players to make the best choices in weighing a player’s health and a team’s need for the player’s on-field contributions. Timing of these policy changes will be kept in mind during this analysis.

# Create a new variable called list that refelcts type of DL referenced in transaction  
new$list <- ifelse(grepl("10-day", new$Transaction, ignore.case = T), "10-day",   
 ifelse(grepl("60-day", new$Transaction, ignore.case = T), "60-day", ifelse(grepl("7-day", new$Transaction, ignore.case = T), "7-day", "Other")))  
  
#make a frequency table of type of disabled list  
list\_count <- plyr::count(new, 'list')  
list\_count

## list freq  
## 1 10-day 1181  
## 2 60-day 168  
## 3 7-day 37

attach(new)  
list\_table <- table(list,action) # A will be rows, B will be columns   
list\_table # print table

## action  
## list Activated Placed Transferred  
## 10-day 408 640 133  
## 60-day 124 44 0  
## 7-day 11 26 0

prop.table(list\_table, 2) # column percentages

## action  
## list Activated Placed Transferred  
## 10-day 0.75138122 0.90140845 1.00000000  
## 60-day 0.22836096 0.06197183 0.00000000  
## 7-day 0.02025783 0.03661972 0.00000000

We see that 85% of the disabled list transactions (90% of placements) involved actions related to the 10-day disabled list. Twelve percent of the DL transactions (6% of placements) involved the 60-day disabled list and 3% of the DL transactions (4% of placements) involved the 7-day disabled list. All of the transfers involved transfers from the 10-day disabled list.

Another important characteristic of DL placements is whether they are retroactive to an earlier date. If a player has not played in up to 5 days, the team can then retroactively place the player on the disabled list and have those days count towards meeting the length requirement of the disabled list. Since teams are protected by being able to retroactively place players on the DL, they tend to list players as day-to-day in case they get better within five days and can avoid having to serve a full ten days on the DL.

# Create new variable to indicate if this DL transaction is a retroactive move?  
new$retro <- ifelse(grepl("retroactive", new$Transaction, ignore.case = T), "Yes", "No")  
  
#make a frequency table of action   
retro\_count <- plyr::count(new, 'retro')  
retro\_count

## retro freq  
## 1 No 1078  
## 2 Yes 308

attach(new)

## The following objects are masked from new (pos = 3):  
##   
## action, Date, list, position, Transaction

retro\_table <- table(retro,action) # A will be rows, B will be columns   
retro\_table # print table

## action  
## retro Activated Placed Transferred  
## No 543 404 131  
## Yes 0 306 2

prop.table(retro\_table, 2) # column percentages

## action  
## retro Activated Placed Transferred  
## No 1.00000000 0.56901408 0.98496241  
## Yes 0.00000000 0.43098592 0.01503759

Examining the table above, we see that 22% of the DL transactions are retroactive. This represents 43% of DL placements in 2017.

Another interesting component that we want to pull out of this transaction text is information about the injuries precipitating DL transactions. Below, we will look at if surgeries are involved.

# Create a variable to indicate if surgery is indicated in DL transaction description  
new$surg <- ifelse(grepl("surgery", new$Transaction, ignore.case = T), "Yes", "No")  
  
#make a frequency table of surgery   
surg\_count <- plyr::count(new, 'surg')  
surg\_count

## surg freq  
## 1 No 1357  
## 2 Yes 29

attach(new)

## The following objects are masked from new (pos = 3):  
##   
## action, Date, list, position, retro, Transaction

## The following objects are masked from new (pos = 4):  
##   
## action, Date, list, position, Transaction

surg\_table <- table(surg,action) # A will be rows, B will be columns   
surg\_table # print table

## action  
## surg Activated Placed Transferred  
## No 543 687 127  
## Yes 0 23 6

prop.table(surg\_table, 2) # column percentages

## action  
## surg Activated Placed Transferred  
## No 1.00000000 0.96760563 0.95488722  
## Yes 0.00000000 0.03239437 0.04511278

There were 29 disabled list transactions in 2017 that mentioned surgery. Surgeries were involved in 23 (3%) of DL placements.

Let’s look at concussions.

