NFL 1st and future: analytics

Introduction

The National Football League (NFL) is one of the four major professional sport leagues and also one of the most popular team sports in North America. The competition consists of 32 teams spread into 2 divisions. The season is divided in a preseason, regular season and postseason. The regular season consists of 256 games, where each team plays 16 games during a 17 week period. Every year, at the end of the postseason, the champions of both divisions compete against each other in the Super Bowl, to determine the league's champion.

Although viewer rates have been lowering steadily since 2015, the NFL increases its revenue each year. According to Forbes, the most valuable team of the NFL, the Dallas Cowboys, had an astounding 950 million dollars in revenue in 2019, which is in fact the most of any U.S. sports team.
The rising revenue for the NFL and the rising value of sport teams in general, also implies a higher cost for injuries. It becomes increasingly important for teams to get their star players on the pitch and it can be critical at times when key players get injured.
The challenge imposed by the NFL, was to investigate the relationship between the playing surface and the injury and performance of NFL athletes and to examine factors that may contribute to lower extremity injuries. Due to time constraints, we restricted ourselves to the first part of the challenge and examined the influence of the field type and environmental factors on the odds of sustaining an injury.

Data cleaning, data wrangling and exploratory analysis

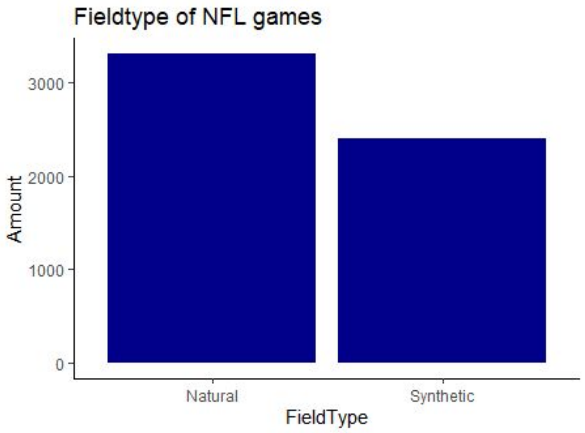
The original dataset used in this analysis is the combination of the Playlist and Injury datasets. It consists of 267.006 observations and 20 variables (table 1).
After data cleaning, 7 types of variables remain (table 4) with 9,7 percent missing values in total (table 5). Since most of the missing data was from the StadiumType variable, and since MAR or MCAR cannot be assumed, the rest of the analysis was based on complete cases to minimize the possible bias. Eventually, the StadiumType variable was omitted as well from the final model, leaving the dataset with 2,67% missing values.
The next section contains a thorough description of the data cleaning and data wrangling process. Additionally, an exploratory analysis of the 7 remaining variable types for analysis has been included as well.

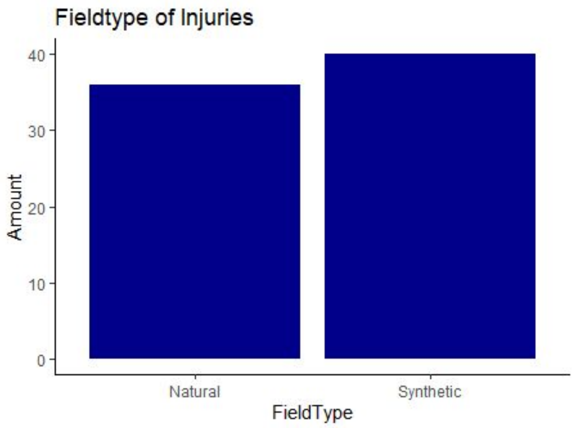
Table 1: Variables from the original dataset

