Effect of Field Surface on NFL Injuries

Anna Niu

Introduction

In sports, there has often been a debate about whether to use artificial turf or natural grass. Artificial turf is advocated for due to it taking fewer resources to maintain; however, most athletes prefer playing on natural grass, claiming that synthetic turf increases their chances of getting injured. In my analysis, I focus on the injuries on different types of turf for professional football players. Currently in the National Football League (NFL), 14 out of 30 fields have artifical turf. Thus, the goal of this paper is to explore if playing on a synthetic field causes an increased probability of an NFL player getting injured.

Data Description

The data I used was provided from the NFL and contains two seasons of injury and game/player data. The diagram below highlights the key variables located in each of the three tables. I cleaned the data by removing null values and joined the tables together using the PlayKey and GameID.

The GameID is in the format of (PlayerKey-Game#), which identifies each instance in the dataset. The Game#'s aren't in temporal order, i.e. a smaller Game# doesn't necessarily indicate an earlier game.

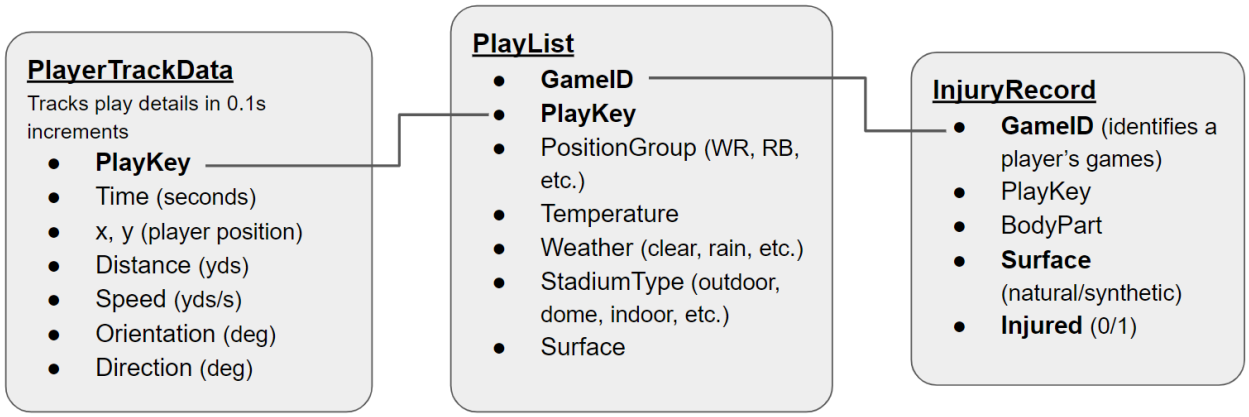


Figure 1: Diagram of Raw Data

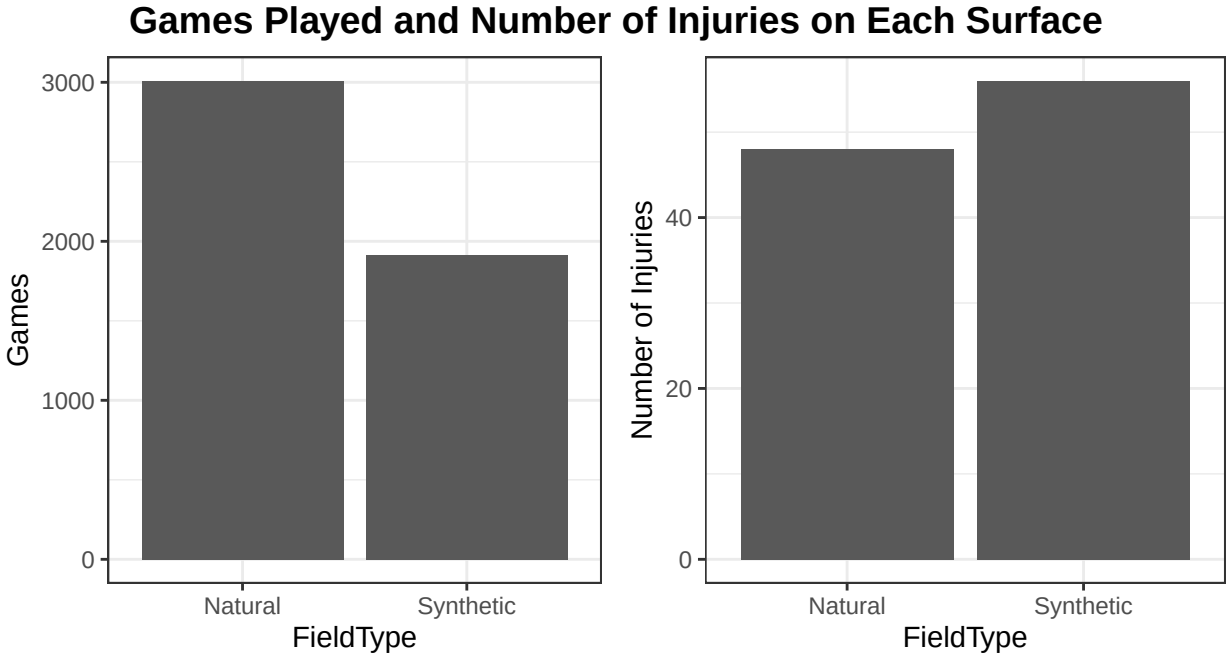
Table 1: Table of variables used in models

