Differences Between Injured and Healthy Players in the NBA

Matthew Verhey

Marquette University
Milwaukee, Wisconsin
matthew.verhey@marquette.edu

ABSTRACT

The goal of my research was to investigate statistical differences between injured and healthy players in the National Basketball Association. The data I worked with was the NBA player statistics for the 2017-2018 season, as well as a data set containing all injury reports from the All-Star break in February, to the end of the season in June. With this data, I set out to answer my research question, are there predictors of points-per-game that are correlated more with healthy or injured players? I ran linear regressions with my predictors being minutes per game, games played, and age. I found all predictors, except games played, to have a strong relation to points per game for both groups. Age was a stronger predictor for points per game for injured players than they were for healthy players. With these results, I began to formulate to how to answer my research questions within the scope of injury research as a whole.

ACM Reference Format:
Matthew Verhey. 2018. Differences Between Injured and Healthy Players in the NBA. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The past few months, I have been particularly interested in injuries and their effects in the National Basketball Association. This interest grew when I read a mock draft article that took a cautious approach towards taller, more slim-built players [8]. With my initial research, the question I wanted to research was "are taller players more likely to be injured than shorter players?"[25]. From my basic statistical analysis, I found that centers were around 15% of total injured players. Assuming the composition of a team is two guards, two forwards, and one center, 20% percent of players are centers [25]. This showed that for the season in which I had injury data, centers were less likely to be injured than other positions.

From here, I transitioned my research towards examining statistical differences between injured and healthy players. I set out to acquire a data set of box score statistics for an entire NBA season. My goal was to use linear regression methods to determine whether certain player factors were strong predictors of points per game. I sought to determine whether or not these predictors would vary across players who were injured and those who were not. To begin to address this question, I must delve into the details of my data.

1.1 The Data

The first data set I used for my research was a set of injury reports from February 22nd to June 6th, 2018. The entries of this data set were the player's name, team of player, data of report, and the reason for missing a game. This included entries for players returning from injury. As injury reports are made public by the National Basketball Association, the data violates no ethical norms. My reason for using this data set was to categorize players into a section for injured and a section for healthy. This categorization will be vital for properly answering my research questions.

Date	Team	Relinquished
2/24/18	Grizzlies	Marc Gasol
2/25/18	Timberwolves	Jimmy Butler
2/28/18	Lakers	Josh Hart
3/2/18	Clippers	Danilo Gallinari
3/2/18	Lakers	Josh Hart
3/5/18	Hawks	Malcolm Delaney
3/5/18	Hornets	Michael Carter-Williams
3/8/18	Magic	Evan Fournier
3/9/18	Celtics	Jaylen Brown
3/9/18	Hornets	Michael Carter-Williams

... (68 rows omitted)

Figure 1: Injury Data Set

The second data set I worked with was the 2017-2018 season statistics for every player in the NBA. This set included age, position, games played, and all box score statistics such as field goal percentage, effective field goal percentage, rebounds per game, assists per game, offensive rating, defensive rating, and player efficiency rating. With the data in this set, I chose my independent

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and dependent variables to conduct linear regressions.

With my first data set, the only purpose I had for it was to create a list of names of injured players. Then, with my second data set of all player's statistics, I could apply add a data field containing a boolean value indicating whether or not that player was injured.

FULL_NAME	TEAM	POS	AGE	GP	MPG
Aaron Brooks	Min	PG	33	32	5.9
Aaron Gordon	Orl	SF	22	58	32.9
Aaron Harrison	Dal	SG	23	9	25.9
Aaron Jackson	Hou	F	31	1	34.5
Abdel Nader	Bos	SF	24	48	10.9
Adreian Payne	Orl	PF	26	5	8.5
Al Horford	Bos	C	31	72	31.6
Al Jefferson	Ind	C	33	36	13.4
Alan Williams	Pho	PF	25	5	14.1
Alec Burks	Uta	SG	26	64	16.5

Figure 2: Player Statistics Data Set

Data Cleaning. For my first data set, the cleaning I performed was fairly straightforward. I dropped the columns that were unnecessary to my research. Within the data, there were entries for injuries that contained 'DTD'. This means that the injury was day-to-day and not long-term or serious. I felt that this group should not be included in my injured data set, so I removed them. The cleaning on my second data set was much less detailed. The main statistics I was interested in investigating were games played, minutes per game, age and points per game. So, to clean it, I removed all irrelevant data from my set.

