nflWAR: A Reproducible Method for Offensive Player Evaluation in Football
(Extended Edition)

Ronald Yurko, Samuel Ventura, and Maksim Horowitz

Department of Statistics & Data Science, Carnegie Mellon University

July 11, 2018

Abstract

Unlike other major professional sports, American football lacks comprehensive statistical ratings for player evaluation that are both reproducible and easily interpretable in terms of game outcomes. Existing methods for player evaluation in football depend heavily on proprietary data, are not reproducible, and lag behind those of other major sports. We present four contributions to the study of football statistics in order to address these issues. First, we develop the R package `nflscrapR` to provide easy access to publicly available play-by-play data from the National Football League (NFL) dating back to 2009. Second, we introduce a novel multinomial logistic regression approach for estimating the expected points for each play. Third, we use the expected points as input into a generalized additive model for estimating the win probability for each play. Fourth, we introduce our *nflWAR* framework, using multilevel models to isolate the contributions of individual offensive skill players, and providing estimates for their individual wins above replacement (*WAR*). We estimate the uncertainty in each player's *WAR* through a resampling approach specifically designed for football, and we present these results for the 2017 NFL season. We discuss how our reproducible *WAR* framework, built entirely on publicly available data, can be easily extended to estimate *WAR* for players at any position, provided that researchers have access to data specifying which players are on the field during each play. Finally, we discuss the potential implications of this work for NFL teams.

Keywords: multilevel model, generalized additive model, multinomial logistic regression, R, reproducibility

1   Introduction

Despite the sport's popularity in the United States, public statistical analysis of American football ("football") has lagged behind that of other major sports. While new statistical research involving player and team evaluation is regularly published in baseball (Albert, 2006; Jensen et al., 2009; Piette and Jensen, 2012; Baumer et al., 2015), basketball (Kubatko et al., 2007; Deshpande and Jensen, 2016), and hockey (Macdonald, 2011; Gramacy et al., 2012; Thomas et al., 2013), there is limited new research that addresses on-field or player personnel decisions for National Football League (NFL) teams. Recent work in football addresses topics such as fantasy football (Becker and Sun, 2016), predicting game outcomes (Balreira et al., 2014), NFL TV ratings (Grimshaw and Burwell, 2014), the effect of "fan passion" and league sponsorship on brand recognition (Wakefield and Rivers, 2012), and realignment in college football (Jensen and Turner, 2014). Additionally, with the notable exception of Lock and Nettleton (2014), recent research relating to on-field or player personnel decisions in football is narrowly focused. For example, Mulholland and Jensen (2014) analyze the success of tight ends in the NFL draft, Clark et al. (2013) and Pasteur and Cunningham-Rhoads (2014) both provide improved metrics for kicker evaluation, Martin et al. (2017) examine the NFL's change in overtime rules, and Snyder and Lopez (2015) focus on discretionary penalties from referees. Moreover, statistical analysis of football that does tackle on-field or player personnel decisions frequently relies on proprietary and costly data sources, where data quality often depends on potentially biased and publicly unverified human judgment. This leads to a lack of reproducibility that is well-documented in sports research (Baumer et al., 2015).

In this paper, we posit that (1) objective on-field and player personnel decisions rely on two fundamental categories of statistical analysis in football: play evaluation and player evaluation, and (2) in order to maintain a standard of objectivity and reproducibility for these two fundamental areas of analysis, researchers must agree on a dataset standard.

1.1   Previous Work: Evaluating Plays

The most basic unit of analysis in football is a single play. In order to objectively evaluate on-field decisions and player performance, each play in a football game must be assigned an appropriate value indicating its success or failure. Traditionally, yards gained/lost have been used to evaluate the success of a play. However, this point of view strips away the importance of context in football (Carter and Machol, 1971; Carroll et al., 1988). For instance, three yards gained on 3rd and 2 are more valuable than three yards gained on 3rd and 7. This key point, that not all yards are created equal, has been the foundation for the development of two approaches for evaluating plays: expected points and win probability. The expected points framework uses historical data to find the number of points eventually scored by teams in similar situations, while the win probability framework uses historical data to find how often teams in similar situations win the game. Using these metrics, one can obtain pre-snap and post-snap values of a play (expected points or win probability) and, taking the difference in these values, the value provided by the play itself – expected points added (EPA) or win probability added (WPA). These approaches have been recently popularized by Brian Burke's work at `www.advancedfootballanalytics.com` and ESPN (Burke, 2009; Katz and Burke, 2017).

Most of the best known approaches for calculating expected points do not provide any level of statistical detail describing their methodology. In most written descriptions, factors such as the down, yards to go for a first down, and field position are taken into account. However, there is no universal standard for which factors should be considered. Carter and Machol (1971) and others essentially use a form of "nearest neighbors" algorithms (Dasarathy, 1991) to identify similar situations based on down, yards to go, and the yard line to then average over the next points scored. Goldner (2017) describes a Markov model and uses the absorption probabilities for different scoring events (touchdown, field goal, and safety) to arrive at the expected points for a play. ESPN has a proprietary expected points metric, but does not detail the specifics of how it is calculated (Pattani, 2012). Burke (2009) provides an intuitive explanation for what expected points means, but does not go into the details of the calculations. Schatz (2003) provides a metric called "defense-adjusted value over average", which is similar to expected points, and also accounts for the strength of the opposing defense. However, specifics on the modeling techniques are not disclosed. Causey (2015) takes an exact-neighbors approach, finding all plays with a set of identical characteristics, taking the average outcome, and conducting post-hoc smoothing to calculate expected points. In this work, Causey explores the uncertainty in estimates of expected points using bootstrap resampling and analyzes the changes in expected point values over time. Causey also provides all code used for this analysis.

Depending on how metrics based on expected points are used, potential problems arise when building an expected points model involving the nature of football games. The main issue, as pointed out by Burke (2014), involves the score differential in a game. When a team is leading by a large number of points at the end of a game, they will sacrifice scoring points for letting time run off the clock. Changes in team behavior in these situations and, more generally, the leverage of a play in terms of its potential effect on winning and losing are not taken into account when computing expected points.

Analyzing changes in win probability for play evaluation partially resolves these issues. Compared to expected points models, there is considerably more literature on different methodologies for estimating the win probability of a play in football. Goldner (2017) uses a Markov model, similar to the approach taken by Tango, Lichtman, and Dolphin (2007) in baseball, by including the score differential, time remaining, and timeouts to extend the expected points model. Burke's approach is primarily empirical estimation by binning plays with adjustments and smoothing. In some published win probability analyses, random forests have been shown to generate well-calibrated win probability estimates (Causey, 2013; Lock and Nettleton, 2014). The approach taken by Lock and Nettleton (2014) also considers the respective strengths of the offensive (possession) and defensive (non-possession) teams.

There are many areas of research that build off of these approaches for valuing plays. For example, analyses of fourth down attempts and play-calling are very popular (Romer, 2006; Alamar, 2010; Goldner, 2012; Quealy et al., 2017). This paper focuses on using play evaluation to subsequently evaluate players, and we discuss prior attempts at player evaluation below.

1.2 Previous Work: Evaluating Players

Due to the complex nature of the sport and the limited data available publicly, the NFL lacks comprehensive statistics for evaluating player performance. While there has been extensive research

on situational analysis and play evaluation as described above, there has been considerably less focus on player evaluation. Existing measures do not accurately reflect a player's value to NFL teams, and they are not interpretable in terms of game outcomes (e.g. points or wins). Similarly, there are no publicly known attempts for developing a *Wins Above Replacement* (*WAR*) measure for every individual NFL player, as made popular in baseball (Schoenfield, 2012) and other sports (Thomas and Ventura, 2015).

Previous methods for player evaluation in football can be broken down into three categories: within-position statistical comparisons, ad hoc across-position statistical comparisons, and across-position statistical comparisons that rely on proprietary data or human judgment.

1.2.1 Within-Position Player Evaluation

Approaches for quantitatively evaluating players who play the same position are numerous, vary by position, and typically lag behind those of other sports. For comparisons of players at offensive skill positions such as quarterback (QB), running back (RB), wide receiver (WR), and tight end (TE), most analysis relies on basic box score statistics. These include yards gained via passing, rushing, and/or receiving; touchdowns via passing, rushing, and/or receiving; rushing attempts for RBs; receptions and targets for RBs, WRs, and TEs; completions, attempts, completion percentage, and yard per attempt for QBs; and other similar derivations of simple box score statistics. These metrics do not account for game situation or leverage. Additionally, they only provide an estimate of a player's relative value to other players at the same position. We cannot draw meaningful conclusions about cross-positional comparisons.

Linear combinations of these box score statistics, such as passer rating (Smith et al., 1973), are often used to compare players at the same position while taking into account more than just a single box score measure. Similarly, Pro Football Reference's adjusted net yards per attempt ("ANY/A") expands upon passer rating in that it accounts for sacks and uses a different linear weighting scheme (Pro-Football-Reference, 2018). These metrics involve outdated and/or ad hoc weights, thresholds, and other features. Passing in the NFL has changed substantially since the conception of the passer rating statistic in 1973, so that the chosen weights and thresholds do not have the same meaning in today's game as they did in 1973. While ANY/A accounts for sacks and uses a different weighting system, it is hardly a complete measure of QB performance, since it does not account for game situation and leverage. Perhaps most importantly, both passer rating and ANY/A are not interpretable in terms of game outcomes like points or wins.

For positions other than QB, RB, WR, and TE, data is limited, since the NFL does not publicly provide information about which players are on the field for a particular play, the offensive and defensive formations (other than the "shotgun" formation on offense), or the pre- and post-snap locations of players on the field. For offensive linemen, very little information is available to statistically compare players, as offensive linemen typically only touch the football on broken plays. For defensive players, the NFL only provides information about which players were directly involved in a play (e.g. the tackler or the defensive back covering a targeted receiver). As such, with these positions, it is difficult to obtain adequate within-positional comparisons of player value, let alone across-position comparisons.

1.2.2 Ad Hoc Across-Position Player Evaluation

Using only box score statistics, it is extremely difficult to ascertain the value of players at different positions. The fantasy sports industry has attempted to provide across-position estimates of player value using box score statistics. These estimates typically use ad hoc linear combinations of box score statistics that differ by position, so as to put the in-game statistical performances of players at different positions on comparable scales. These measures, typically referred to as "fantasy points", are available for all positions except those on the offensive line.

