Analysis of the causal effect of cardiac arrest on the death age of football players

Erica Zhou - 1005687678

December 17, 2021

Abstract

This is a report about the death age of football players and cardiac arrest. The analysis focuses on the estimation of the effect that a cardiac arrest would have on the age of death of football players. The data is collected from Kaggle, and it provides basic information about the deaths of the players. The main method used in the project is propensity score matching, which is often used to reduce or eliminate bias by balancing the characteristics of participants. It turns out that a cardiac arrest causes football players to die younger than other common deadly illnesses. Football players and associations should make more effort on monitoring and the emergency procedures when the problems happen.

Keywords: Football, Cardiac health, Observed data, Casual inference, Propensity Score Matching.

Introduction

When people talk about sports, it is always mentioned that sports are good for health, body shape, and longevity. However, many sports can be risky. Football is one of the high-risk sports. Not like jogging, playing football has a high requirement of muscles, lungs, and energy, especially for professional athletes. Any collision or collapse can be fierce and deadly on the football field. Cardiac arrest is one of the deadly illnesses that frequently happens during football games or training. Although the International Federation of Association Football (FIFA) has applied pre-competition medical assessment to players, cardiac arrest is not effectively prevented. With the popularity of football, people should pay more attention to the health problems of football players behind the fierce matches.

In this project, I will analyze the impact of cardiac arrest on the age of death of football players. In other words, I want to see if cardiac arrest would cause a player to die younger compared to other illnesses such as heart attack and tetanus, which are also potential causes of death in a football game.

Literature

Soccer and Sudden Cardiac Death in Young Competitive Athletes: A Review

The paper of John P. Higgins and Aldo Andino points out that Sudden Cardiac Death (SCD) has been brought to public attention in the recent decade, and athletes have 2.5 times higher risk than others (Higgins and Andino, 2013). According to the paper, there is not any direct relationship between football and SCD, but SCD is usually caused by the increased cardiovascular demand, which usually occurs when doing sports. The authors indicate that there are various reasons for SCD such as CCA, ARVC (in the US), and LVH (in the UK). Since football players in different countries have different major cardiac abnormalities, it raises another potential topic about the reason and the propensity of death of football players due to cardiac arrest in different countries for further studies on this theme.

Christian Eriksen: What can cause a cardiac arrest?

The report published on June 15, 2020, focuses on some cases of cardiac arrest of football players. It points out that one of the most common causes of cardiac arrest is a life-threatening abnormal heart rhythm according to BHF. It introduced that CPR test has been applied on people from 16 to 25 years every two years to prevent potential abnormality, but it is not very effective because the problems barely show up in 16 to 25 and are thus not recognizable. Another important thing mentioned is an example that the player died because of a cardiac arrest after collapsing on the pitch. The relationship between collapse and cardiac arrest is a focus of this project as well.

Terminology

Cardiac arrest: Heart stops beating sunddenly.

Heart attack: The supply of blood and oxygen to heart stops.

Response: Dependent variable.

Predictor: Independent variable.

Treatment effect: The causal effect of a binary variable on an outcome variable.

Linear regression model: A model that reveals the linear relationship between the response and the predictor(s).

Propensity Score Matching: A statistical matching that attempts to estimate the treatment effect.

Hypothesis

As we all know, cardiac arrest is a deadly illness. It is hard to save people's life from a cardiac arrest because it always happens of a sudden and people have no time to react. At the same time, football is an intense and competitive sport, which increases the chance for a player to get a cardiac arrest. Hence, my prediction of mine is that cardiac arrest will cause football players to die earlier compared to other causes of death.

Sections

In section *Data*, some numerical and graphical summaries will be displayed through Exploratory Data Analysis (EDA). The methods used in the analysis will be introduced in section *Methods*. The main estimation results and interpretation will be displayed in section *Results*.

Data

Data Collection Process

The data was posted on Kaggle by Shivam Bansal, who also created the data. He collected these data by manually downloading and web scraping from Google News, Bing News, and Wikipedia. Because this data is collected and cleaned by an individual from the Internet, it is likely to contain selection bias and limited coverage. Therefore, we need to reduce the bias in the process of the analysis. Another potential problem would be the accuracy and reliability of the data. The data is not guaranteed to be corrected and thus the observations may have errors.