1.2 Literature Review

For my literature review, I found myself rereading the journals from my past project, trying to analyze them from a different perspective. The most important takeaways from my first project were regarding past research on injuries and how "the focus has not been on specific sports" and how the absence of an objective injury definition had made "comparison[s] across studies difficult [11]. Also, different, older, factors once believed to have been accurate predictors for injury have been nullified by newer research. For example, one researcher in 1970 found loose-jointed football players suffered more injuries than tight-jointed players. This finding was later countered by research done in 1974, 1975, 1978 and 1982 finding "no such relationship between joint stability and injury"[14]. This repeated theme of past research not being reliable makes it difficult to define my own findings within the field of injury research.

With my new research questions, I began looking for injury research that aimed at predicting injury. The journal "Structural measures as predictors of injury in basketball players" provided me with a good account of how predictive research is done with regards to basketball injuries. The goal of this paper was to "develop equations to predict injury" by examining structural components of the subjects. The variables used for the logistic regression performed in this paper were bilateral weight, quadricep girth, calf girth, Q-angle of the knee, dorsiflexion of the ankle, forefoot varus, rear-foot valgus, and leg length [x]. From their three-variable logistic regression, they formulated an injury equation [Figure 3] that produces a score. Based on this score, subjects were classified into

$$score = weightimbalance*.036+rightabnormalQangle*.48+leftabnormalQangle*.086+intercept$$

Figure 3: Injury Score Equation

a category for either healthy or injured players. The interpretation of the scores was rather straightforward, with a score greater than zero classifying a subject into the injured category, and a score less than zero classifying a subject into the non-injured category. With this equation, they were able to correctly predict the injury status of 91% of the subjects. The final conclusion the authors of this journal came to was that a relationship between bodily structure measurements and injury rates in basketball.

To conclude my review of this article, it has provided me with insight on how to interpret my results. While my data significantly differs from the data used in this research, I have still gained the insight that injury can be accurately predicted.

The work of Loeffelholz et al published in the *Journal of Quantitative Analysis in Sports* titled Predicting NBA Games Using Neural Networks was another piece of literature vital to my research. This journal from 2009 aims at predicting NBA games using neural networks. In this research, they creating their models using two techniques. The first examined "the current season average of each team" taking into account home and away games [17]. The second "used only the average of the previous five games played by each team" to attempt to consider winning or losing streaks [17]. The data collection from this research has provided me some insight into how NBA statistics are mathematically interpreted. This journal made me consider my own data collection techniques and what variables I should consider. For example, with regards to my research, how should I decide which variables to run linear regressions on. The most readily available NBA datasets are season averages, so I knew averages would be the type of numbers I'd be

dealing with. As stated in Loeffelholz et al's work, " a teamâĂŹs average statistics typically provide insight into performance"[17] meaning they are accurate representations of overall output with use for their neural networks. I decided averages would be appropriate numbers to use for my research through this article's emphasis on their importance and based on their success doing the same.

To gain an idea of the current state of basketball statistical analysis, we look at Modelling the scores and performance statistics of NBA basketball games, a journal from 2018 offers takes on approaches to predicting NBA victories. The work done in the journal is the creation of a betting strategy [22], so their research interests bare no similarities. The reason I bring this journal up in my literature review, is the author's mass use of dated research. In their model, they use "performance statistics"[22], such as effective field goal percentage as well as others. The main features they use in their model are cornerstone features of Kubatko et al's 2007 research on sports analytics [**?**]. This is not a flaw, instead it shows that past research has stood the test of time still remaining relevant to present research.

After summarizing appropriate and relevant literature, I developed the research question: are there predictors of points-per-game that are correlated more with healthy or injured players?

