RESEARCH ARTICLE

COVID-19 countermeasures, Major League Baseball, and the home field advantage: Simulating the 2020 season using logit regression and a neural network [version 1; peer review: 3 approved with reservations]

Justin Ehrlich ¹, Shankar Ghimire ²

¹Department of Sport Analytics, Syracuse University, Syracuse, NY, 13244, USA

²Department of Economics and Decision Sciences, Western Illinois University, Macomb, IL, 61455, USA

V1 First published: 20 May 2020, 9:414
<https://doi.org/10.12688/f1000research.23694.1>
 Latest published: 20 May 2020, 9:414
<https://doi.org/10.12688/f1000research.23694.1>

Abstract

Background: In the wake of COVID-19, almost all major league sports have been either cancelled or postponed. The sports industry suffered a major blow with the uncertainty of sporting events being held in the near future. Various scenarios of how and when sports might recommence have been discussed. This paper examines various scenarios of how Major League Baseball team performance is going to be impacted by the presence of fans, or the lack thereof, in the context of physical distancing and other COVID-19 countermeasures

Methods: The paper simulates, using a neural network and a logit regression model, the win-loss probabilities for various scenarios under consideration and also estimates the home effect for each team using data for the 2017-2019 seasons.

Results: The model demonstrates that individual team home effect is symmetric between home and away and teams will not necessarily have a win or loss of any additional games in neutral stadiums, as teams with a high home field effect will lose more neutral games that would have been at home but will win more neutral games that would have been away. However, the result of individual games will be different since home effect is asymmetric between teams. Our simulation demonstrates that these individual game differences may lead to a slight difference in Play-Off Berths between a full season, a half season, or a full season without fans.

Conclusions: Without fans, any advantage (or disadvantage) from home field advantage is removed. Our models and simulation demonstrate that this will reduce the variance. This stabilizes the outcome based upon true team talent, which we estimate will cause a larger divide between the best and worst teams. This estimation helps decision makers understand how individual team performance will be impacted as they prepare for the 2020 season under the new

Open Peer Review

Approval Status ? ? ?

	1	2	3
version 1			
20 May 2020	view	view	view

1. **Garry Kuan**, Brunel University, Uxbridge, UK

2. **Ismail Dergaa** , Primary Health Care Corporation (PHCC), Qatar, Qatar

3. **Federico Fioravanti** , Universiteit van Amsterdam, Amsterdam, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

circumstances.

Keywords

MLB, Baseball, COVID-19, Neural Network, Logit

This article is included in the **Artificial Intelligence and Machine Learning** gateway.

This article is included in the **Coronavirus** collection.

Corresponding author: Justin Ehrlich (jaehrlie@syr.edu)

Author roles: **Ehrlich J:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Ghimire S:** Formal Analysis, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by funds provided by the David B. Falk College of Sport and Human Dynamics, Syracuse University.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Ehrlich J and Ghimire S. This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ehrlich J and Ghimire S. **COVID-19 countermeasures, Major League Baseball, and the home field advantage: Simulating the 2020 season using logit regression and a neural network [version 1; peer review: 3 approved with reservations]** F1000Research 2020, 9:414 <https://doi.org/10.12688/f1000research.23694.1>

First published: 20 May 2020, 9:414 <https://doi.org/10.12688/f1000research.23694.1>

Introduction

The 2019–2020 pandemic from the novel coronavirus (COVID-19) has brought unprecedented countermeasures to every sector of the economy, including individuals, groups, institutions, and industries. The sports industry took one of the biggest hits, with all major leagues in the U.S. cancelling or halting their events. While these actions were necessary to address the public health concern, each segment is now floating various proposals to resume operations and give some relief to the significant portion of the economy that the sports industry comprises.