Of course, these metrics have several issues. First, they involve many unjustified or ad hoc weights. For example, one rushing yard is worth about 40% of one passing yard in ESPN's standard definitions of these metrics (ESPN, 2017), but these relative values are arbitrary. Second, the definitions are inconsistent, with different on-field events having different values for players of different positions. For example, defensive interceptions are typically worth three times as much as quarterback interceptions thrown (Ratcliffe, 2013; ESPN, 2017). Third, these measures do not account for context, such as the game situation or the leverage of a given play. Finally, they are not directly interpretable in terms of game outcomes (e.g. points or wins).

1.2.3 Player Evaluation with Proprietary Data or Human Judgment

Outside of the public sphere, there have been irreproducible attempts at within-position statistical comparisons of NFL players. Pro Football Focus assigns grades to every player in each play, but this approach is solely based on human judgment and proprietary to PFF (Eager et al., 2017). ESPN's total quarterback rating ("QBR") accounts for the situational contexts a QB faces throughout a game (Katz and Burke, 2017; Oliver, 2011). ESPN uses the following approach when computing QBR: First, they determine the degree of success or failure for each play. Second, they divide credit for each play amongst all players involved. Third, additional adjustments are made for plays of very little consequence to the game outcome. This approach has several important advantages. In the first step, the EPA is used to assign an objective value to each play. Another advantage is that some attempt is made to divide credit for a play's success or failure amongst the players involved. In the approach for NFL player evaluation we propose in this paper, we loosely follow these same two steps.

ESPN's QBR has some disadvantages, however. First and most importantly, Total QBR is not directly reproducible, since it relies on human judgment when evaluating plays. "The details of every play (air yards, drops, pressures, etc.) are charted by a team of trained analysts in the ESPN Stats & Information Group. Every play of every game is tracked by at least two different analysts to provide the most accurate representation of how each play occurred" (Katz and Burke, 2017). Additionally, while QBR down-weights plays in low-leverage situations, the approach for doing so is not clearly described and appears to be ad hoc. Finally, QBR is limited only to the QB position.

The only public approach for evaluating players at all positions according to common scale is Pro Football Reference's "approximate value" (AV) statistic (Drinen, 2013). Using a combination of objective and subjective analysis, AV attempts to assign a single numerical value to a player's performance in any season since 1950, regardless of the player's position. AV has some subjective components, such as whether or not a lineman was named to the NFL's "all-pro" team, and whether a running back reaches the arbitrary threshold of 200 carries. Additionally, since AV

uses linear combinations of end-of-season box score statistics to evaluate players, it does not take into account game situation, opponent, or many other contextual factors that may play a role in the accumulation of box score statistics over the course of a season. Finally, although the basis of many AV calculations involves points scored and allowed, AV is not interpretable in terms of game outcomes.

1.3   Our Framework for Evaluating NFL Plays and Players

In order to properly evaluate players, we need to allocate a portion of a play's value to each player involved (Katz and Burke, 2017). Baumer and Badian-Pessot (2017) details the long history of division of credit modeling as a primary driver of research in sports analytics, with origins in evaluating run contributions in baseball. However, in comparison to baseball, every football play is more complex and interdependent, with the 22 players on the field contributing in many ways and to varying degrees. A running play depends not only on the running back but the blocking by the linemen, the quarterback's handoff, the defensive matchup, the play call, etc. A natural approach is to use a regression-based method, with indicators for each player on the field for a play, providing an estimate of their marginal effect. This type of modeling has become common in basketball and hockey, because it accounts for factors such as quality of teammates and competition (Rosenbaum, 2004; Kubatko et al., 2007; Macdonald, 2011; Gramacy et al., 2012; Thomas et al., 2013).

We present four contributions to the study of football statistics in order to address the issues pertaining to play evaluation and player evaluation outlined above:

1. The `R` package `nflscrapR` to provide easy access to publicly available NFL play-by-play data (Section 2).

2. A novel approach for estimating expected points using a multinomial logistic regression model, which more appropriately models the "next score" response variable (Section 3.1).

3. A generalized additive model for estimating the win probability using the expected points as input (Section 3.2).

4. Our *nflWAR* framework, using multilevel models to isolate offensive skill player contribution and estimate their *WAR* (Section 4).

We use a sampling procedure similar to Baumer et al. (2015) to estimate uncertainty in each player's seasonal *WAR*. Due to the limitations of publicly available data, the primary focus of this paper is on offensive skill position players: QB, RB, WR, and TE. However, we present a novel metric that serves as a proxy for measuring a team's offensive line performance on rushing plays. Furthermore, the reproducible framework we introduce in this paper can also be easily extended to estimate *WAR* for all positions given the appropriate data. Researchers with data detailing which players are on the field for every play can use the framework provided in Section 6.4 to estimate *WAR* for players at all positions.

Our *WAR* framework has several key advantages. First, it is fully reproducible: it is built using only public data, with all code provided and all data accessible to the public. Second, our expected points and win probability models are well-calibrated and more appropriate from

a statistical perspective than other approaches. Third, player evaluation with *WAR* is easily interpretable in terms of game outcomes, unlike prior approaches to player evaluation in the NFL discussed above. The replacement level baseline informs us how many wins a player adds over a readily available player. This is more desirable than comparing to average from the viewpoint of an NFL front office, as league average performance is still valuable in context (Baumer et al., 2015). Fourth, the multilevel model framework accounts for quality of teammates and competition. Fifth, although this paper presents *WAR* using our expected points and win probability models for play evaluation, researchers can freely substitute their own approaches for play evaluation without any changes to the framework for estimating player *WAR*. Finally, we recognize the limitations of point estimates for player evaluation and provide estimates of the uncertainty in a player's *WAR*.

2   Play-by-Play Data with `nflscrapR`

Data in professional sports comes in many different forms. At the season-level, player and team statistics are typically available dating back to the 1800s (Lahman, 1996 – 2017, Phillips (2018)). At the game-level, player and team statistics have been tracked to varying degrees of detail dating back several decades (Lahman, 1996 – 2017). Within games, data is available to varying degrees of granularity across sports and leagues. For example, Major League Baseball (MLB) has play-by-play data at the plate appearance level available dating back several decades (Lahman, 1996 – 2017), while the National Hockey League (NHL) only began releasing play-by-play via their real-time scoring system in the 2005-06 season (Thomas and Ventura, 2017).

Play-by-play data, or information specifying the conditions, features, and results of an individual play, serves as the basis for most modern sports analysis in the public sphere (Kubatko et al., 2007, Macdonald (2011), Lock and Nettleton (2014), Thomas and Ventura (2017)). Outside of the public sphere, many professional sports teams and leagues have access to data at even finer levels of granularity, e.g. via optical player tracking systems in the National Basketball Association, MLB, and the English Premier League that track the spatial distribution of players and objects at multiple times per second. The NFL in 2016 began using radio-frequency identification (RFID) technology to track the locations of players and the football (Skiver, 2017), but as of mid 2018, this data is not available publicly, and NFL teams have only just gained accessed to the data beyond their own players. In almost all major professional sports leagues, play-by-play data is provided and includes information on in-game events, players involved, and (usually) which players are actively participating in the game for each event (Thomas and Ventura, 2017, Lahman (1996 – 2017)).

Importantly, this is not the case for the NFL. While play-by-play data is available through the NFL.com application programming interface (API), the league does not provide information about which players are present on the playing field for each play, what formations are being used (aside from the "shotgun" formation), player locations, or pre-snap player movement. This is extremely important, as it limits the set of players for which we can provide estimates of their contribution to game outcomes (e.g. points scored, points allowed, wins, losses, etc).

We develop an `R` package (R Core Team, 2017), called `nflscrapR`, that provides users with clean datasets, box score statistics, and more advanced metrics describing every NFL play since

2009 (Horowitz et al., 2017). This package was inspired largely by other R packages facilitating the access of sports data. For hockey, `nhlscrapR` provides clean datasets and advanced metrics to use for analysis for NHL fans (Thomas and Ventura, 2017). In baseball, the R packages `pitchRx` (Sievert, 2015), `Lahman` (Lahman, 1996 – 2017), and `openWAR` (Baumer et al., 2015) provide tools for collecting MLB data on the pitch-by-pitch level and building out advanced player evaluation metrics. In basketball, `ballR` (Elmore and DeWitt, 2017) provides functions for collecting data from `basketball-reference.com`.

Each NFL game since 2009 has a 10 digit game identification number (ID) and an associated set of webpages that includes information on the scoring events, play-by-play, game results, and other game data. The API structures its data using JavaScript Object Notation (JSON) into three major groups: game outcomes, player statistics at the game level, and play-by-play information. The design of the `nflscrapR` package closely mimics the structure of the JSON data in the API, with four main functions described below:

`season_games()`: Using the data structure outputting end of game scores and team matchups, this function provides end of game results with an associated game ID and the home and away teams abbreviations.

`player_game()`: Accessing the player statistics object in the API's JSON data, this function parses the player level game summary data and creates a box-score-like data frame. Additional functions provide aggregation functionality: `season_player_game()` binds the results of `player_game()` for all games in a season, and `agg_player_season()` outputs a single row for each player with their season total statistics.

`game_play_by_play()`: This is the most important function in `nflscrapR`. The function parses the listed play-by-play data then uses advanced regular expressions and other data manipulation tasks to extract detailed information about each play (e.g. players involved in action, play type, penalty information, air yards gained, yards gained after the catch, etc.). The `season_play_by_play()` binds the results of `game_play_by_play()` for all games in a season.

`season_rosters()`: This function outputs all of the rostered players on a specified team in a specified season and includes their name, position, unique player ID, and other information.

For visualization purposes we also made a dataset, `nflteams` available in the package which includes the full name of all 32 NFL teams, their team abbreviations, and their primary colors[1].

In addition to the functions provided in `nflscrapR`, we provide downloadable versions in comma-separated-value format, along with a complete and frequently updating data dictionary, at `https://github.com/ryurko/nflscrapR-data`. The datasets provided on this website included play-by-play from 2009 – 2017, game-by-game player level statistics, player-season total statistics, and team-season total statistics. These datasets are made available to allow users familiar with other software to do research in the realm of football analytics. Table 1 gives a brief overview of some of the more important variables used for evaluating plays in Section 3.

---

[1]Some of this information is provided through Ben Baumer's R package `teamcolors` (Baumer and Matthews, 2017)

8

Table 1: Description of the play-by-play dataset.