On Kaggle, the data has license CC0: Public Domain, which allows the use of the data. The first version of this data is lasted uploaded and updated on November 30th, 2021, and is expected to be quarterly updated.

Data Summary

The data is collected from Kaggle, and it was uploaded by Shivam Bansal on Nov 30, 2021. It describes the deaths of football players during a game or training. It includes the deaths due to either direct injuries or indirect illness. There are a total of 229 observations and 13 variables, including the incident date, the death age, the player information, and so forth. Variable "player_age" would be the response in this project, and it represents the death age of the player. Other potential predictors include "heart_related", "cardiac_related", "team_country", and so forth. Most of the predictors would be categorical variables.

Before the analysis, I dropped the missing values of the "player_age" variable. Some players' death ages were not collected or recorded, and thus, these observations should not be contained in the analysis.

Important variables:

Name	Defnition
player_name	The name of the player
team_country	The country of the football team
player_age	The age of death of the player
heart_related	Whether there was a heart related cause of death
cardiac_related	Whether there was a cardiac related cause of death
collapsed	Whether the player died from a collapse
lightning	Whether the player died from being struck by the lightning
collision	Whether the player died from a collision

Summary table

Variable	Mean	Standard Deviation
player_age	25.925	6.893
heart_related	0.355	0.48
cardiac_related	0.15	0.357
collapsed	0.463	0.5
lightning	0.028	0.165
collision	0.075	0.264

According to the summary table, the average age of death of the listed football players is 25.925. It represents that football can be a risky sport in which the players are likely to die younger than others. Among the listed dead players, about 35.5% died with a heart attack while playing, 15% died with a cardiac arrest while playing, 46.3% died from a collapse, 2.8% died from lightning (struck by lightning), and 7.5% died from a collision. It turns out that heart attack and cardiac arrest are common causes of death among football

players. A collapse is more likely to be a trigger of a deadly illness than a collision or lightning. In addition, by looking at the data, 19 observations are recorded dead with a cardiac arrest from a collapse.