Major League Baseball (MLB) is likely to be the first American professional sporting league to resume, probably in May or June (Passan, 2020). *Players are willing (although not all players agree to the method), the League is willing, the Arizona government is willing, and health professionals have approved a plan* to move forward, known as the *Arizona Plan*. This plan calls for players, coaches, and staff to be quarantined in hotels around the Phoenix area, and to play in empty ballparks that include the ten Cactus League Spring Training parks, Chase Field, and other Phoenix ballparks. One interesting aspect of this arrangement is that the stadium size will not matter because there will not be any fans. This will be an opportunity for MLB to get back into the spotlight and accumulate massive television viewership that MLB has not seen in decades. The experience will be completely optimized for TV viewing, and so the league will finally be able to experiment with proposed rule changes, including removing mound visits to make the game go faster, adding a *Robo Umpire*, which has already been successfully tested last season via a partnership with the independent Atlantic League (Bogage, 2019), and an expanded roster giving players more rest due to the extremely hot temperatures of Phoenix. While all of this will alter predictions on who's going to the playoffs, probably the biggest impact that this plan will have on the games is the *lack of the home field advantage (HFA): the advantage that the home team has over the visiting team* due to the home team having fans, the familiarity of the home team to their own ball park, and the away team having to travel.

Baseball has been shown in previous studies to be less susceptible to the HFA effect than other professional sports (Edwards & Archambault, 1979; Gómez *et al.*, 2011; Pollard *et al.*, 2017). Despite this, there is a measurable home field advantage in baseball, as shown by Jones (2015); Jones (2018). Building on this, we extend the analysis for the MLB under uncertainty of which scenario the League will be following for the 2020 season. In particular, we simulate the win-loss probabilities for three different scenarios as well as estimate the home advantage for each team using the past three seasons' data. This estimation helps us understand how individual team performance is going to be impacted as they prepare for the 2020 season in the new circumstances.

Methods

Data sources

We use the MLB 2017–2019 season data for the 30 teams represented in the league. The data were obtained from the MLB

Advanced Media's *Baseball Savant Website* using the Python package *PyBaseball* 1.0.4 (LeDoux, 2017/2020). The data shows that out of the 7,290 home games played during the 2017–2019 seasons, 3,881 (53.237%) resulted in wins and the remaining 3,409 (46.763%) resulted in a loss. Next, we seek to quantify the HFA's role in this difference.

Calculating home advantage

There are various techniques to calculate the home advantage depending on the sport, gender, league, and the nature of scoring (Jones, 2015). Pollard *et al.* (2017) use a general linear model to fit the home advantage. However, because we have a categorical variable of win or lose, we need to follow a non-linear approach. To test the hypothesis that teams have home-field advantage, we apply a logit regression model to predict the probability of winning as a function of home game dummy, team fixed effects, opponent fixed effects, and the win-loss records. We estimate the following regression equation:

$$Win_i = \alpha_0 + \alpha_1 Home_i + \alpha_2 Team_i + \alpha_3 Home \times Team_i + \alpha_4 Home \times Opp_i + \eta Z_i + \varepsilon_i \quad (1)$$

where Win_i is a dummy variable that takes the value 1 if the recorded game resulted in a win for the team-opponent pair, and zero otherwise; $Home_i$ accounts for home game; and $Team_i$ controls for the individual team fixed effects, and Opp_i controls for the opponent fixed effects, Z_i represents the win-loss percentage for the team as well as the opponent, ε_i stands for the error term. The HFA is calculated accounting for the team fixed effects as well as the opponent fixed effects by interacting home with team and opponent separately. We run the logit model on all the data, with $Home=1$ for the home team and equal to zero for the travelling team. Doing so separates the team fixed effects and home field advantage. The model in Equation 1 is used to estimate both the win probability and the HFA per team. The HFA is obtained by calculating the marginal effect (ME) of $Home$ on the win probability for each team separately.