Variable in original dataset	Description
PlayerKey	Uniquely identifies a player with a five-digit numerical key
GameID	Uniquely identifies a player's games (not strictly in termporal order)
PlayKey	Uniquely identifies a player's plays within the game (in sequential order)
RosterPosition	Provides the player's roster position (i.e. Running Back)
PlayerDay	An integer sequence that reflects the timeline of a player's participation in game
PlayerGame	Uniquely identifies a player's games
StadiumType	A free text description of the type of stadium
FieldType	A categorical description of the field type (Natural or Synthetic)
Temperature	On-field temperature at the start of the game
Weather	A free text description of the weather at the stadium
PlayType	Categorical description of play type
PlayerGamePlay	An ordered index denoting the running count of plays the player has participated in during the game
Position	A categorical variable denoting the player's position for the play
PositionGroup	A categorical variable denoting the player's position group for the position held during the play
BodyPart	Identifies the injured body part (Knee, Ankle, Foot, etc.)
Surface	Identifies the playing surface at time of injury (Natural or Synthetic)
DM_M1	One-Hot encoding indicating 1 or more days missed due to injury
DM_M7	One-Hot encoding indicating 7 or more days missed due to injury
DM_M28	One-Hot encoding indicating 28 or more days missed due to injury
DM_M42	One-Hot encoding indicating 42 or more days missed due to injury

1) FieldType, Surface and Injury variables

The merger between both aforementioned datasets resulted in 2 variables with similar meaning. Of those 2 variables, the Surface variable was transformed to the Injury variable, which is a binary indicator to denote whether the person sustained an injury to the lower extremities or not. The Injury variable was used subsequently as outcome variable for the analysis. This variable was found to be highly unbalanced. However, since inference is of interest instead of prediction, this was of minor concern. A first glimpse into the relationship between injuries and type of field reveal that most games are played on a natural turf, however, most injuries occur on a synthetic pitch.





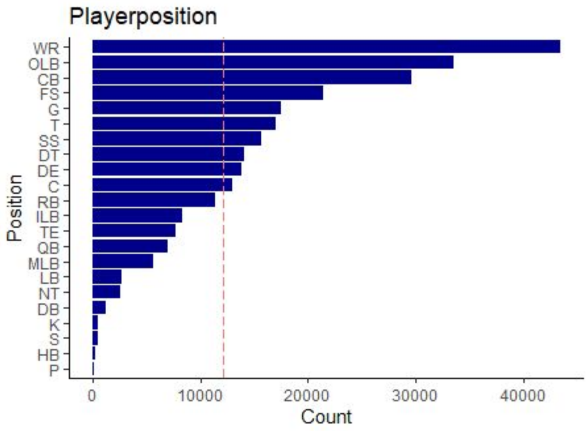
2) PlayerKey, GameID and PlayKey variables

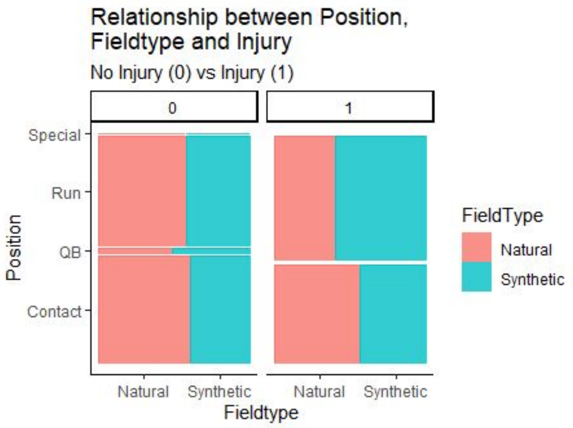
These variables indicate player, play and game specific information. These are solely used for the merging of datasets and do not provide additional information for the analysis and were subsequently deleted.

3) RosterPosition, Position and PositionGroup variables

Only the Position variable has been preserved for the analysis. The RosterPosition and PositionGroup have been omitted.

The dataset comprises 22 possible positions. These positions have been grouped into 4 categories: Special teams (“Special”), Quarterback (“QB”), positions where most of the action is about running (“Run”) and positions where most of the action is about contact (“Contact”).
There are proportionally more injuries for Run-type players in the injury group than in the no injury group. Special teams and QBs are not present in the injury group since these players did not sustain any injury to the lower extremities.