Variable	Description
Possession Team	Team with the ball on offense (opposing team is on defense)
Down	Four downs to advance the ball ten (or more) yards
Yards to go	Distance in yards to advance and convert first down
Yard line	Distance in yards away from opponent's endzone (100 to zero)
Time Remaining	Seconds remaining in game, each game is 3600 seconds long (four quarters, halftime, and a potential overtime)
Score differential	Difference in score between the possession team and opposition

3 Evaluating Plays with Expected Points and Win Probability

As described in Section 1.1, expected points and win probability are two common approaches for evaluating plays. These approaches have several key advantages: They can be calculated using only data provided by the NFL and available publicly, they provide estimates of a play's value in terms of real game outcomes (i.e. points and wins), and, as a result, they are easy to understand for both experts and non-experts.

Below, we introduce our own novel approaches for estimating expected points (EP) and win probability (WP) using publicly available data via `nflscrapR`.

3.1 Expected Points

While most authors take the average "next score" outcome of similar plays in order to arrive at an estimate of EP, we recognize that certain scoring events become more or less likely in different situations. As such, we propose modeling the probability for each of the scoring events directly, as this more appropriately accounts for the differing relationships between the covariates in Table 1 and the different categories of the "next score" response. Once we have the probabilities of each scoring event, we can trivially estimate expected points.

3.1.1 Multinomial Logistic Regression

To estimate the probabilities of each possible scoring event conditional on the current game situation, we use multinomial logistic regression. For each play, we find the next scoring event within the same half (with respect to the possession team) as one of the seven possible events: touchdown (7 points), field goal (3 points), safety (2 points), no score (0 points), opponent safety (-2 points), opponent field goal (-3 points), and opponent touchdown (-7 points). Here, we ignore

Table 2: Description of variables for the *EP* model.

Variable	Variable description
Down	The current down (1st, 2nd, 3rd, or 4th
Seconds	Number of seconds remaining in half
Yardline	Yards from endzone (0 to 100)
log(YTG)	Log transformation of yards to go for a first down
GTG	Indicator for whether or not it is a goal down situation
UTM	Indicator for whether or not time remaining in the half is under two minutes

point after touchdown (PAT) attempts, and we treat PATs separately in Section 3.1.4.

Figure 1 displays the distribution of the different type of scoring events using data from NFL regular season games between 2009 and 2016, with each event located on the y-axis based on their associated point value *y*. This data consists of 304,896 non-PAT plays, excluding QB kneels (which are solely used to run out the clock and are thus assigned an *EP* value of zero). The gaps along the y-axis between the different scoring events reinforce our decision to treat this as a classification problem rather than modeling the point values with linear regression – residuals in such a model will not meet the assumptions of normality. While we use seven points for a touchdown for simplicity here, our multinomial logistic regression model generates the probabilities for the events agnostic of the point value. This is beneficial, since it allows us to flexibly handle PATs and two-point attempts separately. We can easily adjust the point values associated with touchdowns to reflect changes in the league's scoring environment.
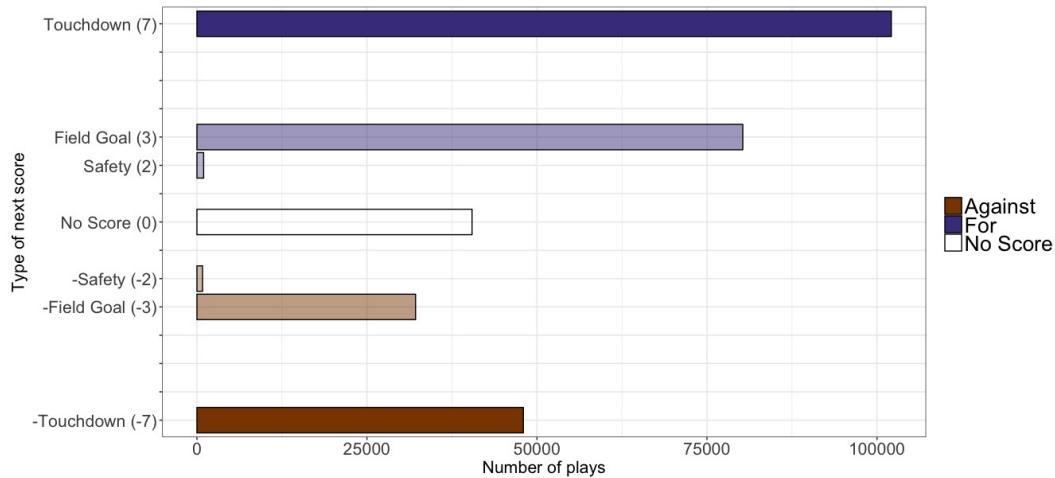

Figure 1: Distribution of next scoring events for all plays from 2009-16, with respect to the possession team.

We denote the covariates describing the game situation for each play as **X**, which are presented in Table 2, and the response variable:

$$Y \in \{\text{Touchdown (7), Field Goal (3), Safety (2), No Score (0),}$$
$$- \text{Touchdown (-7)}, -\text{Field Goal (-3)}, -\text{Safety (-2)}\} \tag{1}$$

The model is specified with six logit transformations relative to the "No Score" event with the following form:

$$\log\left(\frac{P(Y = Touchdown|\mathbf{X})}{P(Y = No\ Score|\mathbf{X})}\right) = \mathbf{X} \cdot \boldsymbol{\beta}_{Touchdown},$$

$$\log\left(\frac{P(Y = Field\ Goal|\mathbf{X})}{P(Y = No\ Score|\mathbf{X})}\right) = \mathbf{X} \cdot \boldsymbol{\beta}_{Field\ Goal},$$

$$\vdots \tag{2}$$

$$\log\left(\frac{P(Y = -Touchdown|\mathbf{X})}{P(Y = No\ Score|\mathbf{X})}\right) = \mathbf{X} \cdot \boldsymbol{\beta}_{-Touchdown},$$

where $\boldsymbol{\beta}_y$ is the corresponding coefficient vector for the type of next scoring event. Using the generated probabilities for each of the possible scoring events, $P(Y = y|\mathbf{X})$, we simply calculate the expected points (*EP*) for a play by multiplying each event's predicted probability with its associated point value y:

$$EP = E[Y|\mathbf{X}] = \sum_y y \cdot P(Y = y|\mathbf{X}). \tag{3}$$

3.1.2 Observation Weighting

Potential problems arise when building an expected points model because of the nature of football games. The first issue, as pointed out by Burke (2014), regards the score differential in a game. When a team is leading by a large number of points at the end of a game they will sacrifice scoring points for letting time run off the clock. This means that plays with large score differentials can exhibit a different kind of relationship with the next points scored than plays with tight score differentials. Although others such as Burke only use the subset of plays in the first and third quarter where the score differential is within ten points, we don't exclude any observations but instead use a weighting approach. Figure 2(a) displays the distribution for the absolute score differential, which is clearly skewed right, with a higher proportion of plays possessing smaller score differentials. Each play $i \in \{1, \ldots, n\}$, in the modeling data of regular season games from 2009 to 2016, is assigned a weight w_i based on the score differential S scaled from zero to one with the following function:

$$w_i = w(S_i) = \frac{\max_i(|S_i|) - |S_i|}{\max_i(|S_i|) - \min_i(|S_i|)}. \tag{4}$$

In addition to score differential, we also weight plays according to their "distance" to the next score in terms of the number of drives. For each play i, we find the difference in the number of

drives from the next score D: $D_i = d_{next\ score} - d_i$, where $d_{next\ score}$ and d_i are the drive numbers for the next score and play i, respectively. For plays in the first half, we stipulate that $D_i = 0$ if the $d_{next\ score}$ occurs in the second half, and similarly for second half plays for which the next score is in overtime. Figure 2(b) displays the distribution of D_i excluding plays with the next score as "No Score." This difference is then scaled from zero to one in the same way as the score differential in Equation 4. The score differential and drive score difference weights are then added together and again rescaled from zero to one in the same manner resulting in a combined weighting scheme. By combining the two weights, we are placing equal emphasis on both the score differential and the number of drives until the next score and leave adjusting this balance for future work.
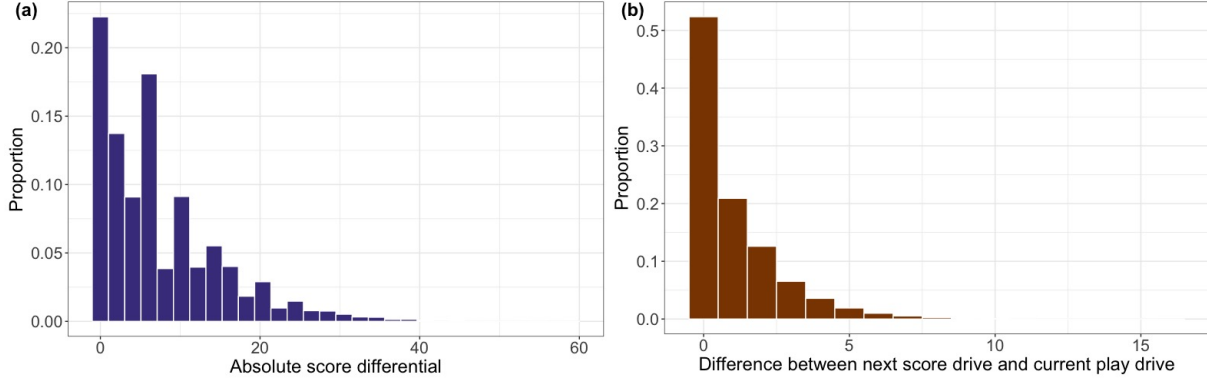
Figure 2: Distributions for (a) absolute score differential and (b) number of drives until next score (excluding plays without a next score event).

3.1.3 Model Selection with Calibration

Since our expected points model uses the probabilities for each scoring event from multinomial logistic regression, the variables and interactions selected for the model are determined via calibration testing, similar to the criteria for evaluating the win probability model in Lock and Nettleton (2014). The estimated probability for each of the seven scoring events is binned in five percent increments (20 total possible bins), with the observed proportion of the event found in each bin. If the actual proportion of the event is similar to the bin's estimated probability then the model is well-calibrated. Because we are generating probabilities for seven events, we want a model that is well-calibrated across all seven events. To objectively compare different models, we first calculate for scoring event y in bin $b \in \{1, \ldots, B\}$ its associated error $e_{y,b}$:

$$e_{y,b} = |\hat{P}_b(Y = y) - P_b(Y = y)|, \tag{5}$$

where $\hat{P}_b(Y = y)$ and $P_b(Y = y)$ are the predicted and observed probabilities, respectively, in bin b. Then, the overall calibration error e_y for scoring event y is found by averaging $e_{y,b}$ over all bins, weighted by the number of plays in each bin, $n_{y,b}$:

$$e_y = \frac{1}{n_y} \sum_b n_{y,b} \cdot e_{y,b}, \tag{6}$$

12

where $n_y = \sum_b n_{y,b}$. This leads to the model's calibration error e as the average of the seven e_y values, weighted by the number of plays with scoring event y, n_y:

$$e = \frac{1}{n} \sum_y n_y \cdot e_y, \tag{7}$$

where $n = \sum_y n_y$, the number of total plays. This provides us with a single statistic with which to evaluate models, in addition to the calibration charts.