Development of the neural network model

A neural network model was also created to act as a robustness check for the logit win prediction model. The software to train the model is hosted on *GitHub* (Ehrlich, 2020a). We used the R package *nnet* 7.3–14 (Ripley & Venables, 2020) for the neural network platform, and trained and tuned the model with the R package *caret* 6.0–86 (Kuhn *et al.*, 2020). We developed a simulator to estimate what might happen if: 1) The full 2020 season continued on in a parallel universe devoid of COVID-19; 2) MLB waits and is able to return and play half a season to packed stadiums around the All Star Break, which is assuming an extremely optimistic timeline of a return to normal life; 3) a full season is played without fans, which is likely the only way they will be able to play this season (i.e., the Arizona Plan). The simulation was executed 100 times, and the logit win prediction model was used as the basis for predicting each win. A random number between 0 and 1 was generated and checked against the win probability provided by the model. If the random number was below the probability, then the team won, otherwise the team lost.

Results

The summary statistics of the training data is contained in [Table 1](#). The logit results from [Equation 1](#), without the fixed effects, are reported in [Table 2](#). Both log odds ratios and the MEs are reported in this table. The results show that the individual regressors included in the model show plausible impacts. Looking at the log-odds ratios, home games and the home team's previous win-loss percentage (WL%) are more likely to result in a win but the opponent's WL% is less likely to result a loss for the home team. These results support the presence of the HFA. The right half of the table shows the MEs for each variable. We are mainly interested in the MEe for the *Home* variable, which is 0.064. This means, the marginal probability of winning a game at home versus away field goes up by 6.4%. This is the average HFA for all of the teams as a whole. The HFA for each team is presented in [Figure 1](#). In our sample, PHI seems to have the highest home advantage and HOU seems to have the lowest (negative, in fact) home advantage.

The model was trained using the 2017–2019 MLB regular season games. The schedule for the 2020 season was estimated using the schedule from the 2019 season. While the dates will be off slightly, the team pairings will be nearly the same. The wins and losses of the 100 simulations were added to form the result of the 2020 season. The overall results are visualized in

[Figure 2](#), while the divisional results shown in [Table 3](#). [Table 4](#) provides statistics calculated during each season and averaged. This includes the correlation between the full season and the half and no-fan seasons using both the overall rankings and the win-loss percent (WL%). The full seasons rank correlations are higher with the no-fans seasons (0.825) than the half seasons (0.735). The correlations using WL% is similar. The standard deviation of the predicted win probabilities is lower for the no-fans seasons (0.073) than the full (0.085) and half seasons (0.085). The home effect was correlated with the win probabilities' standard deviations and is negative for the no fans seasons (-0.221). In other words, the higher the home effect, the lower the variance.

This neural network was also used as the win predictor in 100 simulations and the results are very similar to the logit win prediction model, which shows robustness in the simulation results. [Table 4](#) shows the statistical results of both models. The correlation and standard deviation differences are approximately the same between the two models.

The results of the simulations are available as *Extended data* ([Ehrlich, 2020b](#)) and the code necessary for replicating the results, including training the models, are hosted on [GitHub](#) ([Ehrlich, 2020a](#)).

Table 1. Summary statistics used for training the model.

Variable	Mean	SD	Min	Median	Max
Home	0.500	0.500	0.000	0.500	1.000
Prev WL %	0.500	0.114	0.000	0.500	1.000
PrevWL% Opp	0.500	0.114	0.000	0.500	1.000
Season	2018.000	0.816	2017.000	2018.000	2019.000

Table 2. Results of regression analysis.

Predictors	Logit win prediction model			Logit win prediction marginal effects		
	Odds ratios	Std. error	p	AME	Std. error	p
(Intercept)	0.876	0.174	0.447			
Home	1.304	0.034	<0.001	0.064	0.008	<0.001
PrevWLPerc	2.307	0.167	<0.001	0.203	0.040	<0.001
PrevWLPercOpp	0.433	0.167	<0.001	-0.203	0.040	<0.001
WLPercSeasonPrev	8.251	0.245	<0.001	0.513	0.060	<0.001
WLPercSeasonPrevOpp	0.121	0.245	<0.001	-0.512	0.060	<0.001
Observations	14580					
R ² Tjur	0.029

Figure 1. MLB home field advantage effect of individual teams.

Figure 2. MLB season 2020 change in simulated rank after 100 simulations.