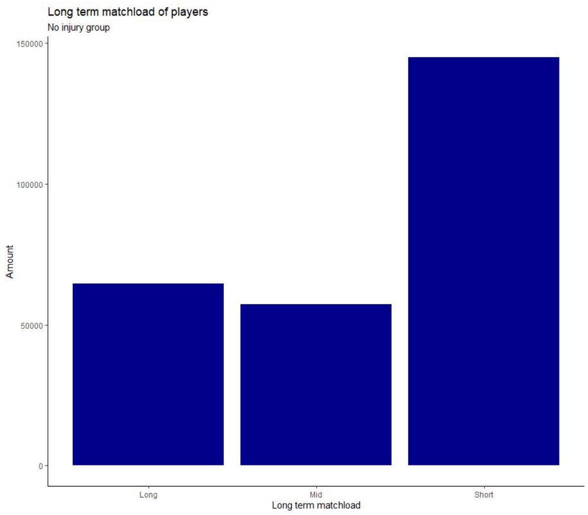


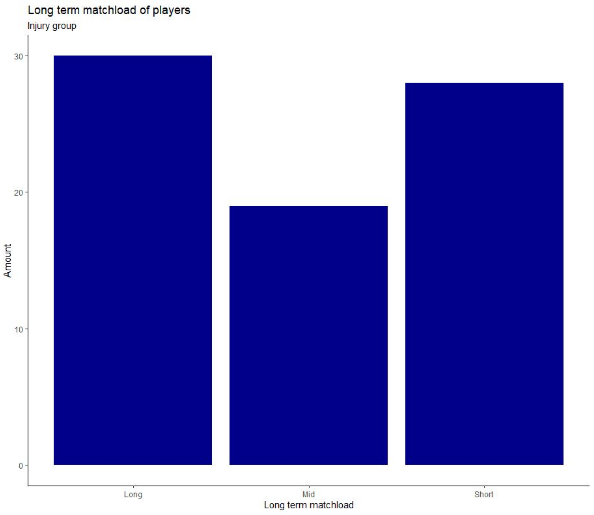
4) PlayerDay, PlayerGame and PlayerGamePlay variables

The PlayerDay variable reflects the timeline of a player’s participation in games and has subsequently been transformed for use in the analysis. The main objective of the transformation was to evaluate the match load as the number days (time span) since the last game (short term), 2 games back (mid term) and 3 games back in time. Thereafter, the 3 variables were binned into three categories to indicate if this timespan is short, medium or rather long. These variables can be found in table 2 with respective cut-off values.

table 2: Matchload variables with cut-off values

Variable	Short cut-off (days)	Medium cut-off (days)	Long cut-off (days)
Short_bin	<= 8	> 8 & <= 16	> 16
Mid_bin	<= 16	> 16 & <= 24	> 24
Long_bin	<= 24	> 24 & <= 48	> 48





The distribution of the different categories in the injury and no injury group for the long term variable are given above as an example. Most players have a short timespan of match load, however, the majority players who sustained an injury are in the long timespan group. Additionally, the categories are more balanced in the injury group.

5) Temperature variables

Approximately 24.000 missing values were recorded for the temperature variable. Most of them were in stadiums that were closed. Since temperature can be controlled better in stadiums that are closed, those values were median imputed. Thereafter, the remaining 1973 missing data were omitted from the analysis.

Temperature, expressed in degrees Fahrenheit, expresses only a minor deviation from normality, with some outliers to the left. Additionally, A smaller spread with an average slightly higher for the injuries in comparison to the non injuries can be found in the plot below.
Grouping has been performed with 3 different levels: “Cold”, “Intermittent” and “Hot” with 2 different cut-off levels (table 3).
All three variables (2 categorical and the original on a continuous scale) will be used for further analysis to see the best fit to the model.