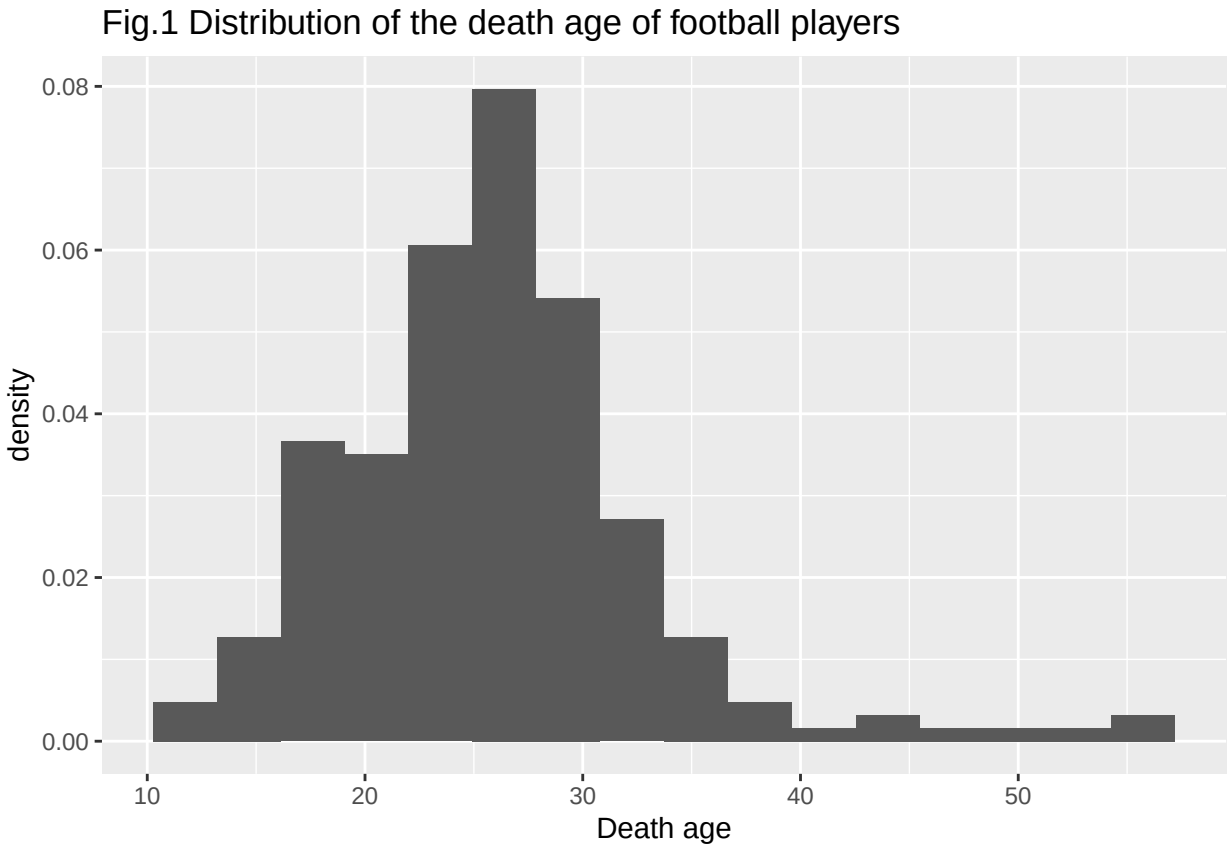

Fig.1 The histogram represents a nearly normal distribution but with a high concentration of age around 25 and some outliers around 50 and 60. A large number of deaths of young players can be explained by the properties of the sports industry that most athletes are young. Also, because athletes generally retire earlier from the industry than other employees, there is only a small portion of players who died in their 50's or 60's and showed as the outliers in this plot.