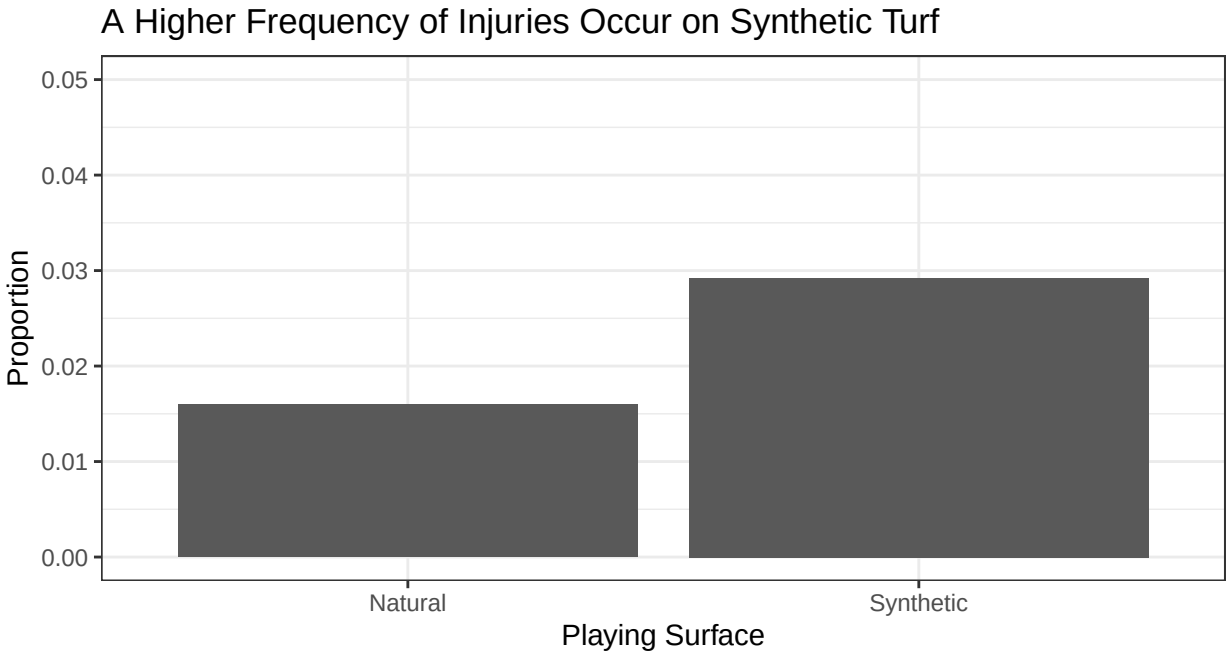
	Variables
Outcome Variable	is_Injured
Treatment	is_Synthetic
Weather	clear, cloudy, rain, snow
StadiumType	outdoor
PlayerPosition	DB, DL, LB, OL, QB, RB, TE, WR
Numeric Variables	AverageSpeed, NumPlays, Temperature

The variables in the final dataset are shown in the Table 1. Each subject in the dataset is a player (that's linked to a specific game). For example, a subject Player5-13 contains information about the position of the player, their average speed during game 13, the conditions of the game, etc. The categorical variables were also split into binary variables. In my analysis, the outcome variable, `is_Injured`, is a binary variable indicating whether or not the player was injured in that game. The treatment variable, `is_Synthetic`, is a binary variable referring to whether or not the field the game was played on was synthetic turf.