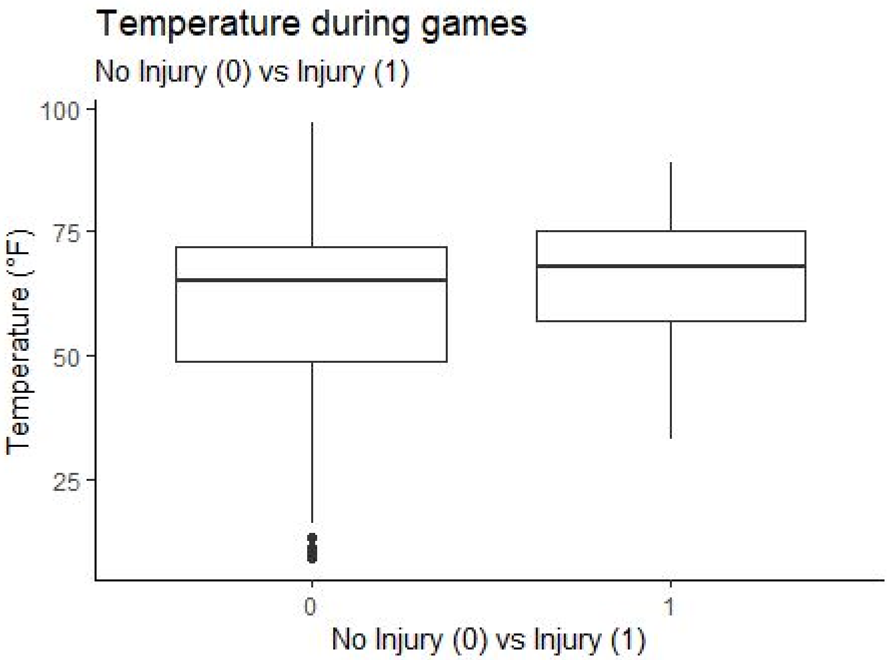


table 3: Grouping of Temperature variable

Variable	Cold (°F)	Intermittent (°F)	Hot (°F)
First grouping	< 59	≥ 59 & < 68	≥ 68
Second grouping	< 59	≥ 59 & < 80	≥ 80

6) Weather and StadiumType variables

Considering both variables have respectively 64 and 30 levels, binning procedures have been performed.

Three different groupings for the weather variable have been assessed in the analysis:

- A grouping of 5 levels (“Sun”, “Rain”, “Cloudy”, “Clear” and “Indoor”)
- A grouping of 3 levels (“Rain”, “No Rain” and “Indoor”)
- A grouping of 2 levels (“Rain”, “No Rain”)

The StadiumType variable has been grouped in two levels (“Open” and “Closed”).

Both grouping variables have some overlap, especially where the variables are binned into 2 levels. The key difference between the two can be distinguished into weather conditions during the match (Weather variable) and if the pitch has been exposed to weather conditions on a long term or not (StadiumType variable).

Imputations for the Weather variables were based on the Stadiumtype variable. Most of the missing data were from the weather variables, mainly from indoor types of stadiums. These could be imputed in the grouping variables quite easily.

7) Injury time variables and Bodypart

The Bodypart variable and the DM variables denoting the amount of days lost to injury, were omitted from the analysis.

After the initial analysis to investigate the associations between sustaining a lower extremity injury and environmental factors, the same analysis has been performed on injuries were the amount of days missed is equal to or more than 28 days. This in order to examine whether there is a difference between long term and short term injuries.

Table 4: Variables used in the analysis

Variable	Description
Injury	Binary indicator for injury (1) or no injury (2)
FieldType	Binary indicator for Synthetic or Natural turf
Temperature	3 different variables examined for analysis: - Continuous scale - Categorical scale with 3 levels: “Cold”, “Intermittent” and “Hot” - Categorical scale with 3 levels: “Cold”, “Intermittent” and “Hot”
Weather	3 different variables examined for analysis: - Categorical variable with 5 levels : “Sun”, “Rain”, “Cloudy”, “Clear” and “Indoor” - Categorical variable with 4 levels : “Rain”, “No Rain” and “Indoor” - Categorical variable with 3 levels : “Rain”, “No Rain”
StadiumType	Binary indicator for Open stadium and Closed stadium
PlayType	Categorical variable with 3 levels: “Kick”, “Rush”, “Pass”
Position	Categorical variable with 4 levels: “Contact”, “QB”, “Rush”, “Special”
Timespan	3 different variables examined for analysis: - Short timespan (difference of days between last game) - Medium timespan (difference of days between 2 games before) - Long timespan (difference of days between 3 games before)