We calculate the model calibration error using leave-one-season-out cross-validation (LOSO CV) to reflect how the `nflscrapR` package will generate the probabilities for plays in a season it has not yet observed. The model yielding the best LOSO CV calibration results uses the variables presented in Table 2, along with three interactions: log(YTG) and Down, Yardline and Down, and log(YTG) and GTG. Figure 3 displays the selected model's LOSO CV calibration results for each of the seven scoring events, resulting in $e \approx 0.013$. The dashed lines along the diagonal represent a perfect fit, i.e. the closer to the diagonal points are the more calibrated the model. Although time remaining is typically reserved for win probability models (Goldner, 2017), including the seconds remaining in the half, as well as the indicator for under two minutes, improved the model's calibration, particularly with regards to the "No Score" event. We also explored the use of an ordinal logistic regression model which assumes equivalent effects as the scoring value increases, but found the LOSO CV calibration results to be noticeably worse with $e \approx 0.022$.
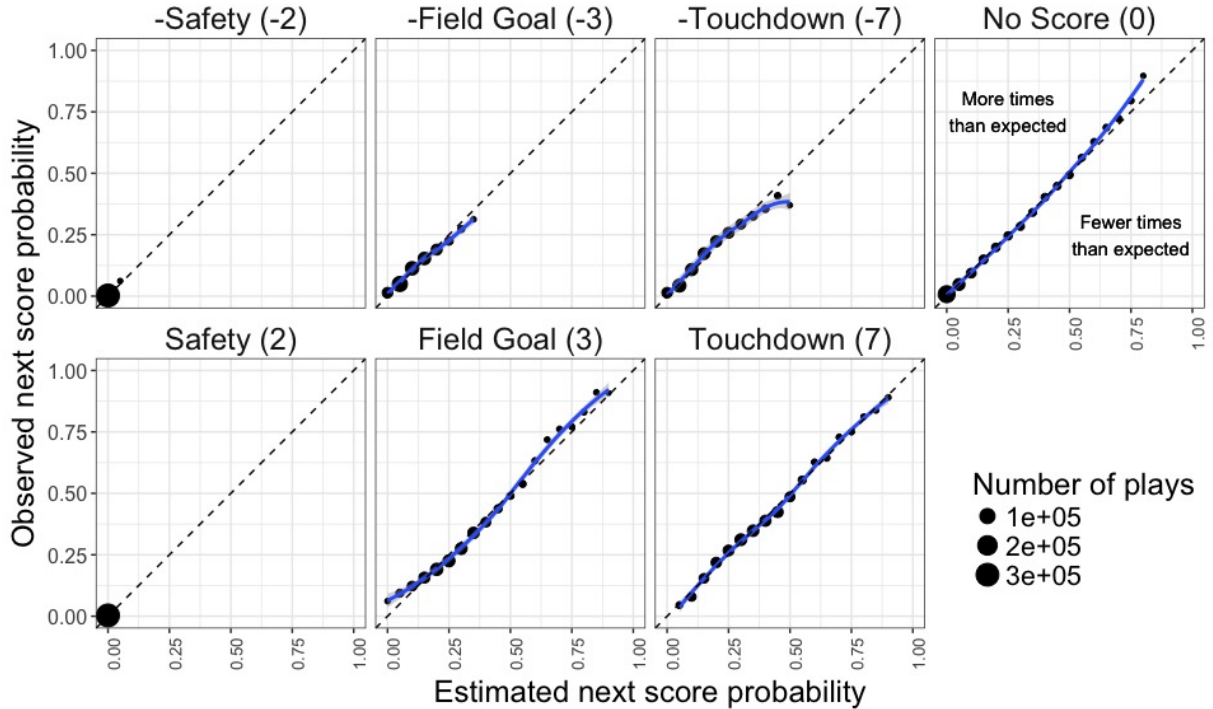

Figure 3: Expected points model LOSO CV calibration results by scoring event.

3.1.4 PATs and Field Goals

As noted earlier, we treat PATs (extra point attempts and two-point attempts) separately. For two-point attempts, we simply use the historical success rate of 47.35% from 2009-2016, resulting in $EP = 2 \cdot 0.4735 = 0.9470$. Extra point attempts use the probability of successfully making the kick from a generalized additive model (see Section 3.2.1) that predicts the probability of making the kick, $P(M)$ for both extra point attempts and field goals as a smooth function of the kick's distance, k (total of 16,906 extra point and field goal attempts from 2009-2016):

$$\log\left(\frac{P(M)}{1 - P(M)}\right) = s(k). \tag{8}$$

The expected points for extra point attempts is this predicted probability of making the kick, since the actual point value of a PAT is one. For field goal attempts, we incorporate this predicted probability of making the field goal taking into consideration the cost of missing the field goal and turning the ball over to the opposing team. This results in the following override for field goal attempts:

$$EP_{field\ goal\ attempt} = P(M) \cdot 3 + (1 - P(M)) \cdot (-1) \cdot E[Y|X = m], \tag{9}$$

where $E[Y|X = m]$ is the expected points from the multinomial logistic regression model but assuming the opposing team has taken possession from a missed field goal, with the necessary adjustments to field position and time remaining (eight yards and 5.07 seconds, respectively, estimated from NFL regular season games from 2009 to 2016), and multiplying by negative one to reflect the expected points for the team attempting the field goal. Although these calculations are necessary for proper calculation of the play values $\delta_{f,i}$ discussed in Section 3.3, we note that this is a rudimentary field goal model only taking distance into account. Enhancements could be made with additional data (e.g. weather data, which is not made available by the NFL) or by using a model similar to that of Morris (2015), but these are beyond the scope of this paper.

3.1.5 Expected Points by Down and Yard Line

For reference, Figure 4 displays the relationship between the field position and the EP for our multinomial logistic regression model available via `nflscrapR` compared to the previous relationships found by Carter and Machol (1971) and Carroll et al. (1988). We separate the `nflscrapR` model by down to show its importance, and in particular the noticeable drop for fourth down plays and how they exhibit a different relationship near the opponent's end zone as compared to other downs. To provide context for what is driving the difference, Figure 5 displays the relationship between each of the next score probabilities and field position by down. Clearly on fourth down, the probability of a field goal attempt overwhelms the other possible events once within 50 yards of the opponent's end zone.

3.2 Win Probability

Because our primary focus in this paper is in player evaluation, we model win probability without taking into account the teams playing (i.e. we do not include indicators for team strength in the
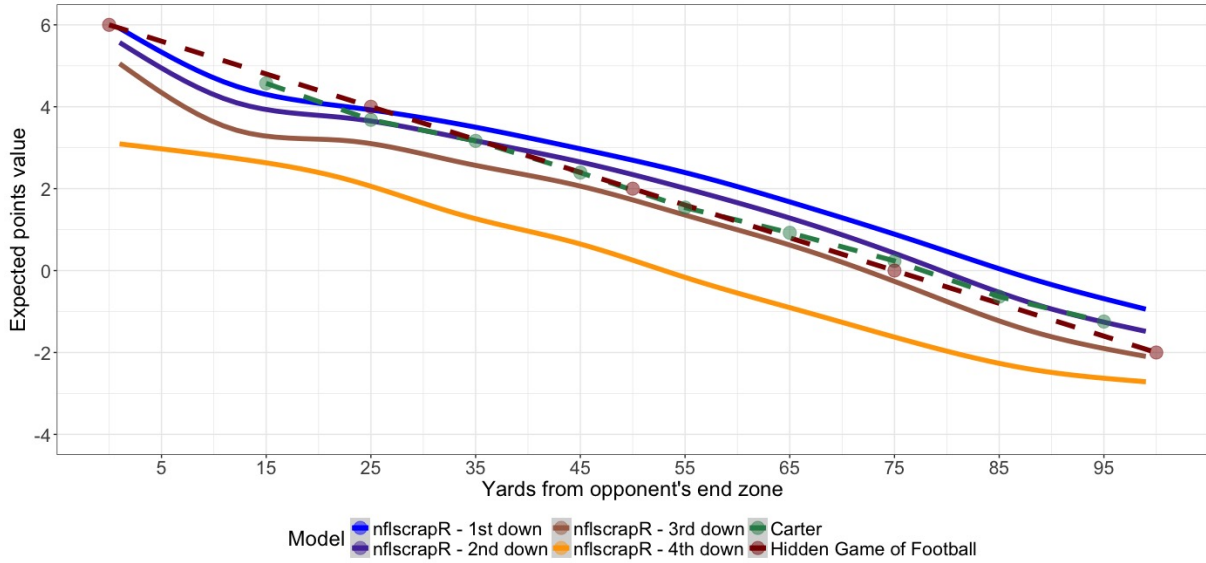
14

Figure 4: Comparison of historical models and `nflscrapR` expected points value, based on distance from opponent's end zone by down.
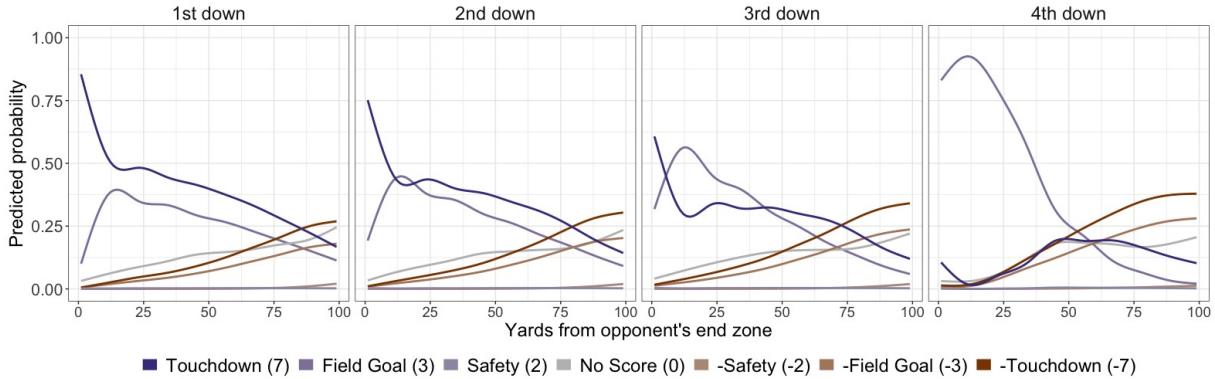

Figure 5: Relationship between next score event probabilities and field position by down.

win probability model). As a result, every game starts with each team having a 50% chance of winning. Including indicators for a team's overall, offensive, and/or defensive strengths would artificially inflate (deflate) the contributions made by players on bad (good) teams in the models described in Section 4, since their team's win probability would start lower (higher).