Table 3. Results of the simulation using the logit win prediction model.

		No fan season			Full season			Half season			
Division	Tm	Rank	Berth	WL%	Rank	Berth	WL%	Rank	Berth	WL%	Home effect
AL Central	CLE	1	y	0.593	1	y	0.587	1	y	0.584	0.034
AL Central	MIN	2	w	0.573	2		0.565	2	w	0.572	0.033
AL Central	CHW	3		0.419	3		0.414	3		0.425	0.065
AL Central	KCR	4		0.390	4		0.397	4		0.402	0.062
AL Central	DET	5		0.337	5		0.340	5		0.341	0.050
AL East	NYN	1	y	0.623	1	y	0.614	1	y	0.607	0.115
AL East	BOS	2		0.569	2	w	0.573	3		0.568	0.003
AL East	TBR	3		0.567	3		0.564	2	w	0.582	0.069
AL East	TOR	4		0.438	4		0.435	4		0.408	0.072
AL East	BAL	5		0.347	5		0.349	5		0.357	0.092
AL West	HOU	1	y	0.653	1	y	0.650	1	y	0.657	-0.014
AL West	OAK	2	w	0.580	2	w	0.576	2		0.566	0.115
AL West	SEA	3		0.473	3		0.468	3		0.491	0.016
AL West	LAA	4		0.469	4		0.467	4		0.462	0.056
AL West	TEX	5		0.463	5		0.454	5		0.436	0.070
NL Central	MIL	1	y	0.565	1	y	0.555	1	y	0.553	0.071
NL Central	CHC	2	w	0.544	2	w	0.545	2		0.550	0.118
NL Central	STL	3		0.534	3		0.542	3		0.544	0.056
NL Central	PIT	4		0.450	4		0.451	4		0.464	0.084
NL Central	CIN	5		0.437	5		0.445	5		0.440	0.099
NL East	WSN	1	y	0.569	1	y	0.565	2	w	0.556	0.016
NL East	ATL	2	w	0.559	2	w	0.559	1	y	0.558	-0.005
NL East	NYM	3		0.494	3		0.485	3		0.505	0.045
NL East	PHI	4		0.475	4		0.477	4		0.479	0.163
NL East	MIA	5		0.394	5		0.393	5		0.392	0.107
NL West	LAD	1	y	0.632	1	y	0.632	1	y	0.634	0.081
NL West	ARI	2		0.540	2		0.539	2	w	0.553	0.049
NL West	COL	3		0.500	3		0.499	3		0.488	0.092
NL West	SFG	4		0.437	4		0.441	4		0.437	0.071
NL West	SDP	5		0.432	5		0.423	5		0.402	0.050

Table 4. Key summary statistics of the simulation using the logit win prediction model.

Model simulated	Logit	NN
Simulated Seasons	100	100
Full-NoFans Rank Correlation	0.825	0.814
Full-Half Rank Correlation	0.735	0.719
Full-NoFans WL% Correlation	0.823	0.817
Full-Half WL% Correlation	0.734	0.718
NoFans WL% SD	0.073	0.073
Full WL% SD	0.085	0.086
Half WL% SD	0.085	0.086
Full WinProb SD-HomeEffect Correlation	0.361	0.251
NoFans WinProb SD – HomeEffect Correlation	-0.221	-0.268

Note: These statistics are calculated for each season and averaged. WL% is predicted based upon the Win Prediction.