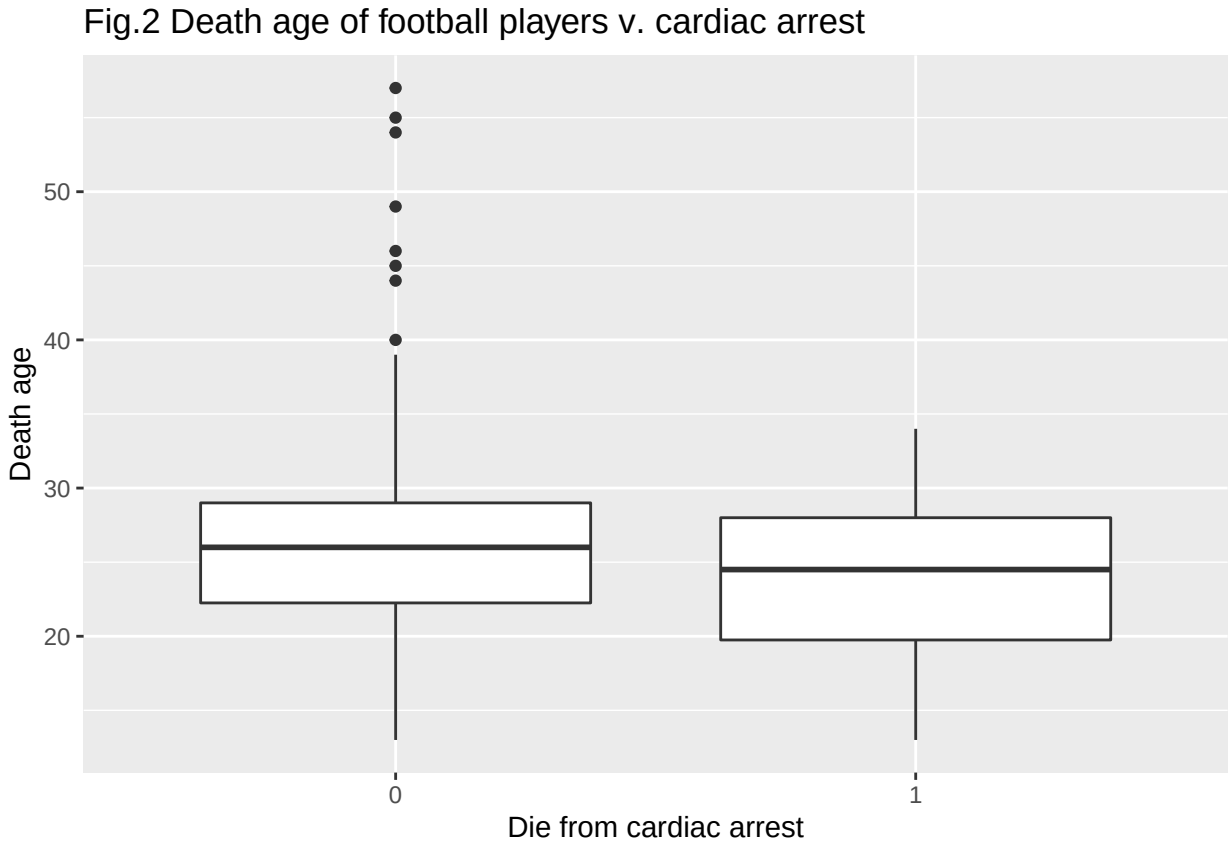

Fig.2 The boxplot describes the death age of players who died from cardiac arrest and who died from other injuries. It turns out that cardiac arrest was likely to cause the deaths of younger players than other injuries. Form the first box which represents the deaths due to other injuries or illness, there are several outliers on the top, but there aren't any on the top of the second box, which represents the deaths due to cardiac arrest. It means that the deaths of older players are more likely to be caused by other injuries or illnesses rather than cardiac arrest, and on the other hand, cardiac arrest is more likely to cause players to die earlier.