Our approach for estimating WP also differs from the others mentioned in Section 1.1 in that we incorporate the estimated EP directly into the model by calculating the expected score differential for a play. Our expected points model already produces estimates for the value of the field position, yards to go, etc without considering which half of the game or score. When including the variables presented in Table 3, we arrive at a well-calibrated WP model.

Table 3: Description of selected variables for the win probability model. Note: S is the score differential at the current play.

Variable	Variable description
$E[S]$	Expected score differential $= EP + S$
s_g	Number of seconds remaining in game
$E[\frac{S}{s_g+1}]$	Expected score time ratio
h	Current half of the game (1st, 2nd, or overtime)
s_h	Number of seconds remaining in half
u	Indicator for whether or not time remaining in half is under two minutes
t_{off}	Time outs remaining for offensive (possession) team
t_{def}	Time outs remaining for defensive team

3.2.1 Generalized Additive Model

We use a generalized additive model (GAM) to estimate the possession team's probability of winning the game conditional on the current game situation. GAMs have several key benefits that make them ideal for modeling win probability: They allow the relationship between the explanatory and response variables to vary according to smooth, non-linear functions. They also allow for linear relationships and can estimate (both ordered and unordered) factor levels. We find that this flexible, semi-parametric approach allows us to capture nonlinear relationships while maintaining the many advantages of using linear models. Using a logit link function, our WP model takes the form:

$$\log\left(\frac{P(\text{Win})}{P(\text{Loss})}\right) = s(E[S]) + s(s_h) \cdot h + s\left(E\left[\frac{S}{s_g+1}\right]\right) + h \cdot u \cdot t_{off} + h \cdot u \cdot t_{def}, \qquad (10)$$

where s is a smooth function while h, u, t_{off}, and t_{def} are linear parametric terms defined in Table 3. By taking the inverse of the logit we arrive at a play's WP.

3.2.2 Win Probability Calibration

Similar to the evaluation of the EP model, we again use LOSO CV to select the above model, which yields the best calibration results. Figure 6 shows the calibration plots by quarter, mimicking the approach of Lopez (2017) and Yam and Lopez (2018), who evaluate both our WP model and that of Lock and Nettleton (2014). The observed proportion of wins closely matches the expected proportion of wins within each bin for each quarter, indicating that the model is well-calibrated across all quarters of play and across the spectrum of possible win probabilities. These findings match those of Yam and Lopez (2018), who find "no obvious systematic patterns that would signal a flaw in either model."
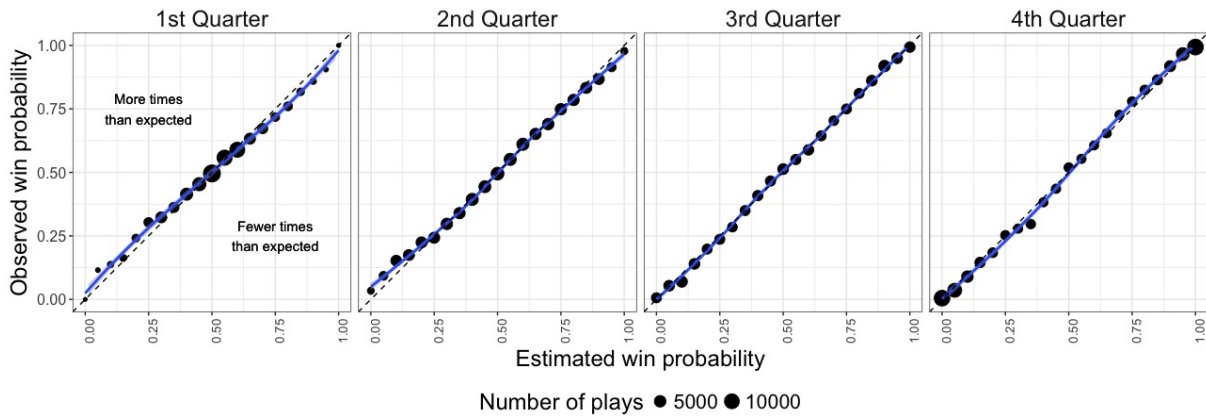
Figure 6: Win probability model LOSO CV calibration results by quarter.

3.2.3 Win Probability Example

An example of a single game WP chart is provided in Figure 7 for the 2017 American Football Conference (AFC) Wild Card game between the Tennessee Titans and Kansas City Chiefs. The game starts with both teams having an equal chance of winning, with minor variations until the score differential changes (in this case, in favor of Kansas City). Kansas City led 21-3 after the first half, reaching a peak win probability of roughly 95% early in the third quarter, before giving up 19 unanswered points in the second half and losing to Tennessee 22-21.
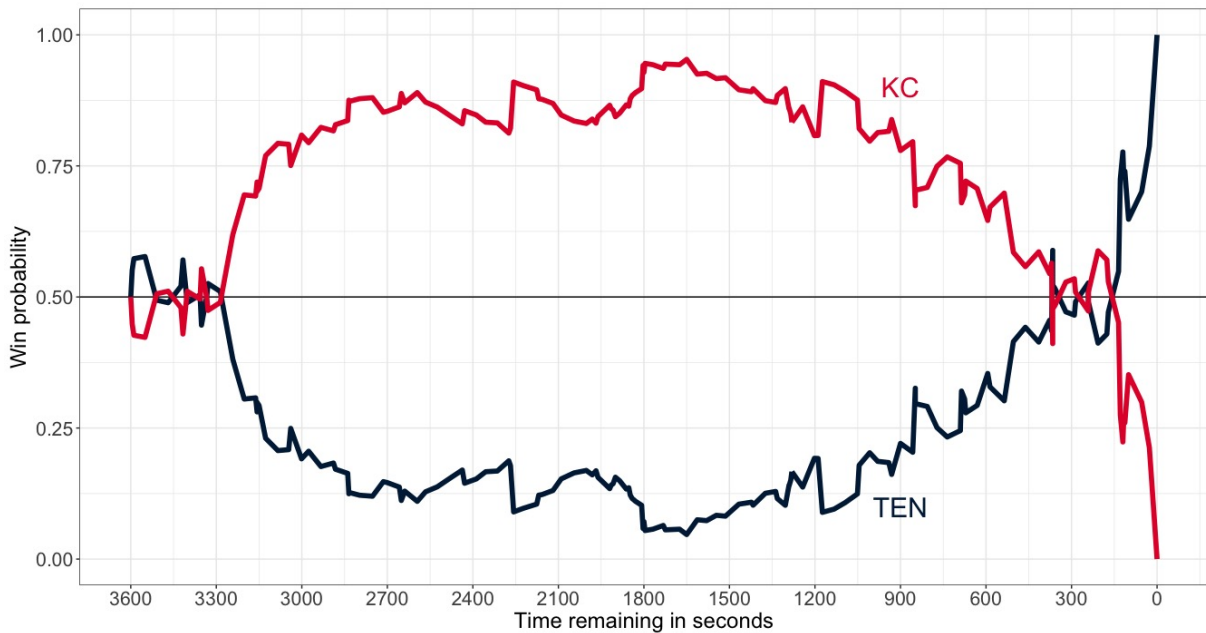
Figure 7: Win probability chart for 2017 AFC Wild Card game.

3.3 Expected Points Added and Win Probability Added

In order to arrive at a comprehensive measure of player performance, each play in a football game must be assigned an appropriate value $\delta_{f,i}$ that can be represented as the change from state i to state f:

$$\delta_{f,i} = \boldsymbol{V}_f - \boldsymbol{V}_i, \tag{11}$$

where $\boldsymbol{V}_f$ and $\boldsymbol{V}_i$ are the associated values for the ending and starting states respectively. We represent these values by either a play i's expected points (EP_i) or win probability (WP_i).

Plugging our EP and WP estimates for the start of play i and the start of the following play f into Equation 11's values for $\boldsymbol{V}_i$ and $\boldsymbol{V}_f$ respectively provides us with the two types of play valuations $\delta_{f,i}$: (1) the change in point value as expected points added (EPA), and (2) the change in win probability as win probability added (WPA). For scoring plays, we use the associated scoring event's value y as $\boldsymbol{V}_f$ in place of the following play's EP to reflect that the play's value is just connected to the difference between the scoring event and the initial state of the play. As an example, during Super Bowl LII the Philadelphia Eagles' Nick Foles received a touchdown when facing fourth down on their opponent's one yard line with thirty-eight seconds remaining in the half. At the start of the play the Eagles' expected points was $\boldsymbol{V}_i \approx 2.78$, thus resulting in $EPA \approx 7 - 2.78 = 4.22$. In an analogous calculation, this famous play known as the "Philly special" resulted in $WPA \approx 0.1266$ as the Eagles' increased their lead before the end of the half.

For passing plays, we can additionally take advantage of *air yards* (perpendicular distance in yards from the line of scrimmage to the yard line at which the receiver was targeted or caught the ball) and *yards after catch* (perpendicular distance in yards from the yard line at which the receiver caught the ball to the yard line at which the play ended), for every passing play available with `nflscrapR`. Using these two pieces, we can determine the hypothetical field position and whether or not a turnover on downs occurs to separate the value of a play from the air yards versus the yards after catch. For each completed passing play, we break the estimation of EP and WP into two plays – one comprising everything leading up to the catch, and one for the yards after the catch. Because the models rely on the seconds remaining in the game, we make an adjustment to the time remaining by subtracting the average length of time for incomplete passing plays, 5.7 seconds[2]. We then use the EP or WP through the air as $\boldsymbol{V}_f$ in Equation 11 to estimate $EPA_{i,air}$ or $WPA_{i,air}$, denoting these as $\delta_{f,i,air}$. We estimate the value of yards after catch, $\delta_{f,i,yac}$, by taking the difference between the value of the following play $\boldsymbol{V}_f$ and the value of the air yards, $\delta_{f,i,air}$. We use this approach to calculate both $EPA_{i,yac}$ and $WPA_{i,yac}$.