Discussion

Based on the above results, since the team-home effect is symmetric between home and away, teams will not necessarily win or lose any additional games in neutral stadiums as teams with a high home field effect will lose more neutral games that would have been at home but will win more neutral games that would have been away. The greater the home-team ME, the less variance there will in of the predicted win probabilities. To verify this assumption, we calculated the correlation of HomeEffects and the standard deviation (SD) of win probabilities between a full (0.361) and no-fan season (-0.221). Since the home effect is symmetric for each team (the away field disadvantage = -the home field advantage), decreasing the variance does not affect the overall expected WL% for each team. However, the result of individual games will be different since home effect is asymmetric between teams. For example, if the Cubs (highest home effect in the NL Central) plays the Cardinals (lowest home effect in the NL Central), the Cubs will have a larger advantage playing at home then the Cardinals will have playing at home (besides team fixed effects). These differences are removed with the No-Fan scenario and the outcome will be solely based upon the talent of the teams. However, on average there only a slight change of overall WL% (or playoff berth), just the SD of the results (see Table 3). Without fans, any advantage (or disadvantage) from home field advantage, which cause higher levels of variance, is removed. This stabilizes the outcome based upon true team talent. As fewer games have been played, the half-season will have more upsets, but the SD is close to the same as the full season.

Conclusion

This paper analyzes the previous season MLB data to estimate the win-loss probabilities for the 2020 season for each of the 30 teams in the League using logit regressions and a neural network. The *Arizona Plan*'s neutralization of HFA would not significantly affect the overall outcome of the season. In fact, our model predicts that the *Arizona Plan* season will produce season results that are based more on the true talent of the teams. Further, our simulation demonstrates that there

will be less variance in the win probability between any two teams, which we estimate will cause a larger divide between the best and worst teams. In conclusion, we believe that the results of the *Arizona Plan* will be similar to a regular season with fans, and that the teams' standings at the end of the regular season will be more predictable than a normal season.

Data availability

Source data

Zenodo: Syracuse-University-Sport-Analytics/MLBCovid19. <https://doi.org/10.5281/zenodo.3775959> (Ehrlich, 2020a).

This project contains the following source data files:

- data/2008_2019Games.csv. (Input data scraped using the scrapingMLB.ipynb.)
- data/divisions.csv. (Input team division data for grouping by division.)
- data/mlbTeamColors.csv. (Input team colors for the visualizations.)

Source data are also available on GitHub: <https://github.com/Syracuse-University-Sport-Analytics/MLBCovid19>.

Extended data

Harvard Dataverse: Replication Data for: COVID-19 Countermeasures, Major League Baseball, and the Home Field Advantage. <https://doi.org/10.7910/DVN/OOMWSD> (Ehrlich, 2020b).

This project contains the following extended data files:

- divisionRankings. (Results of simulation using the logit model.)
- divisionRankingsNN. (Results of simulation using the neural network model.)
- homeEffectLogit. (Team home effects using the logit model.)
- homeEffectNN. (Team home effects using the neural network model.)
- modelCorrelationsSummaryWithNNResults. (Simulation statistics from both the logit and neural network models.)

Zenodo: Syracuse-University-Sport-Analytics/MLBCovid19. <https://doi.org/10.5281/zenodo.3775959> (Ehrlich, 2020a).

This project contains the following source files:

- pythonStatcastScraper/scrapingMLB.ipynb. (Python Jupiter Notebook code for scraping Statcast.)
- halfSeasonPrediction.Rmd. (R Markdown Notebook code for developing the logit and neural network models. Also contains the code for running the simulations.)
- All the other data is intermediate output from the simulations. The important output files are located in the above Harvard Dataverse repository.

Source code is also available on GitHub: <https://github.com/Syracuse-University-Sport-Analytics/MLBCovid19>.

Mixed data and code hosted on GitHub and Zenodo are available under the terms of the [GNU General Public License v3.0](#).

Data hosted on Harvard Dataverse are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

References

Bogage J: **National Baseball Hall of Fame accepts Atlantic League ‘robo ump’**. items. Washington Post. 2019.

[Reference Source](#)

Edwards J, Archambault D: **The home field advantage**. *Sports, games, and play: Social and psychological viewpoints*. 1979; 409–438.

[Reference Source](#)

Ehrlich J: **Syracuse-University-Sport-Analytics/MLBCovid19: First Release (Version v1.0.0)**. Zenodo. 2020a.