# Create variable for concussion involvement  
new$concuss <- ifelse(grepl("concussion", new$Transaction, ignore.case = T), "1", "0")  
  
#make a frequency table of action   
concuss\_count <- plyr::count(new, 'concuss')  
concuss\_count

## concuss freq  
## 1 0 1342  
## 2 1 44

attach(new)

## The following objects are masked from new (pos = 3):  
##   
## action, Date, list, position, retro, surg, Transaction

## The following objects are masked from new (pos = 4):  
##   
## action, Date, list, position, retro, Transaction

## The following objects are masked from new (pos = 5):  
##   
## action, Date, list, position, Transaction

con\_table <- table(concuss,action) # A will be rows, B will be columns   
con\_table # print table

## action  
## concuss Activated Placed Transferred  
## 0 543 670 129  
## 1 0 40 4

prop.table(con\_table, 2) # column percentages

## action  
## concuss Activated Placed Transferred  
## 0 1.00000000 0.94366197 0.96992481  
## 1 0.00000000 0.05633803 0.03007519

There were 44 disabled list transactions in 2017 that mentioned concussions. Forty (6%) of DL placements involved concussions. This is around the number we would expect since we saw there were 37 instances where the 7-day disabled list was used, which is specfically for concussions.

Next, we will look at which side of the body is involved in the injury.

# Create new variable to indicate side of body of injury  
new$side <- ifelse(grepl("right", new$Transaction, ignore.case = T), "Right", ifelse(grepl("left", new$Transaction, ignore.case = T), "Left", "Unknown"))  
  
#make a frequency table of action   
side\_count <- plyr::count(new, 'side')  
side\_count

## side freq  
## 1 Left 240  
## 2 Right 385  
## 3 Unknown 761

attach(new)

## The following objects are masked from new (pos = 3):  
##   
## action, concuss, Date, list, position, retro, surg,  
## Transaction

## The following objects are masked from new (pos = 4):  
##   
## action, Date, list, position, retro, surg, Transaction

## The following objects are masked from new (pos = 5):  
##   
## action, Date, list, position, retro, Transaction

## The following objects are masked from new (pos = 6):  
##   
## action, Date, list, position, Transaction

side\_table <- table(side,action) # A will be rows, B will be columns   
side\_table # print table

## action  
## side Activated Placed Transferred  
## Left 0 210 30  
## Right 4 311 70  
## Unknown 539 189 33

prop.table(side\_table, 2) # column percentages

## action  
## side Activated Placed Transferred  
## Left 0.000000000 0.295774648 0.225563910  
## Right 0.007366483 0.438028169 0.526315789  
## Unknown 0.992633517 0.266197183 0.248120301

For those DL placements that included the side of the body involved in the injury, 60% involved the right side of the body. However, 27% of all DL placements did not list side of the body.

Next, let’s look at DL transactions that mentioned Tommy John surgery, a common procedure for pitchers.

# Tommy John Surgery indicator  
new$tom <- ifelse(grepl("Tommy John", new$Transaction, ignore.case = T), "1", ifelse(grepl(" UCL ", new$Transaction, ignore.case = T), "1", ifelse(grepl("ulnar collateral", new$Transaction, ignore.case = T), "1", "0")))  
  
#make a frequency table of action   
tom\_count <- plyr::count(new, 'tom')  
tom\_count

## tom freq  
## 1 0 1368  
## 2 1 18

attach(new)

## The following objects are masked from new (pos = 3):  
##   
## action, concuss, Date, list, position, retro, side, surg,  
## Transaction

## The following objects are masked from new (pos = 4):  
##   
## action, concuss, Date, list, position, retro, surg,  
## Transaction

## The following objects are masked from new (pos = 5):  
##   
## action, Date, list, position, retro, surg, Transaction

## The following objects are masked from new (pos = 6):  
##   
## action, Date, list, position, retro, Transaction

## The following objects are masked from new (pos = 7):  
##   
## action, Date, list, position, Transaction

tom\_table <- table(tom,action) # A will be rows, B will be columns   
tom\_table # print table

## action  
## tom Activated Placed Transferred  
## 0 543 698 127  
## 1 0 12 6

prop.table(tom\_table, 2) # column percentages

## action  
## tom Activated Placed Transferred  
## 0 1.00000000 0.98309859 0.95488722  
## 1 0.00000000 0.01690141 0.04511278

Overall, 1.7% of DL placements and 4.5% of DL transfers involve Tommy John surgery.

# create variable for month of transaction - later it will be useful to have this  
new$month <- format(new$Date,"%B")

## Create word cloud of disabled list descriptions

Next, let’s see if we can make a wordcloud of the disabled list descriptions.