4 Evaluating Players with nflWAR

We use the play values calculated in Section 3 as the basis for a statistical estimate of wins above replacement (*WAR*) for each player in the NFL. To do this, we take the following approach:

- estimate the value of each play (Section 3),
- estimate the effect of each player on play value added (Section 4.1),

---

[2]This estimate could be improved in future work if information about the time between the snap and the pass becomes available.

- evaluate relative to replacement level (Section 4.2),
- convert to a wins scale (Section 4.3), and
- and estimate the uncertainty in *WAR* (Section 4.4).

This framework can be applied to any individual season, and we present results for the 2017 season in Section 5. Due to data restrictions, we currently are only able to produce *WAR* estimates for offensive skill position players. However, a benefit of our framework is the ability to separate a player's total value into the three components of WAR_{air}, WAR_{yac}, and WAR_{rush}. Additionally, we provide the first statistical estimates for a team's rush blocking based on play-by-play data.

4.1   Division of Credit

In order to properly evaluate players, we need to allocate the portion of a play's value $\delta_{f,i}$ to each player on the field. Unfortunately, the NFL does not publicly specify which players are on the field for every play, preventing us from directly applying approaches similar to those used in basketball and hockey discussed in Section 1.2, where the presence of each player on the playing surface is treated as an indicator covariate in a linear model that estimates the marginal effect of that player on some game outcome (Kubatko et al., 2007; Macdonald, 2011; Thomas et al., 2013). Instead, the data available publicly from the NFL and obtained via `nflscrapR` is limited to only those players directly involved in the play, plus contextual information about the play itself. For rushing plays, this includes:

- Players: rusher and tackler(s)
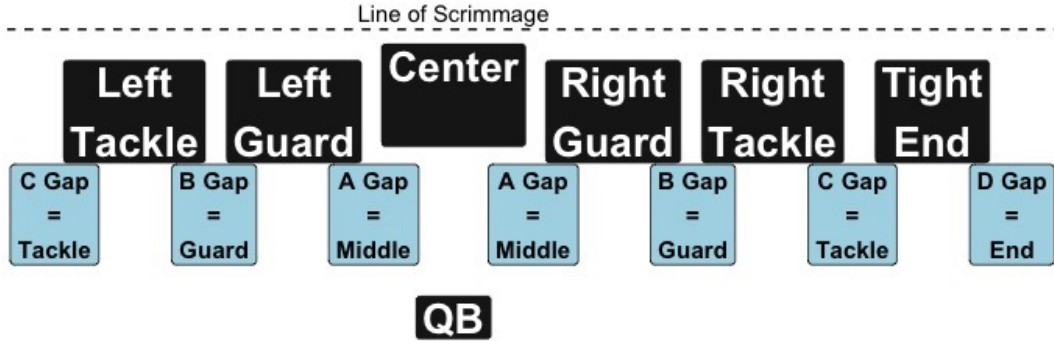- Context: run gap (end, tackle, guard, middle) and direction (left, middle, right)

Figure 8: Offensive Line Gaps for Rushing Plays.

Figure 8 provides a diagram of the run gaps (in blue) and the positions along the offensive line (in black). In the NFL play-by-play, the gaps are not referred to with letters, as they commonly are by football players and coaches; instead, the terms "middle", "guard", "tackle", and "end" are used. For the purposes of this paper, we define the following linkage between these two nomenclatures:

- "A" Gap = "middle"

19

- "B" Gap = "guard"
- "C" Gap = "tackle"
- "D" Gap = "end"

For passing plays, information about each play includes:

- Players: passer, targeted receiver, tackler(s), and interceptor
- Context: air yards, yards after catch, location (left, middle, right), and if the passer was hit on the play.

4.1.1 Multilevel Modeling

All players in the NFL belong to positional groups that dictate how they are used in the context of the game. For example, for passing plays we have the QB and the targeted receiver. However, over the course of an NFL season, the average QB will have more pass attempts than the average receiver will have targets, because there are far fewer QBs (more than 60 with pass attempts in the 2017 NFL season) compared to receivers (more than 400 targeted receivers in the 2017 season).

Because of these systematic differences across positions, there are differing levels of variation in each position's performance. Additionally, since every play involving the same player is a repeated measure of performance, the plays themselves are not independent.

To account for these structural features of football, we use a multilevel model (also referred to as hierarchical, random-effects, or mixed-effects model), which embraces this positional group structure and accounts for the observation dependence. Multilevel models have recently gained popularity in baseball statistics due to the development of catcher and pitcher metrics by Baseball Prospectus (Judge et al., 2015a,b), but have been used in sports dating back at least to 2013 (Thomas et al., 2013). Here, we novelly extend their use for assessing offensive player contributions in football, using the play values $\delta_{f,i}$ from Section 3 as the response.

In order to arrive at individual player effects we use varying-intercepts for the groups involved in a play. A simple example of modeling $\delta_{f,i}$ with varying-intercepts for two groups, QBs as Q and receivers as C, with covariates X_i and coefficients β is as follows:

$$\delta_{f,i} \sim Normal(Q_{q[i]} + C_{c[i]} + X_i \cdot \beta, \ \sigma_\delta^2), \ for \ i \ = \ 1, \dots, n \text{ plays}, \tag{12}$$

where the key feature distinguishing multilevel regression from classical regression is that the group coefficients vary according to their own model:

$$Q_q \sim Normal(\mu_Q, \ \sigma_Q^2), \text{ for } q \ = \ 1, \dots, \# \text{ of QBs},$$
$$C_c \sim Normal(\mu_C, \ \sigma_C^2), \text{ for } c \ = \ 1, \dots, \# \text{ of receivers}. \tag{13}$$

By assigning a probability distribution (such as the Normal distribution) to the group intercepts, Q_q and C_c, with parameters estimated from the data (such as μ_Q and σ_Q for passers), each estimate is pulled toward their respective group mean levels μ_Q and μ_C. In this example, QBs and receivers involved in fewer plays will be pulled closer to their overall group averages as compared to those involved in more plays and thus carrying more information, resulting in partially pooled estimates (Gelman and Hill, 2007). This approach provides us with average individual effects on

Table 4: Description of variables in the models assessing player and team effects.

Variable name	Variable description
Home	Indicator for if the possession team was home
Shotgun	Indicator for if the play was in shotgun formation
NoHuddle	Indicator for if the play was in no huddle
QBHit	Indicator for if the QB was hit on a pass attempt
PassLocation	Set of indicators for if the pass location was either middle or right (reference group is left)
AirYards	Orthogonal distance in yards from the line of scrimmage to where the receiver was targeted or caught the ball
RecPosition	Set of indicator variables for if the receiver's position was either TE, FB, or RB (reference group is WR)
RushPosition	Set of indicator variables for if the rusher's position was either FB, WR, or TE (reference group is RB)
PassStrength	EPA per pass attempt over the course of the season for the possession team
RushStrength	EPA per rush attempt over the course of the season for the possession team

play value added while also providing the necessary shrinkage towards the group averages. All models we use for division of credit are of this varying-intercept form, and are fit using penalized likelihood via the `lme4` package in R (Bates et al., 2015). While these models are not explicitly Bayesian, as Gelman and Hill (2007) write, "[a]ll multilevel models are Bayesian in the sense of assigning probability distributions to the varying regression coefficients", meaning we're taking into consideration all members of the group when estimating the varying intercepts rather than just an individual effect.

Our assumption of normality for $\delta_{f,i}$ follows from our focus on *EPA* and *WPA* values, which can be both positive and negative, exhibiting roughly symmetric distributions. We refer to an intercept estimating a player's average effect as their *individual points/probability added* (*iPA*), with points for modeling *EPA* and probability for modeling *WPA*. Similarly, an intercept estimating a team's average effect is their *team points/probability added* (*tPA*). Tables 4 and 5 provide the notation and descriptions for the variables and group terms in the models apportioning credit to players and teams on plays. The variables in Table 4 would be represented by X, and their effects by β in Equation 12.

Table 5: Description of groups in the models assessing player and team effects.

Group	Individual	Description
Q	q	QB attempting a pass or rush/scramble/sack
C	c	Targeted receiver on a pass attempt
H	ι	Rusher on a rush attempt
T	τ	Team-side-gap on a rush attempt, combination of the possession team, rush gap and direction
F	v	Opposing defense of the pass

4.1.2 Passing Models

Rather than modeling the $\delta_{f,i}$ (*EPA* or *WPA*) for a passing play, we take advantage of the availability of air yards and develop two separate models for $\delta_{f,i,air}$ and $\delta_{f,i,yac}$. We are not crediting the QB solely for the value gained through the air, nor the receiver solely for the value gained from after the catch. Instead, we propose that both the QB and receiver, as well as the opposing defense, should have credit divided amongst them for both types of passing values. We let Δ_{air} and Δ_{yac} be the response variables for the air yards and yards after catch models, respectively. Both models consider all passing attempts, but the response variable depends on the model:

$$\Delta_{air} = \delta_{f,i,air} \cdot \mathbf{1}(\text{completion}) + \delta_{f,i} \cdot \mathbf{1}(\text{incompletion}),$$
$$\Delta_{yac} = \delta_{f,i,yac} \cdot \mathbf{1}(\text{completion}) + \delta_{f,i} \cdot \mathbf{1}(\text{incompletion}), \tag{14}$$

where $\mathbf{1}(\text{completion})$ and $\mathbf{1}(\text{incompletion})$ are indicator functions for whether or not the pass was completed. This serves to assign all completions the $\delta_{f,i,air}$ and $\delta_{f,i,yac}$ as the response for their respective models, while incomplete passes are assigned the observed $\delta_{f,i}$ for both models. In using this approach, we emphasize the importance of completions, crediting accurate passers for allowing their receiver to gain value after the catch.