<http://www.doi.org/10.5281/zenodo.3775959>

Ehrlich J: **Replication Data for: COVID-19 Countermeasures, Major League Baseball, and the Home Field Advantage**. Harvard Dataverse, V1, UNF:6: LiInpTKr15iER0wC31Bb9g== [fileUNF]. 2020b.

<http://www.doi.org/10.7910/DVN/OOMWSD>

Gómez MA, Pollard R, Luis-Pascual JC: **Comparison of the home advantage in nine different professional team sports in Spain**. *Perceptual and motor skills*. 2011; 113(1): 150–156.

[PubMed Abstract](#) | [Publisher Full Text](#)

Jones MB: **The home advantage in major league baseball**. *Percept Mot Skills*.

2015; 121(3): 791–804.

[PubMed Abstract](#) | [Publisher Full Text](#)

Jones MB: **Differences in home advantage between sports**. *Psychol Sport Exerc*. 2018; 34: 61–69.

[Publisher Full Text](#)

Kuhn M, Wing J, Weston S, et al.: **caret: Classification and Regression Training (Version 6.0-86) [Computer software]**. 2020.

[Reference Source](#)

LeDoux J: **Jldbc/pybaseball [Python]**. 2020; (Original work published 2017).

[Reference Source](#)

Passan J: **Sources: MLB, players eye May return in Arizona**. ESPN.Com. 2020.

[Reference Source](#)

Pollard R, Prieto J, Gómez MÁ: **Global differences in home advantage by country, sport and sex**. *Int J Perform Anal Sport*. 2017; 17(4): 586–599.

[Publisher Full Text](#)

Ripley B, Venables W: **nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models (Version 7.3-14) [Computer software]**. 2020;

[Reference Source](#)

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 28 March 2023

<https://doi.org/10.5256/f1000research.26143.r164101>

© 2023 Fioravanti F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Federico Fioravanti

Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam, North Holland, The Netherlands

General comments

The work examines three possible scenarios for the Major League Baseball 2020 season, motivated by the fact that due to Covid-19 countermeasures, there will be no presence of fans. The topic is interesting, but the authors should work more on describing the motivation and the results they find.

Specific comments

Abstract: Line 13 - first sentence is too long.

Introduction: It is worth noticing that Covid-19 was a worldwide problem. So it is preferable to start telling that, and then say the study will focus on a particular sport in a particular country. A bit of effort should be done to explain the baseball terminology (or reduce its use), in order to facilitate the understanding for people from countries where baseball is not mainstream. Even how the Baseball environment is, requires an explanation, so it is easier to interpret the results. Use papers such as Schwarz and Basky (1977)¹ or Agnew and Carron (1994)² to expand the explanation of the HFA and the possible factors causing it.

Line 10 - it says: "*Major League Baseball (MLB) is likely to be the first American professional sporting league to resume*" and it should say: "Major League Baseball (MLB) is likely to be the first professional sporting league in the United States" (if it is the first one in the American continent then cite accordingly).

Methods: This section needs further development. The regression equation needs more explanation and why every variable is introduced in the equation. What does every variable mean? What is the range of every variable?

It is recommendable to describe a bit more the different simulations that are considered.

Results: For ease of understanding, there must be a description of the abbreviations that are used

in tables and figures (not the name of the teams). Some statistics are missing, such as sample size. Line 6 is difficult to understand. What do we look at? The log-odds ratio? The log-odds ratio, home games and WL%? What is less likely to result in a win?

Discussion: Why is the home effect symmetric for each team? If it is a hypothesis, it must be justified.

Line 7 - it says: "*the less variance there will in of the predicted*" and it should say: "the less variance will be in the predicted" or "the less variance of the predicted"

Line 18 - it says: "*at home then the Cardinals*" and it should say: "at home than the Cardinals"

Line 22 - it says: "*there only*" and it should say: "there is only"

There is a lack of discussion in this section, commenting on limitations, etc...

Conclusion: The conclusion leads me to think that there is not a real HFA, as the standings will be similar between a regular season and a no-fans season. The message in this section is not so clear.