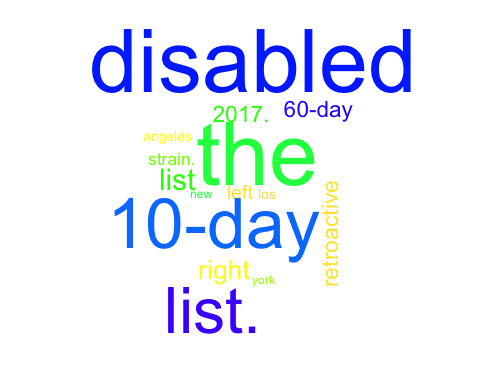
DL\_lines <- grep("disabled list", transactions$Transaction, value = TRUE)  
length(DL\_lines)

## [1] 1386

head(DL\_lines)

## [1] "Cincinnati Reds placed RHP Homer Bailey on the 60-day disabled list. Right elbow surgery."   
## [2] "Texas Rangers placed DH Prince Fielder on the 60-day disabled list. Disc herniation in neck."   
## [3] "Texas Rangers placed LHP Jake Diekman on the 60-day disabled list. Colon surgery."   
## [4] "Atlanta Braves placed LHP Jacob Lindgren on the 60-day disabled list. Recovery from Tommy John surgery."   
## [5] "Kansas City Royals placed LHP Brian Flynn on the 60-day disabled list. Stable lumbar vertebral fracture."   
## [6] "Los Angeles Dodgers placed RHP Yimi Garcia on the 60-day disabled list. Recovering from Tommy John Surgery"

# Make wordcloud of words used in disabled list transactions   
  
wordcloud(VCorpus(VectorSource(DL\_lines)), max.words = 15, scale =c(5.5,.4), colors = topo.colors(n=30), random.color = TRUE)



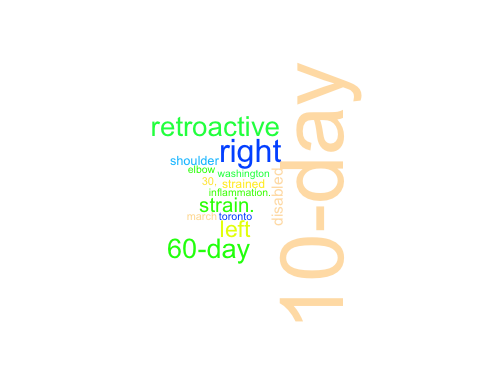
#Get rid of common words so that word cloud is more interesting  
# took out common words like "the" and also word that are part of multiple teams' names   
#(e.g., Los, York, Chicago)  
  
pattern <- "disabled"  
DL\_lines2 <- sub(pattern, "", DL\_lines)  
  
tail(DL\_lines2)

## [1] "Los Angeles Dodgers activated LF Andrew Toles from the 60-day list."   
## [2] "Philadelphia Phillies activated RHP Zach Eflin from the 60-day list."   
## [3] "Philadelphia Phillies activated RHP Jerad Eickhoff from the 60-day list."   
## [4] "Philadelphia Phillies activated RHP Vince Velasquez from the 60-day list."  
## [5] "Pittsburgh Pirates activated 2B Josh Harrison from the 60-day list."   
## [6] "St. Louis Cardinals activated RHP Alex Reyes from the 60-day list."

DL\_lines2[200:220]

## [1] "San Francisco Giants placed LHP Madison Bumgarner on the 10-day list. Bruised ribs and sprained left shoulder."   
## [2] "Cleveland Indians activated 2B Jason Kipnis from the 10-day list."   
## [3] "Detroit Tigers placed SS Jose Iglesias on the 7-day list retroactive to April 20, 2017. Concussion."   
## [4] "Los Angeles Angels placed RHP Mike Morin on the 10-day list retroactive to April 20, 2017. Neck tightness."   
## [5] "Chicago White Sox placed RHP James Shields on the 10-day list retroactive to April 18, 2017. Strained right lat."   
## [6] "Tampa Bay Rays placed LHP Xavier Cedeno on the 10-day list retroactive to April 18, 2017. Left forearm tightness."   
## [7] "Texas Rangers placed RHP A.J. Griffin on the 10-day list retroactive to April 18, 2017. Gout in left ankle."   
## [8] "Detroit Tigers placed 1B Miguel Cabrera on the 10-day list. Right groin strain."   
## [9] "Chicago White Sox activated C Geovany Soto from the 10-day list."   
## [10] "Chicago White Sox transferred CF Charlie Tilson from the 10-day list to the 60-day disabled list. Stress reaction in right foot."   
## [11] "Philadelphia Phillies transferred RHP Clay Buchholz from the 10-day list to the 60-day disabled list. Torn flexor tendon in right elbow."  
## [12] "Los Angeles Angels transferred RHP Garrett Richards from the 10-day list to the 60-day disabled list. Right biceps strain."   
## [13] "Los Angeles Angels placed RHP Cam Bedrosian on the 10-day list. Right groin strain."   
## [14] "Minnesota Twins placed RHP Justin Haley on the 10-day list. Right bicep tendinitis."   
## [15] "Tampa Bay Rays placed RHP Tommy Hunter on the 10-day list. Right calf strain."   
## [16] "San Diego Padres activated RHP Luis Perdomo from the 10-day list."   
## [17] "Oakland Athletics transferred RHP Chris Bassitt from the 10-day list to the 60-day disabled list. Recovering from Tommy John surgery."   
## [18] "Toronto Blue Jays transferred CF Dalton Pompey from the 10-day list to the 60-day disabled list. Concussion."   
## [19] "Toronto Blue Jays placed SS Troy Tulowitzki on the 10-day list. Strained right hamstring."   
## [20] "San Francisco Giants placed CF Denard Span on the 10-day list."   
## [21] "Philadelphia Phillies placed RHP Aaron Nola on the 10-day list retroactive to April 21, 2017. Lower back strain."