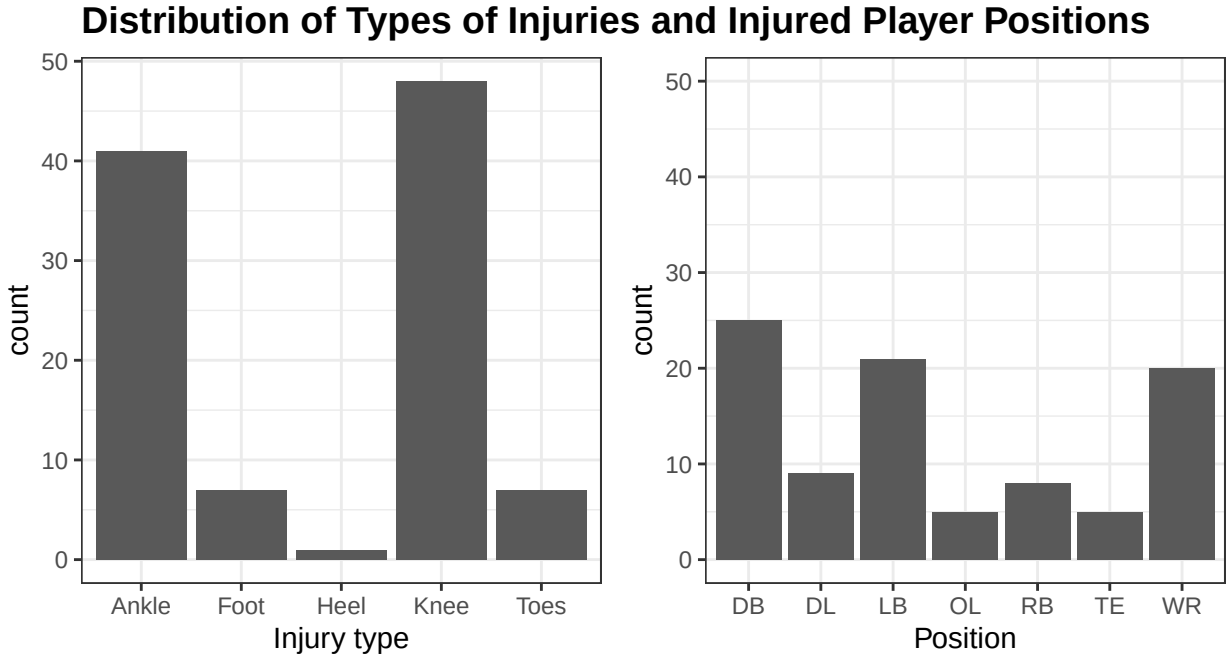
Exploratory Data Analysis

Before implementing any methods, I first perform some exploratory data analysis. In the dataset, more games were played on natural grass fields, but more injuries occured on synthetic turf. This is also highlighted by looking at the frequency of injuries on each field type, where there's a higher frequency of injuries occuring on synthetic turf. I look at if this treatment effect is statistically significant in the methods later in this paper.





I also look at the distribution of types of injuries and which positions were injured the most. Knee and ankle injuries were the most common. Defensive backs, linebackers, and wide receivers were also most frequently injured.