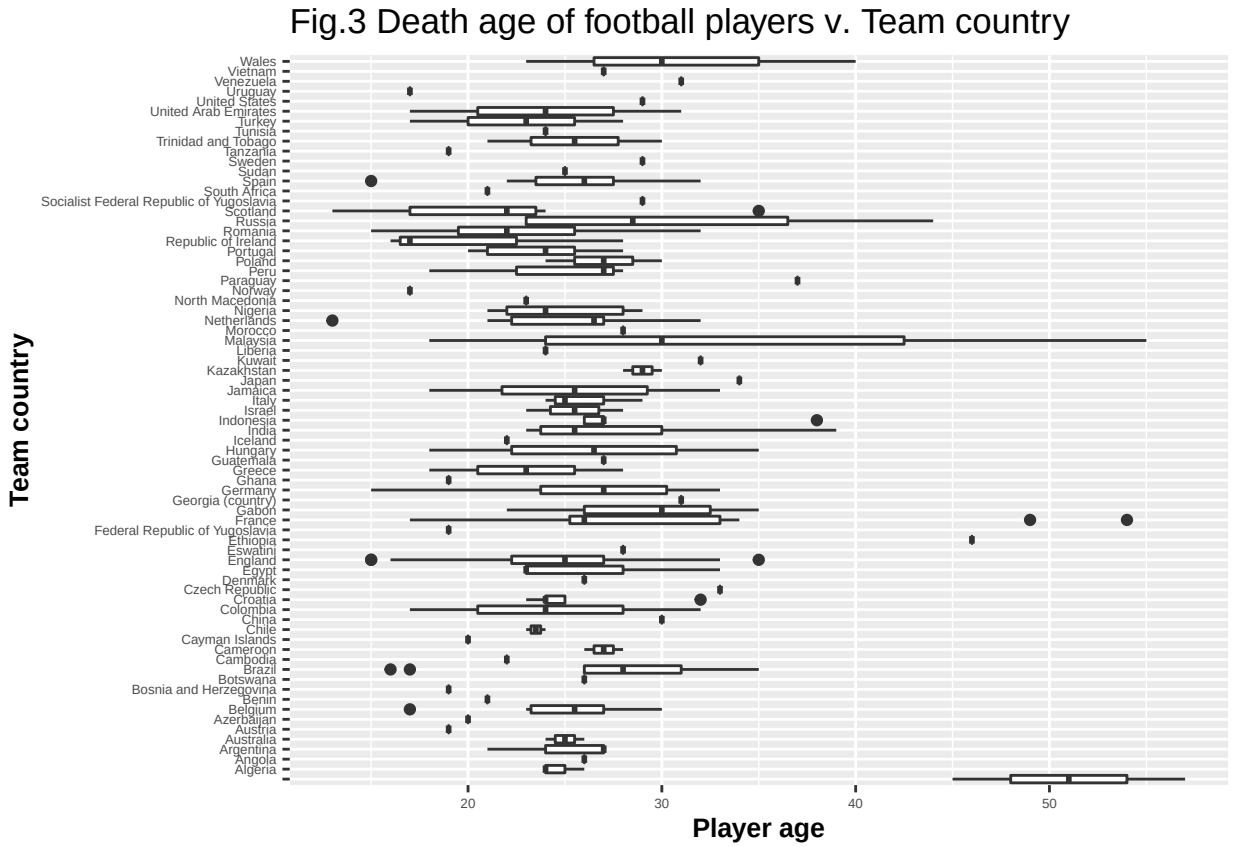

Fig.3 shows the death age of football players according to the country of his team. We see that there are pretty many differences among various team countries. However, there is no strong evidence about the relationship between death age and country because some countries may have more teenager teams than others. Other factors such as laws and rules may also affect the overall players' ages in a country, and thus, we cannot make any conclusion about the death age either.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

Linear regression model

A linear regression model is a statistical approach for analyzing the linear relationship between a response and one or more explanatory predictors. The general form of a linear regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$, where Y is the response, Xs are the predictors, β_0 is the intercept, and $\beta_{1,...,n}$ are the coefficients of corresponding predictors.

Propensity Score Matching

As mentioned in **Data collection**, the data was collected and cleaned by an individual, and thus it probably has bias and limited coverage. The data only includes 229 players. Some districts such as south Asia are not covered. Based on these facts, it is suitable to use propensity score matching to reduce or eliminate the bias by balancing the characteristics of participants. Basically, for every player who died from cardiac arrest, we want those who died from other injuries to have similar characteristics. Then I will examine the causal effect of cardiac arrest on the death age of football players. To observe the treatment effect of cardiac arrest, it is impossible to assign people to the treatment "having a cardiac arrest". Thus, PSM would be an alternative way to estimate the treatment effect. Based on the empirical information from **Literature**, PSM is needed to control the bias of "collapse" and "team country", which are factors that predict cardiac arrest. The backgrounds indicate that collapse can cause a cardiac arrest and the different medical standards of different countries can also affect the probability of getting a cardiac arrest in practice. In this situation, PSM will achieve the goal by making the treatment group, and non-treatment group similar and comparable to the control variables. Nevertheless, PSM has disadvantages. It has been proved to increase model inefficiency, imbalance, and dependence because it does not account for latent characteristics. The true propensity score is unknown, which will cause inaccuracy.

Logistic regression

Logistic regression is a statistical model used for predicting the probability of an event that has a binary outcome. As the first step of PSM, we need to generate a logistic regression that explains the propensity of being treatment i.e. dying from a cardiac arrest. As mentioned before, "collapse" and "team country" may both affect the propensity of getting a cardiac arrest. Thus, "collapse" and "team_country" would be the predictors of the logistic regression, and "cardiac_related" would be the response. Based on the logistic regression, we know the propensity of dying from a cardiac arrest of each player listed.

Matching

For each player that has a propensity p of dying from a cardiac arrest and did die from a cardiac arrest, we want to match him to a player that also has the propensity p of dying from a cardiac arrest but died from other illnesses. In other words, for every player that died from a cardiac arrest, we want to match him to one who had the same propensity to die from a cardiac arrest but died from other reasons. Then we want a dataset containing only those matched pairs. This process will reduce many observations and may cause inaccuracy in the estimation.