The passing model for Δ_{air} is as follows:

$$\Delta_{air} \sim Normal(Q_{air,q[i]} + C_{air,c[i]} + F_{air,v[i]} + \mathbf{A}_i \cdot \boldsymbol{\alpha}, \ \sigma_{\Delta_{air}}) \text{ for } i = 1,\ldots, n \text{ plays},$$
$$Q_{air,q} \sim Normal(\mu_{Q_{air}}, \sigma^2_{Q_{air}}), \text{ for } q = 1,\ldots,\# \text{ of QBs},$$
$$C_{air,c} \sim Normal(\mu_{C_{air}}, \sigma^2_{C_{air}}), \text{ for } c = 1,\ldots,\# \text{ of receivers}, \tag{15}$$
$$F_{air,v} \sim Normal(\mu_{F_{air}}, \sigma^2_{F_{air}}), \text{ for } v = 1,\ldots,\# \text{ of defenses},$$

where the covariate vector $\mathbf{A}_i$ contains a set of indicator variables for Home, Shotgun, NoHuddle, QBHit, Location, RecPosition, as well as the RushStrength value while $\boldsymbol{\alpha}$ is the corresponding coefficient vector. The passing model for Δ_{yac} is of similar form:

$$\Delta_{yac} \sim Normal(Q_{yac,q[i]} + C_{yac,c[i]} + F_{yac,v[i]} + \boldsymbol{B}_i \cdot \boldsymbol{\beta},\ \sigma_{\Delta_{yac}}) \text{ for } i = 1, \ldots, n \text{ plays,}$$

$$Q_{yac,q} \sim Normal(\mu_{Q_{yac}},\ \sigma^2_{Q_{yac}}), \text{ for } q = 1, \ldots, \# \text{ of QBs,}$$

$$C_{yac,c} \sim Normal(\mu_{C_{yac}},\ \sigma^2_{C_{yac}}), \text{ for } c = 1, \ldots, \# \text{ of receivers,} \tag{16}$$

$$F_{yac,v} \sim Normal(\mu_{F_{yac}},\ \sigma^2_{F_{yac}}), \text{ for } v = 1, \ldots, \# \text{ of defenses,}$$

where the covariate vector $\boldsymbol{B}_i$ contains the same set of indicator variables in $\boldsymbol{A}_i$ but also includes the AirYards and interaction terms between AirYards and the various RecPosition indicators, with $\boldsymbol{\beta}$ as its respective coefficient vector. We include the RushStrength in the passing models as a group-level predictor to control for the possession team's rushing strength and the possible relationship between the two types of offense. For QBs, their estimated $Q_{air,q}$ and $Q_{yac,q}$ intercepts represent their iPA_{air} and iPA_{yac} values respectively (same logic applies to receivers). Likewise, the opposing defense values of $F_{air,v}$ and $F_{yac,v}$ are their tPA_{air} and tPA_{yac} values.

4.1.3 Rushing Models

For rushing plays, we again model the play values $\delta_{f,i}$. However, we build two separate models, with one rushing model for QBs and another for all non-QB rushes. This is because we cannot consistently separate (in the publicly available data) designed QB rushes from scrambles on broken plays, the characteristics of which result in substantially different distributions of play value added. It is safe to assume all non-QB rushes are designed rushes. Our rushing model for QBs consists of all scrambles, designed runs, and sacks (to account for skilled rushing QBs minimizing the loss on sacks). The QB rushing model is as follows:

$$\delta_{f,i} \sim Normal(Q_{rush,q[i]} + F_{rush_Q,\,v[i]} + \boldsymbol{\Gamma}_i \cdot \boldsymbol{\gamma},\ \sigma_{\delta_{rush_Q}}) \text{ for } i = 1, \ldots, n \text{ plays,}$$

$$Q_{rush,q} \sim Normal(\mu_{Q_{rush}},\ \sigma^2_{Q_{rush}}), \text{ for } q = 1, \ldots, \# \text{ of QBs,} \tag{17}$$

$$F_{rush_Q,v} \sim Normal(\mu_{F_{rush_Q}},\ \sigma^2_{F_{rush_Q}}), \text{ for } v = 1, \ldots, \# \text{ of defenses,}$$

where the covariate vector $\boldsymbol{\Gamma}_i$ contains a set of indicator variables for Home, Shotgun, NoHuddle, as well as the PassStrength variable where $\boldsymbol{\gamma}$ is the corresponding coefficient vector.

For the designed rushing plays of non-QBs, we include an additional group variable T. As detailed in Table 4 and Figure 8, T serves as a proxy for the offensive linemen or blockers involved in the rushing attempt. Each team has seven possible T levels of the form team-side-gap. For example, the Pittsburgh Steelers (PIT) have the following levels: PIT-left-end, PIT-left-tackle, PIT-left-guard, PIT-middle-center, PIT-right-guard, PIT-right-tackle, PIT-right-end. The non-QB rushing model is as follows:

$$\delta_{f,i} \sim Normal(H_{\iota[i]} + T_{\tau[i]} + F_{rush,v[i]} + \boldsymbol{P}_i \cdot \boldsymbol{\rho},\ \sigma_{\delta_{rush}}) \text{ for } i = 1, \ldots, n \text{ plays,}$$

$$H_\iota \sim Normal(\mu_H,\ \sigma^2_H), \text{ for } \iota = 1, \ldots, \# \text{ of rushers,}$$

$$T_\tau \sim Normal(\mu_T,\ \sigma^2_T), \text{ for } \tau = 1, \ldots, \# \text{ of team-side-gaps,} \tag{18}$$

$$F_{rush,v} \sim Normal(\mu_{F_{rush}},\ \sigma^2_{F_{rush}}), \text{ for } v = 1, \ldots, \# \text{ of defenses,}$$

where the covariate vector $\boldsymbol{P}_i$ contains a set of indicator variables for Home, Shotgun, NoHuddle, RushPosition, and PassStrength, and where $\boldsymbol{\rho}$ is the corresponding coefficient vector. The resulting $Q_{rush,q}$ and $H_{rush,\iota}$ estimates are the iPA_{rush} values for the QB and non-QB rushers, respectively. Additionally, the T_τ estimate is the $tPA_{rush,side-gap}$ for one of the seven possible side-gaps for the possession team, while $F_{rush,v}$ and $F_{rush_Q,v}$ are the tPA_{rush} and tPA_{rush_Q} values for the opposing defense for non-QB and QB rushes.

4.1.4 Individual Points/Probability Added

Let κ refer to the number of attempts for a type of play. Using an estimated type of iPA value for a player p and multiplying by the player's associated number of attempts provides us with an *individual points/probability above average* ($iPAA_p$) value. There are three different types of $iPAA_p$ values for each position:

$$
\begin{aligned}
iPAA_{p,air} &= \kappa_{p,pass} \cdot iPA_{p,air}, \\
iPAA_{p,yac} &= \kappa_{p,pass} \cdot iPA_{p,yac}, \\
iPAA_{p,rush} &= \kappa_{p,rush} \cdot iPA_{p,air},
\end{aligned}
\tag{19}
$$

where the values for $\kappa_{p,pass}$ and $\kappa_{p,rush}$ depend on the player's position. For QBs, $\kappa_{p,pass}$ equals their number of pass attempts, while $\kappa_{p,rush}$ is the sum of their rush attempts, scrambles, and sacks. For non-QBs $\kappa_{p,pass}$ equals their number of targets and $\kappa_{p,rush}$ is their number of rush attempts. Summing all three components provides us with player p's total individual points/probability above average, $iPAA_p$.

4.2 Comparing to Replacement Level

As described in Section 1.2, it is desirable to calculate a player's value relative to a "replacement level" player's performance. There are many ways to define replacement level. For example, Thomas and Ventura (2015) define a concept called "poor man's replacement", where players with limited playing time are pooled, and a single effect is estimated in a linear model, which is considered replacement level. Others provide more abstract definitions of replacement level, as the skill level at which a player can be acquired freely or cheaply on the open market (Tango et al., 2007).

   We take a similar approach to the *openWAR* method, defining replacement level by using a roster-based approach (Baumer et al., 2015), and estimating the replacement level effects in a manner similar to that of Thomas and Ventura (2015). Baumer et al. (2015) argue that "replacement level" should represent a readily available player that can replace someone currently on a team's active roster. Due to differences in the number of active players across positions in football, we define replacement level separately for each position. Additionally, because of usage for the different positions in the NFL, we find separate replacement level players for receiving as compared to rushing. In doing so, we appropriately handle cases where certain players have different roles. For example, a RB that has a substantial number of targets but very few rushing attempts can be considered a replacement level rushing RB, but not a replacement level receiving RB.

Accounting for the 32 NFL teams and the typical construction of a roster (Lillibridge, 2013), we consider the following players to be "NFL level" for each the non-QB positions:

- rushing RBs = $32 \cdot 3 = 96$ RBs sorted by rushing attempts,
- rushing WR/TEs = $32 \cdot 1 = 32$ WR/TEs sorted by rushing attempts,
- receiving RBs = $32 \cdot 3 = 96$ RBs sorted by targets,
- receiving WRs = $32 \cdot 4 = 128$ WRs sorted by targets,
- receiving TEs = $32 \cdot 2 = 64$ TEs sorted by targets.

Using this definition, all players with fewer rushing attempts or targets than the NFL level considered players are deemed replacement-level. This approach is consistent with the one taken by Football Outsiders (Schatz, 2003). We combine the rushing replacement level for WRs and TEs because there are very few WRs and TEs with rushing attempts.
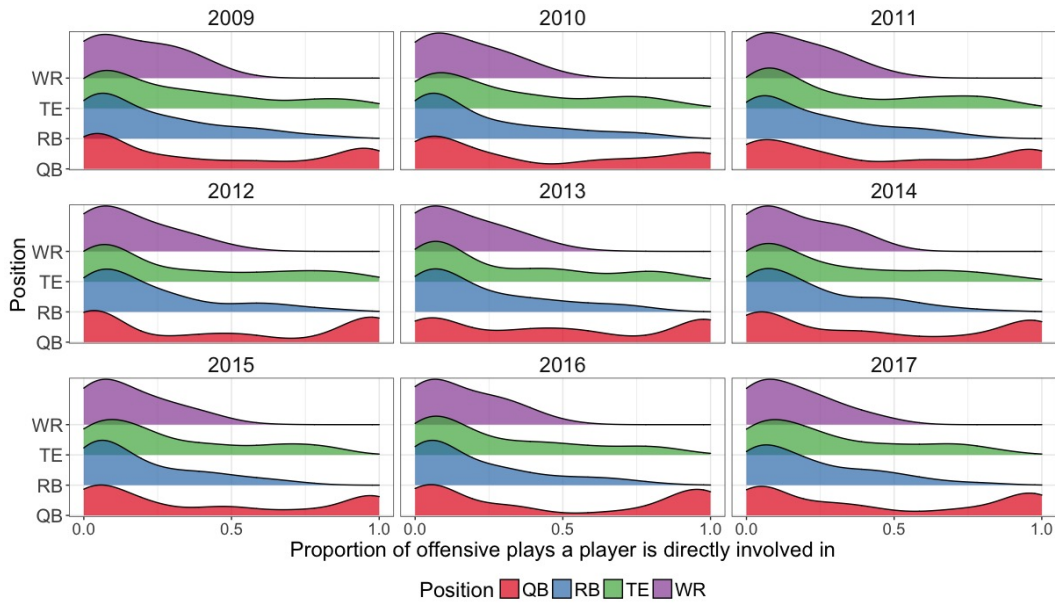
Figure 9: Distribution of the proportion of offensive plays a player is directly involved in by position (2009-2017).