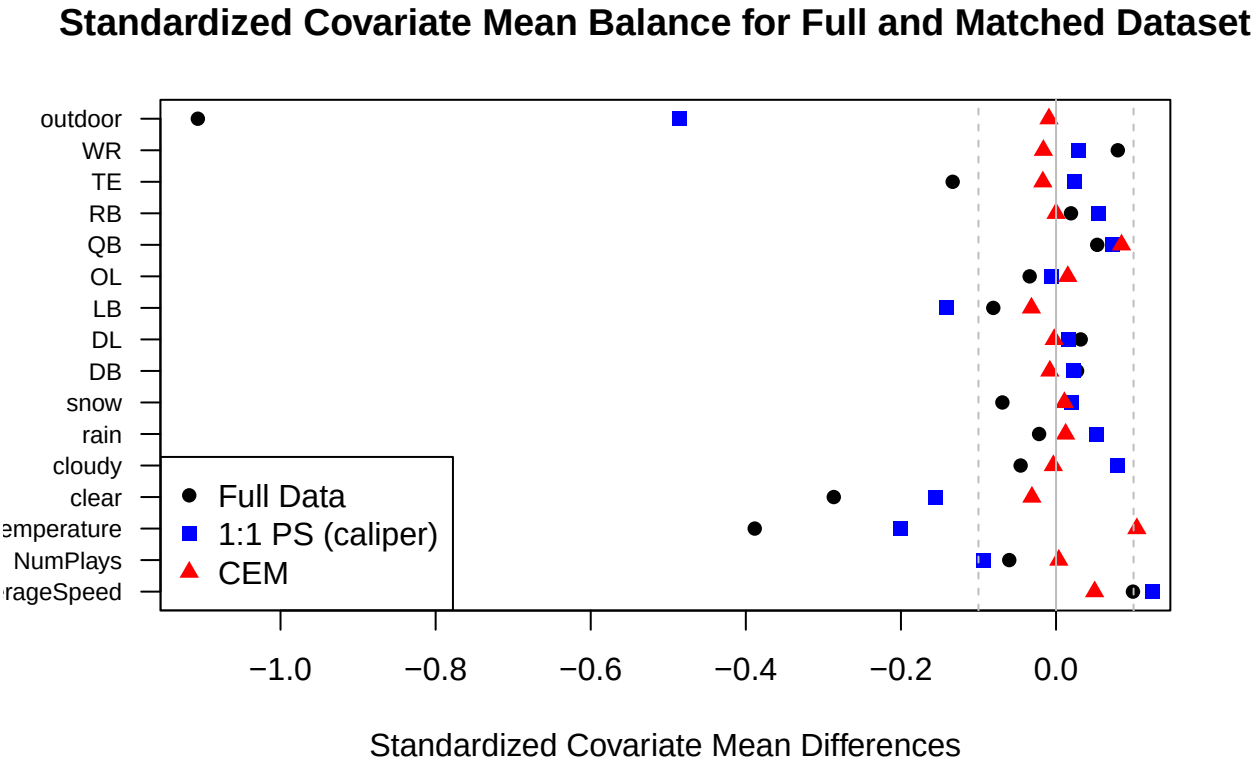


Methods

In this report, I explore the four causal inference methods listed below:

- 1:1 Propensity Score Matching
  - Caliper of $0.2 * sd \approx 0.0507$
- Coarsened Exact Matching (CEM)
- Plug-in Estimator
- Doubly-Robust Estimator (DR)

To estimate the propensity scores for the 1:1 PS Matching and DR Estimator, I used a logistic regression model. I also used a logistic regression model in the plug-in and DR estimator to predict the outcome variable (`is_Injured`) using the covariates listed in Table 1. For the one-to-one propensity score matching, I implemented a caliper of approximately 0.0507 to prevent bad matches.