Results

By taking collapse and team countries into consideration, we extracted a purer causal effect of cardiac arrest on the death age of football players than other common illnesses. Overall, the propensity score model tells us that cardiac arrest harms the age of death of football players, which proves the hypothesis. The result is reasonable because cardiac arrest is a deadly illness that even the youth that has strong body and organs can barely survive from it. It also turns out that collapse can be a cause of cardiac arrest. There are always collapses during a game or training and it is unavoidable. Therefore, players should always avoid fierce collapses though the game is competitive.

Model for matching

$$Cardiac\,related = \alpha_0 + \alpha_1 Collapsed + \alpha_2 Team\,country$$

Model for the topic

$$Death\,age\,of\,player = \beta_0 + \beta_1 Cardiac\,related + \beta_2 Collapsed + \beta_3 Team\,country$$

Propensity Score Matching table

Parameter	Coefficient	SE	95% CI	t(43)	p
(Intercept)	27.67	4.43	[18.73, 36.61]	6.24	< .001
cardiac related	-3.69	1.48	[-6.68, -0.71]	-2.50	0.016
collapsed	-0.82	2.00	[-4.85, 3.20]	-0.41	0.682
team country [Brazil]	1.12	4.45	[-7.86, 10.09]	0.25	0.803
team country [Chile]	-1.91	5.60	[-13.21, 9.38]	-0.34	0.735
team country [Croatia]	-1.62	5.04	[-11.77, 8.54]	-0.32	0.750
team country [England]	-1.29	4.47	[-10.30, 7.73]	-0.29	0.775
team country [France]	-4.00	5.51	[-15.11, 7.11]	-0.73	0.472
team country [Gabon]	1.00	5.51	[-10.11, 12.11]	0.18	0.857
team country [Germany]	1.23	4.61	[-8.07, 10.53]	0.27	0.791
team country [India]	4.38	5.04	[-5.77, 14.54]	0.87	0.389
team country [Italy]	1.85	6.79	[-11.85, 15.54]	0.27	0.787
team country [Japan]	10.85	6.79	[-2.85, 24.54]	1.60	0.117
team country [Portugal]	-1.54	4.93	[-11.49, 8.40]	-0.31	0.756
team country [Romania]	-2.50	5.51	[-13.61, 8.61]	-0.45	0.652
team country [Scotland]	-5.03	4.88	[-14.87, 4.82]	-1.03	0.309
team country [Spain]	-1.03	4.47	[-10.03, 7.98]	-0.23	0.819
team country [Sweden]	5.85	6.79	[-7.85, 19.54]	0.86	0.394
team country [Uruguay]	-6.98	7.08	[-21.26, 7.31]	-0.99	0.330
team country [Venezuela]	7.85	6.79	[-5.85, 21.54]	1.16	0.254
team country [Wales]	6.07	5.21	[-4.44, 16.58]	1.16	0.251

According to the model, there is a significantly negative relationship between the death age of players and cardiac arrest. It proved the hypothesis that cardiac arrest is more likely for football players to die younger than other deadly illnesses. While there is no significant evidence that team countries or collapses were directly related to the death age. Overall, the treatment effect of cardiac arrest on death age is that a player who got a cardiac arrest would approximately die 3.69 years earlier than players who get the other deadly accidents or illnesses.

However, in the model in the propensity matching process, the sample size got reduced to 64 (32 pairs). In addition, $R^2 = 36.9\%$ (not shown) represents that 36.9% of the variance of the response is explained by the predictors, and it means that the explanation is not sufficiently efficient.

Conclusions

With the higher and higher popularity of football, the health problems of football players are also more concerned. However, many illnesses and injuries are unavoidable in competitive games and training. Before the analysis, the hypothesis is that a more negative impact of cardiac arrest on the longevity of football players than other common deadly accidents or illnesses such as heart attack. In the project, a linear regression model of the death age and cardiac arrest through a propensity score matching ensure a purer treatment effect of a cardiac arrest. The results have proved the hypothesis that **cardiac arrest will cause football players to die earlier compared to other causes of death**. It turns out that football associations, teams as well as players themselves should pay more attention to cardiac health. It is difficult to reduce the competitiveness of a match, thus, some potential ways of improvement may be more medical or physical care and so forth. While there is no quick solution to it since there are a lot more considerations.