2 METHODS

To begin my research, I created visualizations that were distributions of the variables I wanted to use as my predictors in my regression (minutes per game, age, and games played). My reasoning for this is one assumption of linear regression is that all variables must be multivariate normal. While the distributions of minutes per game and age appeared bell-shaped and normal [Figure 4,5], the distribution of games played did not [Figure 6]. These

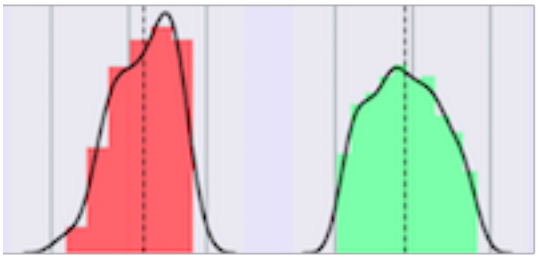

Figure 4: Minutes per game distribution

distributions served as visual aids to understanding my data sets. Visual data in my research is extremely important. Because each data point is a player in the NBA, if there is, for example, an outlier in blocks well-above the mean, the player's wingspan and height could be examined as causes for the extremity in the data. This is also unfortunately, a perfect opportunity for bias to enter the picture. Because a viewer of the NBA has their own opinions of player's, those perceptions could lead towards inaccurate predictions. I'll explore this theme more in the discussion section of this paper.

Table 1: Non-injured statistics

	μ	min	max	σ
PPG	7.59	0.00	30.40	5.76
Age	26.19	19	41	4.13
GP	42.37	1	82	28.83
MPG	18.10	0.50	36.90	9.43

Table 2: Injured statistics

	μ	min	max	σ
PPG	10.75	1.30	26.40	6.14
Age	26.17	19	39	4.26
GP	49	1	79	22.14
MPG	23.85	3.70	36.70	7.85

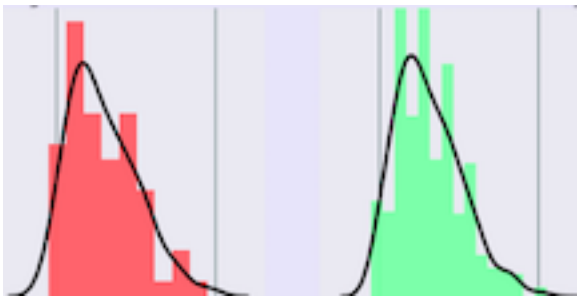

Figure 5: Age distribution

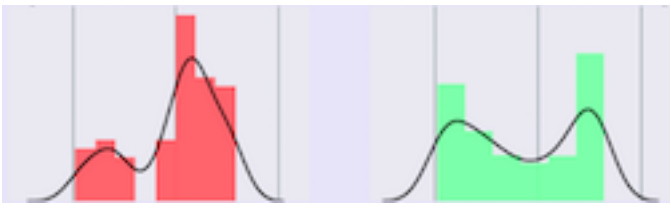

Figure 6: Games played distribution

From here I carried about my linear regression testing without using games played as a predictor variable. As the goal of my research is to determine the strength of relations between points scored and other factors, I was interested in examining coefficients

of determination, or r^2. I wanted to know both the strength of the relationship and whether or not these relationships differed between injured and healthy players.

Table 3: r^2 values

	Non-injured	*Injured*
	r^2	r^2
age	.6333	.7604
mpg	.9116	.9173

3 ANALYSIS

With my methods being explained, it is now important to examine their effect on my data. Beginning my analysis of my data, I found it important to consider all variables and their attributes. For example, in Tables 1 and 2, I noticed differences between means, minimums, and maximums. Already some differences between the two groups are noticeable. The difference in means in points per game could be explained by the fact that the more time you spend on the court, the more likely you are to be injured. From this fact, the more time you spend on the court highly correlates to the amount of points you score in a game. This helps in explaining the difference of 3.16 between the two categories.

Injured player's maximum points per game of 26.40, only 4.00 ppg less than the scoring leader for that season raising some interesting questions. One could see that statistic and interpret it to mean that injury has no effect on a players performance when they return. However, because my data set includes injuries of all kinds and not only specifically career-altering injuries, that interpretation would be invalid.

One statistic that surprised me was injured players having an average of 49 games played. One aspect of the data should be considered here. Since the data for injuries does not include injuries before the All-Star break, it is important to clarify some data. The length of the 2017-2018 season was 167 days. The data for injuries was collected past February 18th, the 115^{th} day of the season. Meaning an average of 56 (56.46) games were played by each team. The minimum of games played being 1 is interesting and unfortunate because it means someone was injured on their first game of the season 68% of the way into the season. The large standard deviations are explainable by the fact that minutes are not uniformly distributed amongst players.

4 RESULTS

The results of my linear regression testing and overall data analysis provided me with an answer to my research question. Based on Figures 7 and 8, some visual insights are revealed.