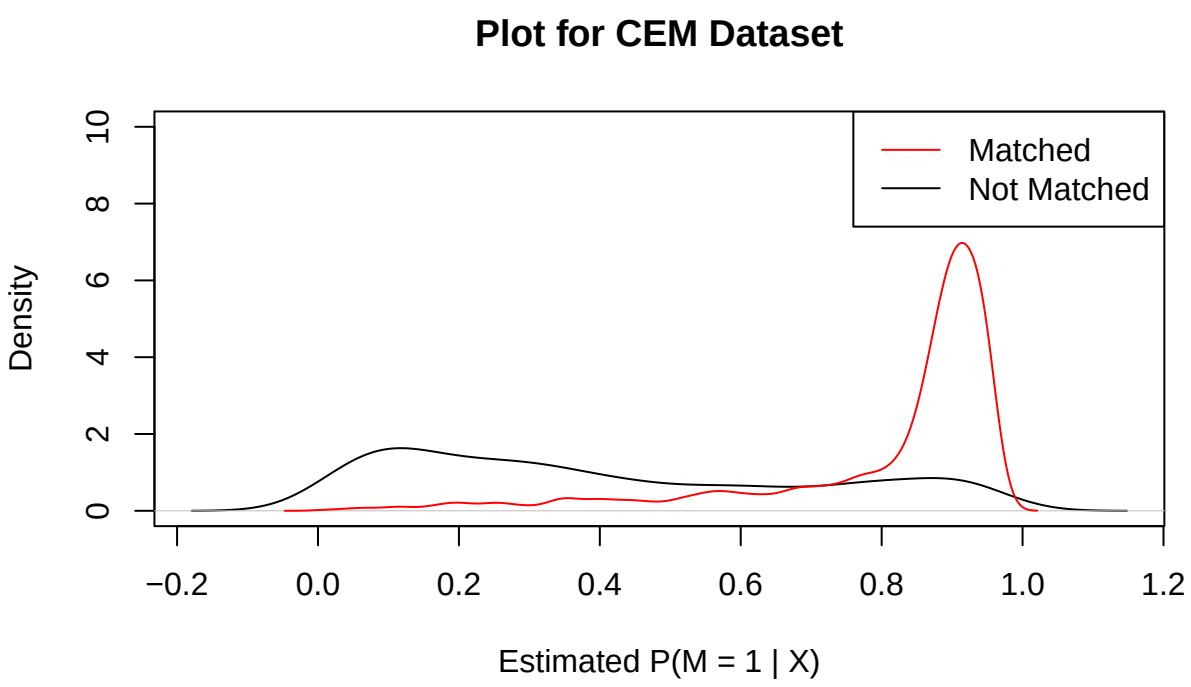


The love plot above compares the standardized covariate mean differences of the full dataset and matched datasets. The CEM dataset had the best balance, with all of the covariates being within the 0.1 rule-of-thumb. This was followed by the PS matched dataset which had most covariates being well-balanced. The worst-balanced dataset was the full dataset, which was to be expected. However, as shown in the Table 2, the full dataset has the most number of subjects, having almost double the number of treated subjects as the CEM dataset. Thus, there is a tradeoff between the number of instances and covariate balance.

Table 2: Number of instances in each dataset

	Control	Treatment
Full	3009	1916
1:1 PS (Caliper)	1155	1155
CEM	2302	1045

Overall, the CEM method seems to obtain the best balance in terms of having more total subjects than the PS matched dataset and having all predictors being balanced. However, it's important to note that it targets the ATT in the matched dataset, rather than the ATE. In addition, the estimate calculated may not be generalizable. The figure below shows the distribution of the probability of being matched, given the covariates, for the matched and not-matched subjects. We see that subjects that are more likely to be matched appear more frequently in the CEM dataset, so we note the issue of generalizability of the estimate.



Since each subject of the dataset refers to a player in a specific game, it's possible that the same player is matched to themselves. For example, PlayerA-Game2 could be matched with PlayerA-Game11. I explore this by seeing how many strata have the same player appear in the treatment and control group. For the CEM dataset, about 40% of strata have a player appear in the treatment and control group at least once. There is an average of 9-10 subjects in each stratum, so I also looked at the number of strata that had the same player appear at least 5 times in the stratum, resulting around 6% (Table 3). For the PS matched dataset, around 1.7% of pairs matched a player to themselves (Table 4).

Table 3: Proportion of strata that have the same player in the treatment and control group

Minimum number of times the same player appears in a group	Number of strata with same player	prop
2	142	0.4057143
5	22	0.0628571

Table 4: PS matched pairs containing the same player

Number of pairs with same player	Proportion
20	0.017316

Results

After implementing the models, I present the point estimates of the average treatment effect (ATE) and the 95% confidence intervals in the Table 5. There's an exception for the CEM estimate as it refers to the average treated effect for the treated (ATT) in the matched dataset, rather than the ATE.

Table 5: Table of estimates

	Point Estimate	95% Confidence Interval
1:1 PS (Caliper)	0.0061	[-0.00557, 0.01769]
CEM	0.0066	[-0.00217, 0.01537]
Plug-in Estimator	0.0074	[-0.0017, 0.01657]
Doubly-Robust	0.0081	[-0.00362, 0.01983]

The point estimates for each of the different methods range from 0.0061 to 0.0081. The confidence intervals for all of the methods include zero, signifying that the estimated ATE's (or ATT) aren't statistically significantly different from zero. In other words, there isn't enough evidence to conclude that playing on artificial turf causes a higher probability of an NFL player getting injured. However, I note that zero is on the edge of the confidence intervals and all the point estimates are greater than zero, which prompts further research in seeing if there is a higher frequency of injuries on artifical turf compared to natural grass.

Conclusion

The dataset used for this analysis only contained 2 seasons of data and the matched datasets contained approximately half of the number of subjects available. Thus, data collected from more than 2 seasons could decrease the variance of the point estimates and produce narrower confidence intervals. The GameID's for each subject were also not time sequential, i.e. there was no way of identifying whether the player was in a game near the beginning of the season or near the end of the season. Another potential weakness was that the data lacked variables signifiying the pace of play in the overall game. For further studies, having measure of pace of play in the game, since it could be the case that players inherently play differently when on different fields, which could lead to injuries. Having information on humidity or field condition could improve the analysis as well, because the weather could be clear but the field could still be slippery. It would also be interesting to focus solely on non-contact injuries, which could increase the focus on the effects of artificial vs natural turf.