Table 5: Missing Data patterns

Variable	Amount missing	Percentage missing
Temperature	1.973	0,81 %
Weather	5.106	2,098 %
PlayType	646	0,26 %
Position	45	< 0,01 %
StadiumType	16.910	6,95 %

Analysis, discussion and conclusion

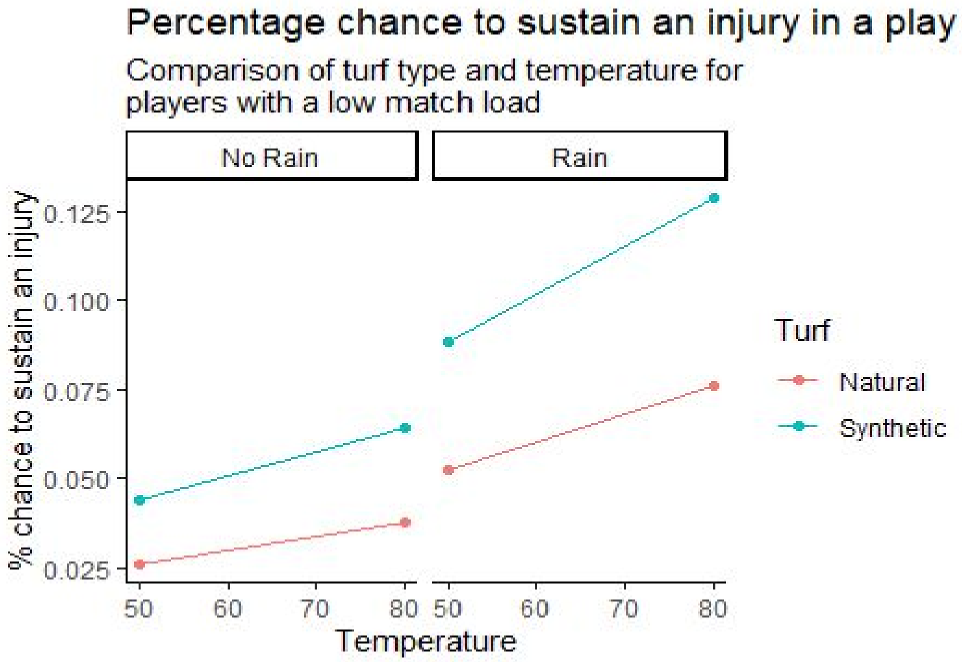
Since inference is the main objective, the dataset has been approached from a statistical point of view. Moreover, since the outcome is a binary variable (Injury or no injury), the choice for logistic regression with logit link function was made to assess the effects on sustaining an injury to the lower extremities, although, Poisson regression would be an equal viable choice for the analysis.
After the exploratory analysis and data cleaning, 7 remaining factors were taking into account for the analysis. Influences from the type of field, the type of weather (multiple possibilities), the positional roles of the players on the pitch, the temperature (multiple possibilities), the type of play, the type of stadium and the match load (multiple possibilities). Stepwise, both forward and backward, regression analysis has been performed to assess the associations between the probability of sustaining an injury to the lower extremities and the explanatory variables. The AIC criterion was used for model selection and possible confounding factors were also assessed. Models were limited to those of which the type of field was included. Aside of the aforementioned factors, also interaction effects were taken into account in the analysis.