The resultant best fit lines for Figure 7 show a positive linear relationship between points per game and minutes per game,

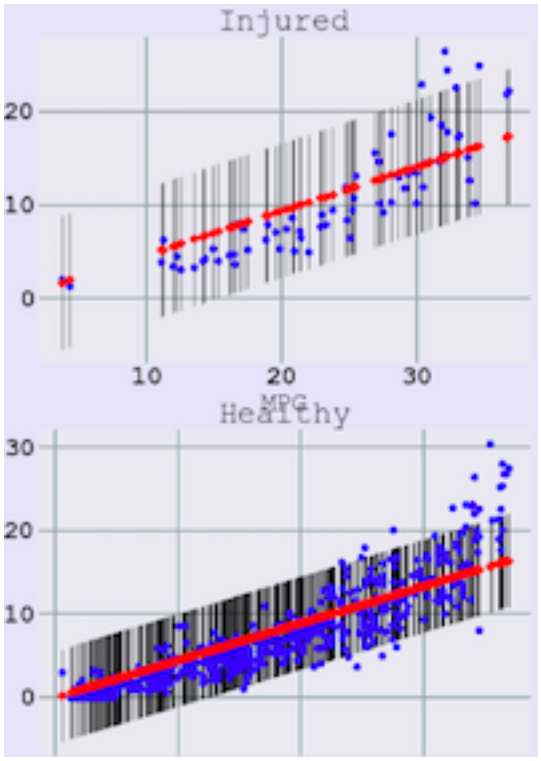

Figure 7: MPG Linear Regression

as expected. The r^2 values seen in Table 3 are very high meaning minutes per game is a strong predictor of points per game. This result is not especially significant to me as it reveals nothing new. However, the difference of r^2 between the two categories is interesting as it is only a difference of .0057. Therefore to answer my research question, there is no significant difference in minutes per game's effect on points per game between injured and healthy players.

Figure 8 visualizes the linear regression of age on points per game. The resultant visualizations reveal scattered plots of data. No groups are apparent in the data. Again from Table 3, the r^2 values are significant, but not as significant as minutes per game. The values of .6333 and .7604 are interesting results, mainly because they differ across the categories significantly.

5 DISCUSSION

Before discussing my results and overall research, I first want to address a limitation on my data set previously mentioned. Had my data set contained all injuries for the 2017-2018 season, my results would have probably been much different. With that as the backdrop upon which my findings are presented upon, I bring up my results.

My resultant r^2 values do indicate that age and minutes per game are solid predictors of points per game. However, the

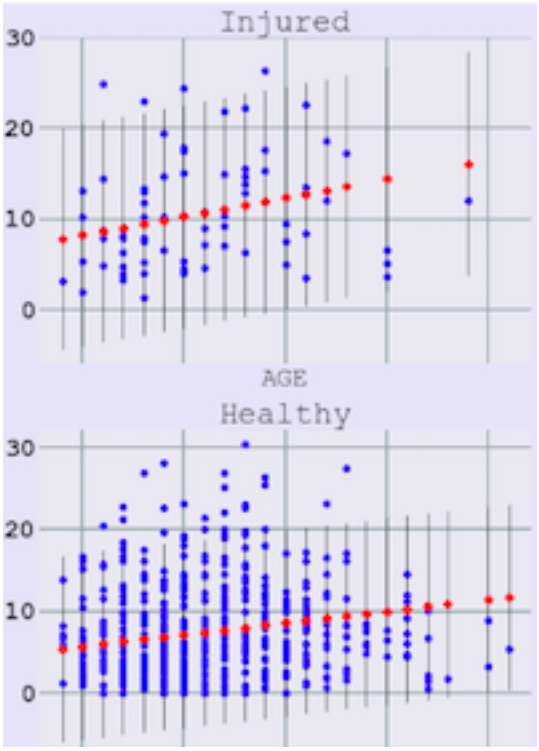

Figure 8: Age Linear Regression

result predictor for minutes per game is not necessarily surprising. As I have mentioned before, the more minutes you spend on the floor, the more chances you have to score, so it seems valid that both injured and healthy players had an $r^2 > .90$. I was surprised by the results for age and how strong the predictor was. This probably is a result of age and points per game having a rather weak linear relation. As seen in Figure 8, the regression plots were randomly scattered.