References

1. Schwartz B, Barsky S: The Home Advantage. *Social Forces*. 1977; **55** (3): 641-661 [Publisher Full Text](#)
2. Agnew GA, Carron AV: Crowd effects and the home advantage. *International Journal of Sport Psychology*. 1994; **25** (1): 53-62

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Mathematics, Sports, Social Choice

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 28 March 2023

<https://doi.org/10.5256/f1000research.26143.r164098>

© 2023 Dergaa I. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ismail Dergaa

Primary Health Care Corporation (PHCC), Qatar, Qatar

The aim of this study is to examine various scenarios of how Major League Baseball team performance is impacted by the presence or absence of fans in the context of physical distancing and COVID-19. While the article is well-structured and falls within the scope of Psychology Research and Behavior Management, it requires major revisions before it can be accepted for indexing.

1. The duplication of keywords with the title needs to be adjusted to increase the visibility of the manuscript. Keywords should not be duplicated with the title. For example, if COVID-19 is in the title and the keywords, it should be replaced with SARS-CoV-2.
2. The biggest concern is that the article presents a simulation of the 2020 season using logit regression and a neural network, while it is now 2023. To address this, the authors should add a section with the actual results of the season and compare them to the study's outcome. They should keep the analysis as it is and change the aim of the study. This will provide two aims for the study, the actual one and an assessment of the accuracy of the simulation.
3. The manuscript needs to be rewritten in past tense as we are in 2023.
4. In addition to the analysis, the authors should compare their results with the actual outcomes of the season.
5. The references need to be updated.
6. The methodology is not clear, and the authors are requested to add a flowchart explaining the study protocol to make it easier for the reader.
7. For the first time in my life, I have seen a discussion section without any references. The authors need to compare their results with similar articles. Additionally, the authors need to mention that sports team performance is a complicated component that depends on several factors, such as players' mentality, team play, motivation, injury of key players, etc. These points need to be supported by credible references to strengthen the arguments presented in the discussion and conclusion section. Special emphasis should be given to the limitation section, which is absent in the study, to provide a more comprehensive analysis of the study's scope and potential impact.

Overall, the manuscript has potential, but it requires significant revisions before it can be accepted for indexing.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sports Medicine and exercise science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 10 June 2022

<https://doi.org/10.5256/f1000research.26143.r140121>

© 2022 Kuan G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Garry Kuan

Brunel University, Uxbridge, UK

GENERAL COMMENTS

The aim of this paper was to examine various scenarios of how Major League Baseball team performance is going to be impacted by the presence of fans, or the lack thereof, in the context of physical distancing and other COVID-19 countermeasures. Although this article addresses an interesting topic, some issues should be addressed before indexing.

SPECIFIC COMMENTS

INTRODUCTION

The introduction needs major revision and clarification. First, the aim of the study is not clear. The manuscript should answer a specific research question, which is lacking in this study. Also, the way the authors build up their introduction does not lead to the research question. The introduction is too shallow, and the context of the manuscript is not well-elaborated. Thus, those who are not living in the USA might have no clue about what the authors are explaining. I suggest that the authors should re-structure their introduction, explaining why their research is important and how is it related to COVID-19.

It is recommended that the authors expand this part: What are the environmental factors, besides the presence of the fans, that might affect the players' performance. More importantly, this should lead to a clear research question.

METHODS

The methods section needs major revision. As it stands, it is not possible to replicate their study. Firstly, COVID-19 started in early 2020 and if data from the 2017-2019 seasons were used it could not represent the COVID-19 situation.

Why is only home advantage calculated? Please explain.

What is the sample size, and what is the effect size?

RESULTS

The results section seemed okay. Probably, including a research question would help the authors to structure their results. I think the authors put too much information and abbreviation in the tables, but no explanation of the abbreviation, making the entire manuscript hard to follow. Figure 2 needs more clarification.

DISCUSSION

The discussion is only a paragraph - the authors should further discuss their findings and the implication of these findings in more depth. Is there any limitation of this study? Suggest revising this section.

Thank you.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Exercise and sports psychology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research