The analysis resulted in the following model with the optimal AIC criterion:

Injury = “FieldType” + “Temperature” + “Weather” + “Longtermmatchload”

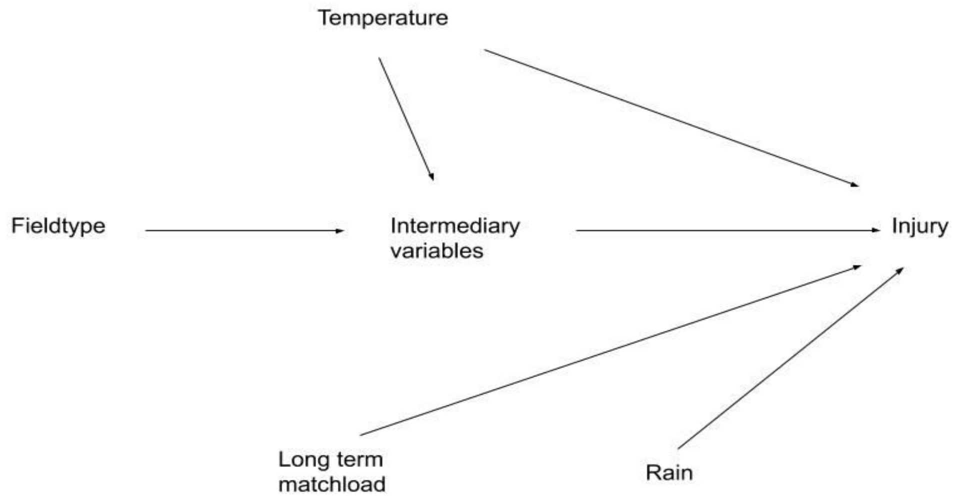
log(odds of sustaining an injury) = 0.52x”FieldType” + 0.012x”Temperature” +
-0.16x”Long_mid” + -0.69x”Long_short” + 0.69x”WeatherRain”

Both the type of field and the short vs long term match load were significant at the 0.05 level. The weather variable was marginally significant at the 0.1 level. The weather variable is the 2-level variable (Rain/No rain) and the Temperature on a continuous scale was the best fit to the data. The figure below shows the differences in probability to sustain an injury to the lower extremities between the type of field, the temperature and the weather conditions. The long term player load has been held constant in this case.



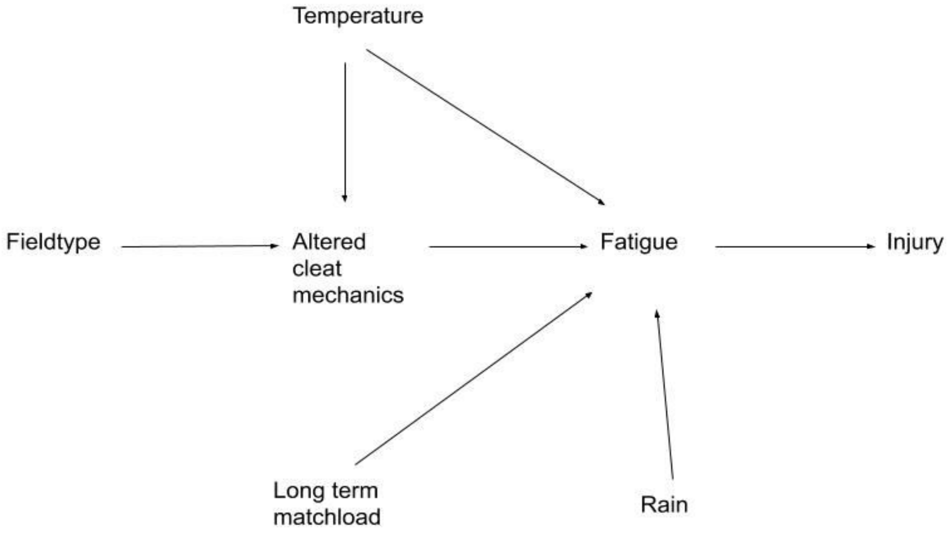
The odds of sustaining an injury rise with a factor of 1.68 times when playing on a synthetic pitch in comparison to a natural field, while controlling for environmental factors such as temperature and weather during the match and adjusting for long term match load. No interaction effects were of value to the model (based on the AIC).