As I mentioned in my Methods section, bias players a significant role in sports analytics. One way to avoid bias from this kind of research would be to remove the player's names from all data sets. The reason I did not do this in my research was because I used the player's names to categorize them as either injured or not injured. The problem names of players bring are researchers, readers, or fans preconceived opinions about a player. Back to my example that I mentioned earlier about an outlier in blocks per game. If said player associated with that statistic has an abnormally long wingspan, people may look to that as an easy explanation for the outlier. However, there is more context beyond basketball statistics that make them almost dangerous to work with. With basketball statistics, an explanation for why a variable is what it is, is a lot more in depth than simply stating, the player is taller and that is why he has a large amount of blocks.

I now propose another example of how simple explanations can distort our understanding of basketball statistics and their significance. An assist is basketball is awarded to the player who passed the ball leading to a made basket. Separating basketball from ice hockey, an assist is only awarded to one player. As in the NHL (National Hockey League), passes leading to assists are also recorded as assists. So let's imagine that Player A in the NBA averages 1 assist per game, but off the record averages 12 "hockey assists", while Player B averages 4.4 assists per game with zero "hockey assists". Upon first inspection, the box score makes the better passer/play maker out to be Player B, while Player A is the actually, on average, leading to more baskets made. This is why analyzing what box score statistics mean is extremely important to this type of research. To summarize my ideas from this discussion on interpreting basketball statistics I state that a player's performance and overall contribution to the game is not always measured by the box score statistics.

These results and following discussion will certainly shape my research in the future. Kubatko et. al described the qualitative analysis of sports as a "melting pot for ideas"[15]. A point to which I can not disagree. As NBA statistics interest me, I plan to continue searching for trends in the data even after this class ends. One question that I should consider addressing regards what makes a box score statistic valid for statistical research. For example, as I learned from my research, minutes per game has a strong positive correlation with points per game. But this is not a significant finding because it is only proof of a rather common assumption. I am curious as to what a significant finding from NBA statistics would look like. Had I approached my research question with this is mind, my predictors chosen may have been different, more statistically determined numbers such as player efficiency rating, abbreviated as PER. Creator of the formula said "PER sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance"[20] which obviously is a much more complex calculation than age. Having a better understanding of that would have probably resulted in my research having more interesting findings.

6 CONCLUSION

The initial goal of my research was to identify differences in predicting points per game between healthy and injured players. Overall, my goal was achieved as I found that age is a stronger predictor for points per game for injured players. Minutes per game was found to be an equal predictor for points per game across both of the categories. To conclude my research, I think it is important to focus on what I could have done to improve or produce higher quality research.

Improvements. One aspect of my project where I could have improved was my choice of data sets. At the time of my data analysis, where I was producing visualizations and beginning to conduct linear regressions, my data set had no flaws apparent to me. However, during my analysis of my results and while writing this report, I began to notice room for error to enter the picture. The main takeaway I hope to learn from my data collection for this project is

determining what the data set is accurately representing. With my subsequent analysis, I noticed how having a data set for injuries beginning in February heavily limits my research. The main flaw is that players injured from the beginning of the season to the All Star Break in February were not included in my data set. The problem with this is that these are the players I should have been most interested in, as assuming their injuries were not season or career ending, had season statistics that were probably impacted by their injury. With this being said, what I hope to improve upon in my future research is my process of collecting data and determining significance and importance of the data set prior to analysis.

Overall, I am looking back on this project, noticing my flaws and look to make future improvements. This project has certainly been a learning experience, and I feel that what I've learned will shape my future work for the better.

REFERENCES

[1] Henry Abbott. 2011. The real story of Manute Bol. http://www.espn.com/blog/truehoop/post/_/id/31423/the-real-story-of-manute-bol

[2] Anonymous. 2017. Oregon lands 7-foot-2 high school basketball star Bol Bol. http://www.maxpreps.com/news/1ba8OF4o1Uihm1PNxjANYw/oregon-lands-7-foot-2-high-school-basketball-star-bol-bol.htm

[3] Anonymous. 2018. http://www.espn.com/nba/draft/rounds

[4] Anonymous. 2018. 2017-2018 NBA season. https://en.wikipedia.org/wiki/2017%E2%80%9318_NBA_season