Weaknesses

1. Assumptions of a linear model: The assumptions of a linear model are not checked in this project. Therefore, it may be unreasonable to make any further inference before checking the assumptions. If there is any violation of assumptions, the model needs to be transformed, and thus, the results may be different from the recent ones.
2. Data: As mentioned before, the data was collected and cleaned by an individual. Hence, there may be incorrect or missing information, which may impact the results of this project. Also, the information provided by the data is not sufficient, and more information about the health condition of the dead players such as medical histories would be necessary for more detailed analysis.
3. PSM: As mentioned before, PSM has drawbacks. It reduces the sample size a lot in the process, and it may affect the accuracy of the estimates. This would also increase the imbalance of the data. Also, PSM only accounts for observable covariates but not latent characteristics, and thus the hidden bias of the model may remain through the processes.
4. Sex: The data only contains male football players. We may also collect information on female players to compare and to estimate the impact of sex on cardiac arrest, and to draw a more complete conclusion, etc.

Next Steps

The most required step is firstly to check the assumptions to ensure the appropriateness of the linear models. If possible, we also need to collect more detailed information such as the medical histories and syndromes of the players. We need more predictors to make a more complete and accurate estimation. In addition, more supplementary methods should be used to compensate for the shortcomings of PSM.

Discussion

The report *Analysis of the causal effect of a cardiac arrest on the death age of football players* gives an introduction of cardiac arrest and the death of football players during a game. The analysis was made on the data collected and posted by Shivam Bansal on Kaggle. The main method is Propensity Score Matching (PSM), which is a statistical matching technique that estimates the treatment effect by accounting for the covariates that affect the propensity of being treated. This method allows us to get a purer treatment effect with less influence from the covariates. However, it has some drawbacks that add imbalance and bias to the model. Under these conditions, the results show that cardiac arrest causes players to die earlier than other deadly illnesses. It is the basic outcome of the analysis. In the future, more data and methods should be added to the analysis to get a more accurate and complete estimation.

Bibliography

Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: December 17, 2021)

Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: December 17, 2021)

(Data) Bansal, Shivam. "Football Players Deaths." Kaggle, November 30, 2021. https://www.kaggle.com/shivamb/football-deaths. (Last Accessed: December 17, 2021)

Higgins, Patrick John, and Aldo Andino. " Soccer and Sudden Cardiac Death in Young Competitive Athletes: A Review." Research Gate, January 2013. https://www.researchgate.net/publication/258400862_Soccer_and_Sudden_Cardiac_Death_in_Young_Competitive_Athletes_A_Review. (Last Accessed: December 17, 2021)

Roxby, Philippa. "Christian Eriksen: What Can Cause a Cardiac Arrest?" BBC News. BBC, June 14, 2021. https://www.bbc.co.uk/news/health-57469627. (Last Accessed: December 17, 2021)

Appendix

A1: Ethics Statement

The original data has license CC0: Public Domain, which allows the use of the data with no copyright. The original data and codes are both uploaded and are permitted for reproducing. All the resources including data are completely cited in *Bibliography*. This project contains no knowing or unknowing p-hacking, gender-sex adjustment, or unpublished privacy information.

A2: Materials

Dataset:

```
## Rows: 229
## Columns: 13
## $ row_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ incident_date        <fct> 1889-01-13, 1892-01-11, 1893-11-12, 1893-11-23, 1~
## $ player_name          <fct> William Cropper, James Dunlop, John Henry Morris,~
## $ player_country       <fct> England, Scotland, England, England, England, Sco~
## $ team_country         <fct> England, Scotland, England, England, England, Eng~
## $ player_age           <dbl> 26, 21, 26, 24, 27, 25, 25, NA, 35, 23, 25, 26, 2~
## $ player_team_name     <fct> Staveley, St Mirren, Shrewsbury Town, Chesterfiel~
## $ incident_description <fct> "Ruptured bowel in a match against Grimsby Town l~
## $ heart_related        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0~
## $ cardiac_related      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ collapsed            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ lightning            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ collision            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
```