pattern2 <- "list"  
DL\_lines2 <- sub(pattern2, "", DL\_lines2)  
pattern3 <- "Chicago"  
DL\_lines2 <- sub(pattern3, "", DL\_lines2)  
pattern4 <- "San"  
DL\_lines2 <- sub(pattern4, "", DL\_lines2)  
pattern5 <- "Los"  
DL\_lines2 <- sub(pattern5, "", DL\_lines2)  
pattern6 <- "New"  
DL\_lines2 <- sub(pattern6, "", DL\_lines2)  
pattern7 <- " on"  
DL\_lines2 <- sub(pattern7, "", DL\_lines2)  
pattern8 <- " the"  
DL\_lines2 <- sub(pattern8, "", DL\_lines2)  
pattern9 <- "from"  
DL\_lines2 <- sub(pattern9, "", DL\_lines2)  
pattern10 <- " The "  
DL\_lines2 <- sub(pattern10, "", DL\_lines2)  
pattern11 <- "Angeles"  
DL\_lines2 <- sub(pattern11, "", DL\_lines2)  
pattern12 <- "list."  
DL\_lines2 <- sub(pattern12, "", DL\_lines2)  
pattern13 <- "2017"  
DL\_lines2 <- sub(pattern13, "", DL\_lines2)  
pattern14 <- "York"  
DL\_lines2 <- sub(pattern14, "", DL\_lines2)  
pattern15 <- "Red"  
DL\_lines2 <- sub(pattern15, "", DL\_lines2)  
pattern16 <- "Blue"  
DL\_lines2 <- sub(pattern16, "", DL\_lines2)  
pattern17 <- "Bay"  
DL\_lines2 <- sub(pattern17, "", DL\_lines2)  
pattern18 <- "to the"  
DL\_lines2 <- sub(pattern18, "", DL\_lines2)  
pattern19 <- " and"  
DL\_lines2 <- sub(pattern19, "", DL\_lines2)  
pattern20 <- "Disabled"  
DL\_lines2 <- sub(pattern20, "", DL\_lines2)  
pattern21 <- "Sox"  
DL\_lines2 <- sub(pattern21, "", DL\_lines2)  
  
wordcloud(VCorpus(VectorSource(DL\_lines2)), max.words = 15, scale =c(5.5,.4), colors = topo.colors(n=30), random.color = TRUE)



### Next steps

* Finish text analysis
  + Pull out player name
  + Pull out team name
* De-duplicate transaction data so that placements, transfers, and activations for the same player are included in the same row for a single player
  + Calculate time on DL
* Merge baseball reference data on hitter and pitcher stats and characteristics that has already been downloaded
  + Match on multiple characteristics to ensure that player is correct since there are many comon names
* Run logistic regression on 20% sample of data to train
* Use model to predict outcome for the remaining 80% of data

## Schedule of Weekly Tasks/Goals

|  |  |
| --- | --- |
| Week | Task |
| March 26 | Finish text analysis of transactio data and merge Baseball Reference data on player characteristics of interest and add to disabled list data matching by player name, position, and team |
| April 2 | Create Table 1 for descriptive statistics, create tables/graphs for model results |
| April 2 | Write first draft of introduction and results sections of manuscript |
| April 9 | Write discussion/conclusions section of manuscript |
| April 16 | Submit manuscript |
| April 23 | Submit presentation |

## Link

The link for this file is located at <https://github.com/lpgleason/Project.git>.