[5] Anonymous. 2018. 2018 NBA All-Star Game. https://en.wikipedia.org/wiki/2018_NBA_All-Star_Game

[6] Anonymous. 2018. Bol Bol's High School Basketball Stats. http://www.maxpreps.com/athlete/bol-bol/TrlmDd-dEeaT-Oz0u-e-FA/gendersport/basketball-stats.htm

[7] Anonymous. 2018. NBA Season Leaders. https://stats.nba.com/leaders/

[8] Tommy Beer. 2018. An Early 2019 NBA Mock Draft: Duke's Barrett and Reddish On Track To Go 1-2. https://www.forbes.com/sites/tommybeer/2018/09/12/early-2019-nba-mock-draft/#35b97eef19eb

[9] Jesse C. Delee Eduardo Gomez and William C. Farney. 1996. Incidence of Injury in Texas Girls High School Basketball. , 684-687 pages. https://doi.org/10.1177/036354659602400521

[10] Adam M. Gonzalez, Jay R. Hoffman, Joseph P. Rogowski, William Burgos, Edwin Manalo, Keon Weise, Maren S. Fragala, and Jeffrey R. Stout. 2013. Performance Changes in NBA Basketball Players Vary in Starters vs. Nonstarters Over a Competitive Season. *Journal of Strength and Conditioning Research* 27, 3 (2013), 611âĂŞ615. https://doi.org/10.1519/jsc.0b013e31825dd2d9

[11] Peter A. Harmer. 2005. Basketball Injuries. *Epidemoiology of Pediatric Sports Injuries Medicine and Sport Science* (2005), 31–61. https://doi.org/10.1159/000085341

[12] H.Bates Noble John A. Zelisko and Marianne Porter. 1982. A comparison of men's and women's professional basketball injuries. , 297-299 pages.

[13] Sheri L. Walters John R. Deitch, Chad Starkey and J.Bruce Moseley. 2006. Injury Risk in Professional Basketball Players. , 1077-1083 pages. https://doi.org/10.1177/0363546505285383

[14] Andrew Klein J.Philip Shambaugh and John H. Herbert. 1991. https://doi.org/10.1249/00005768-199105000-00003

[15] Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan T Rosenbaum. 2007. A Starting Point for Analyzing Basketball Statistics. *Journal of Quantitative Analysis in Sports* 3, 3 (Sep 2007). https://doi.org/10.2202/1559-0410.1070

[16] Sarah K. Fields Laurel A. Borowski, Ellen E. Yard and R.Dawn Comstock. 2008. The Epidemiology of US High School Basketball Injuries, 2005âĂŞ2007. , 2328-2335 pages. https://doi.org/10.1177/0363546508322893

[17] Bernard Loeffelholz, Earl Bednar, and Kenneth W Bauer. 2009. Predicting NBA Games Using Neural Networks. *Journal of Quantitative Analysis in Sports* 5, 1 (2009). https://doi.org/10.2202/1559-0410.1156

[18] G.D. Mckay. 2001. Ankle injuries in basketball: injury rate and risk factors. *British Journal of Sports Medicine 35* (2001), 103–108. https://doi.org/10.1136/bjsm.35.2.103

[19] A.J. Neuharth-Keusch. 2017. NBA big men at heart of three-point shooting revolution. https://www.usatoday.com/story/sports/nba/2017/02/16/nba-three-point-shooting-cousins-gasol-lopedavis/97248784/

[20] Basketball Reference. [n. d.]. Calculating PER. https://www.basketball-reference.com/about/per.html

[21] Statistics Solution. [n. d.]. Assumptions of Linear Regression. https://www.statisticssolutions.com/assumptions-of-linear-regression/

[22] Kai Song, Qingrong Zou, and Jian Shi. 2018. Modelling the scores and performance statistics of NBA basketball games. *Communications in Statistics - Simulation and Computation* (2018), 1âĂŞ13. https://doi.org/10.1080/03610918.2018.1520878

[23] Chad Starkey. 2000. Injuries and Illnesses in the National Basketball Association: A 10-Year Perspective. , 161-167 pages.

[24] J.E. Taunton. 2002. A retrospective case-control analysis of 2002 running injuries. , 95-101 pages. https://doi.org/10.1136/bjsm.36.2.95

[25] Matthew Verhey. 2018. How Important is Height and Weight in Staying Healthy in the NBA?