Temperature was found to have a confounding effect on the relation of sustaining an injury and the type of field. Since no other interactions proved to be of value to add to the model (based on the AIC), only temperature seemed to have an influence on the relation between the type of field and the chance of sustaining an injury. The other variables only had an effect on the odds of sustaining an injury. An increasing temperature, whether there was rain during the game and a higher long term match load, proved to increase the odds of injury in American Football players. From this, it can be concluded that exposure to outdoor weather conditions such as rain and high temperatures, has an influence on sustaining an injury.



Fatigue is one of the biggest risk factors for sustaining an injury in multiple sports (Ekstrand et al., 2009, Van Tiggelen et al., 2016). Biomechanical studies of football cleat-surface interactions have shown that synthetic turf surfaces do not release cleats as readily as natural turf and may contribute to the incidence of non-contact lower limb injuries (Kent el al., 2015). It can be hypothesised that these differences in cleat-turf interactions can increase the level of fatigue for a player (by possible intermediary variables such as altered movement patterns or increased muscle load). In conjunction with this hypothesis, the following model has been created to view the relationship between the variables (figure below). From the analysis, it can be concluded that the type of field, the temperature or the presence of rain during the game and the long term match load have an effect on sustaining an injury to the lower extremities.

This is likely an indirect relation. It can be hypothesised that fatigue (physical, mental, internal, external or a combination of all) may be one of the main intermediary variables in that relationship. Further research is warranted to test this hypothesis.
The type of field may have an influence on an individual's fatigue through the altered cleat mechanics as found in previous research (Kent et al., 2015). Temperature was found to not only influence the odds of sustaining an injury, but also to confound the relation between the type of field and injury. Since the outdoor temperature can not change the type of field itself, we believe the temperature can have an effect on an intermediary variable such as the aforementioned alteration of cleat mechanics.



This primary analysis indicates a higher chance of sustaining an injury to the lower extremities when playing on a synthetic turf, while adjusting for the most common environmental factors that could influence this relation. However, further research is warranted to assess the influence of other possible confounding variables. Additionally, the causality of this relation has to be examined as well.

Furthermore, the long term match load variable, is maybe a badly chosen name. It tends to factor in the amount of days the 3 games sequence spans. Considering that professional athletes often have a combination of talent and physical preparedness, and fatigue is a combination of the load one puts on itself, and the load capacity of that body. Therefore it can be hypothesised that players who play often, also have a better physique, and thus are better prepared against injury. The variable captures in that sense the physical fitness of the athlete, and not the match load over a given number of games. Additionally, load on a player can be assessed in a broader sense. The number of games and trainings of an athlete in a certain timeframe, the external load or the internal load on a player are some examples thereof. Moreover, short and long term interventions to restore and improve the load capacity of a player (such as diet, sleep, cryotherapy etc.) should be researched as well when looking at the full picture of load. Additionally, a distinction or combination of physical and psychological load can be made.

As hinted in the data preparation part, the analysis has been done again after filtering for injuries that lasted for 28 days or more in order to investigate whether the effect of playing surface is also apparent for the long term injuries. While short term injuries such as contusions, minor muscle strains and type 1 ankle sprains are often more easily manageable, long term injuries of athletes can have a bigger impact on team and individual performance.

The influence of the type of pitch has been found to be even higher for long term injuries. The odds of sustaining a long term injury to the lower extremities on a synthetic field is 2,78 times the odds when playing on a natural pitch.
Additional research to the length of injury, return to play and return to sport times is warranted.

This research found a higher risk for sustaining an injury when playing on a synthetics pitch in comparison to a natural grass field. However, only match data and environmental factors have been taken into account. To get a more comprehensive knowledge of the injury causation mechanism, further research is warranted. A possible confounding factor could be that only players who are not accustomed to synthetics pitches (because their trainings and home games are played on natural turf), have a higher risk for sustaining an injury to the lower extremities. Additionally, the hypothesis of causation suggested in this analysis, should be investigated.