In order to find replacement level QBs, we proceed in a different manner, due to the nature of QB usage in the NFL. Figure 9 displays the distribution of the percentage of a team's plays in which a player is directly involved (passer, receiver, or rusher) by position using data from 2009 to 2017. This does not represent the percentage of team snaps by a player, but rather for a given position that is directly involved in a play, it shows the distribution of team play percentages for every player of that position (e.g. New Orleans Saints' RB Alvin Kamara was involved in 38.39% of all Saints plays that directly involved a RB). While the distributions for RB, WR, and TE are unimodal and clearly skewed right, the distributions for QBs are bimodal for each season. This is an unsurprising result, since most NFL teams rely on a single QB for an entire season, resulting in them being involved in more than 80% of the team's plays at QB.

Observing this clear difference in the distribution for QBs, we consider two definitions of replacement level for QBs. The first is to define a replacement-level as any QB with less than ten

percent involvement in their team's plays that directly involve QBs. This approach essentially asserts that backup QBs with limited playing time should represent replacement level for QBs, rather than assuming all NFL teams have at least a certain number of NFL level QBs on their roster. The second option we consider is to limit NFL level to be the 32 QBs that attempted a pass in the first quarter of the first game of the season for each team, and label all remaining QBs as replacement level. The logic here is that NFL teams typically do not sign free agent QBs outside of their initial roster during the course of the season because it takes time to learn a team's playbook and offensive schemes. We recognize that these definitions are far from perfect, but we hope they provide a starting point for defining replacement level from which researchers can improve upon in the future.

Prior to fitting the models discussed in Section 4.1, every player who is identified as replacement level is replaced in their corresponding play-by-play data with their replacement label (e.g. Replacement QB, Replacement RB-rushing, Replacement RB-receiving, etc). By doing so, all replacement level players for a particular position and type (receiving versus rushing) have the same iPA^{repl} estimate. We then calculate a player's value above replacement, *individual points/probability above replacement* ($iPAR_p$) in the same manner as Baumer et al. (2015) and Thomas and Ventura (2015), by calculating a replacement level "shadow" for a particular player. For a player p, this is done by first calculating their replacement "shadow" value, $iPAA_p^{repl}$ by using their respective number of attempts:

$$
\begin{aligned}
iPAA_{p,air}^{repl} &= \kappa_{pass} \cdot iPA_{air}^{repl}, \\
iPAA_{p,yac}^{repl} &= \kappa_{pass} \cdot iPA_{yac}^{repl}, \\
iPAA_{p,rush}^{repl} &= \kappa_{rush} \cdot iPA_{rush}^{repl},
\end{aligned}
\tag{20}
$$

which leads to natural calculations for the three *iPAR* values:

$$
\begin{aligned}
iPAR_{p,air} &= iPAA_{p,air} - iPAA_{p,air}^{repl}, \\
iPAR_{p,yac} &= iPAA_{p,yac} - iPAA_{p,yac}^{repl}, \\
iPAR_{p,rush} &= iPAA_{p,rush} - iPAA_{p,rush}^{repl}.
\end{aligned}
\tag{21}
$$

Taking the sum of the three, we arrive at a player's total $iPAR_p$.

4.3   Conversion to Wins

If the play's value used for modeling purposes was *WPA* based, then the final *iPAR* values are an individual's win probability added above replacement, which is equivalent to their *wins above replacement* (*WAR*). However, for the *EPA*-based play value response, the *iPAR* values represent the individual expected points added above replacement, and thus require a conversion from points to wins. We use a linear regression approach, similar to that of Zhou and Ventura (2017) for football and Thomas and Ventura (2015) for hockey, to estimate the relationship between a team t's regular season win total and their score differential (S) during the season,

$$
Wins_t = \beta_0 + \beta_S S_t + \varepsilon_t, \text{ where } \varepsilon_t \sim N(0, \sigma^2) \ (iid).
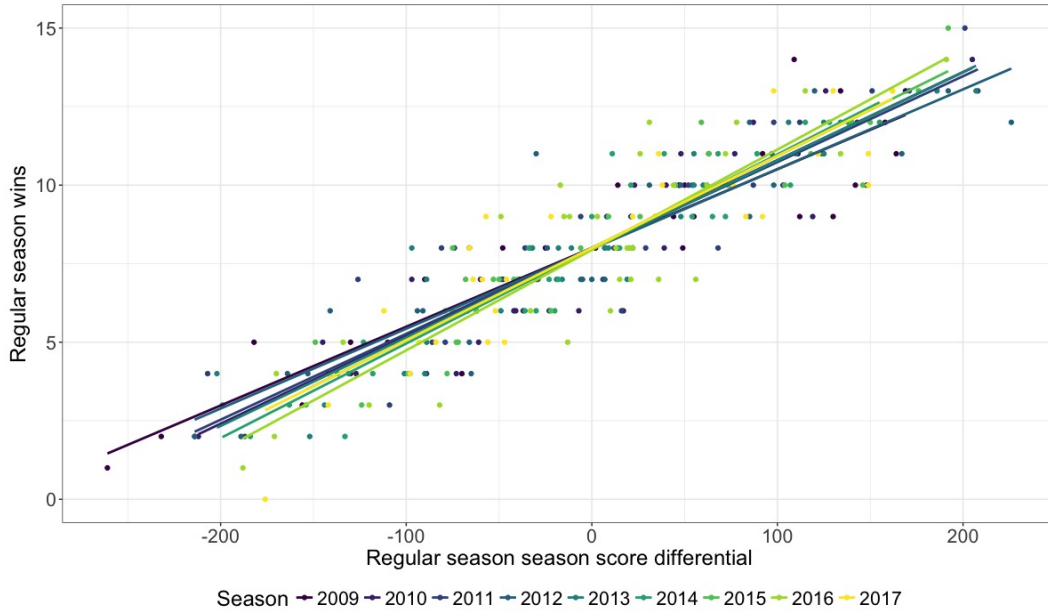\tag{22}
$$

Figure 10: Relationship between number of wins and score differential in the regular season by year (2009-2017).

Figure 10 displays the estimated linear regression fits for each season from 2009 to 2017. The resulting coefficient estimate $\hat{\beta}_S$ represents the increase in the number of wins for each one point increase in score differential. Thus we take the reciprocal, $\frac{1}{\hat{\beta}_S}$ to arrive at the number of points per win. We estimate *WAR* for the *EPA* based approach by taking the *iPAR* values and dividing by the estimated points per win (equivalent to multiplying *iPAR* by $\hat{\beta}_S$).

4.4 Uncertainty

Similar to the approach taken by Baumer et al. (2015) for estimating the variability in their *open-WAR* metric, we use a resampling strategy to generate distributions for each individual player's *WAR* values. Rather than resampling plays in which a particular player is involved to arrive at estimates for their performance variability, we resample entire team drives. We do this to account for the fact that player usage is dependent on team decision making, meaning that the random variation in individual events is dependent upon the random variation in team events. Thus, we must resample at the team level to account for the variability in a player's involvement. The decision to resample whole drives instead of plays is to represent sampling that is more realistic of game flows due to the possibility of dependencies within a drive with regards to team play-calling. We recognize this is a simple viewpoint of possible play correlations and consider exploration of this concept for future work. In Section 5, all uncertainty estimation uses this drive-resampling approach, with 1000 simulated seasons.

5 Results

Given the definitions in Section 4.2, we found the following replacement level designations for the 2017 NFL season for non-QB positions:

- rushing: 52 of the 148 RBs are replacement level,
- rushing: 278 of the 310 of the WR/TEs are replacement level,
- receiving: 52 of the 148 RBs are replacement level
- receiving: 73 of the 201 WRs are replacement level,
- receiving: 45 of the 109 TEs are replacement level.

For the QB position, we consider both approaches discussed in Section 4.2. With the "ten percent of QB plays cutoff" approach resulting in 25 replacement level QBs, and the "one QB for each team" approach resulting in 39 replacement level QBs out of the 71 in total.

First we compare the distributions of both types of WAR estimates, EPA-based and WPA-based, for the two considered definitions of replacement level QBs in Figure 11. It is clear that the "one QB for each team" approach for defining replacement level leads to lower WAR values in general, likely because some QBs who begin the season as back-ups perform better than those who begin the season as starters, yet are designated replacement level with this approach. For simplicity we only consider the ten percent cutoff rule for the rest of the paper.

We compare the distributions for both types of WAR estimates, EPA-based and WPA-based, by position in Figure 12. For all positions, the EPA-based WAR values tend be higher than the WPA-based values. This could be indicative of a player performing well in meaningless situations due to the score differential, particularly for QBs. It is clear that QBs have larger WAR values than the other positions, reflecting their involvement in every passing play and potentially providing value by rushing. Although this coincides with conventional wisdom regarding the importance of the QB position, we note that we have not controlled for all possible contributing factors, such as the specific offensive linemen, the team's offensive schemes, or the team's coaching ability due to data limitations. Researchers with access to this information could easily incorporate their proprietary data into this framework to reach a better assessment of QB value.

Following Major League Baseball's 2017 MVP race, WAR has received heavy criticism for its unclear relationship with wins (James, 2017; Tango, 2017). For this reason, we focus on the WPA-based version of WAR, with its direct relationship to winning games. Figure 13 displays the top five players based on total WAR for each position in the 2017 season. Each chart is arranged in descending order by each player's estimated WAR, and displays the three separate WAR values of WAR_{air}, WAR_{yac} and WAR_{rush}. By doing this separation, we can see how certain types of players vary in their performances. Tom Brady for instance is the only QB in the top five with negative WAR_{rush}. Alvin Kamara appears to be providing roughly equal value from rushing and receiving, while the other top RB performances are primarily driven by rushing success.

Elaborating on this separation of types of players, we can use the random intercepts from the multilevel models, the iPA values, to see the underlying structure of players in terms of their efficiency. Figures 14 and 15 reveals the separation of types of QBs and RBs respectively. The origin point for both charts represents league averages. For QBs, we plot their estimates for iPA_{air} against iPA_{yac}, providing an overview of the types of passers in the NFL. The two components represent different skills of being able to provide value by throwing deep passes through the
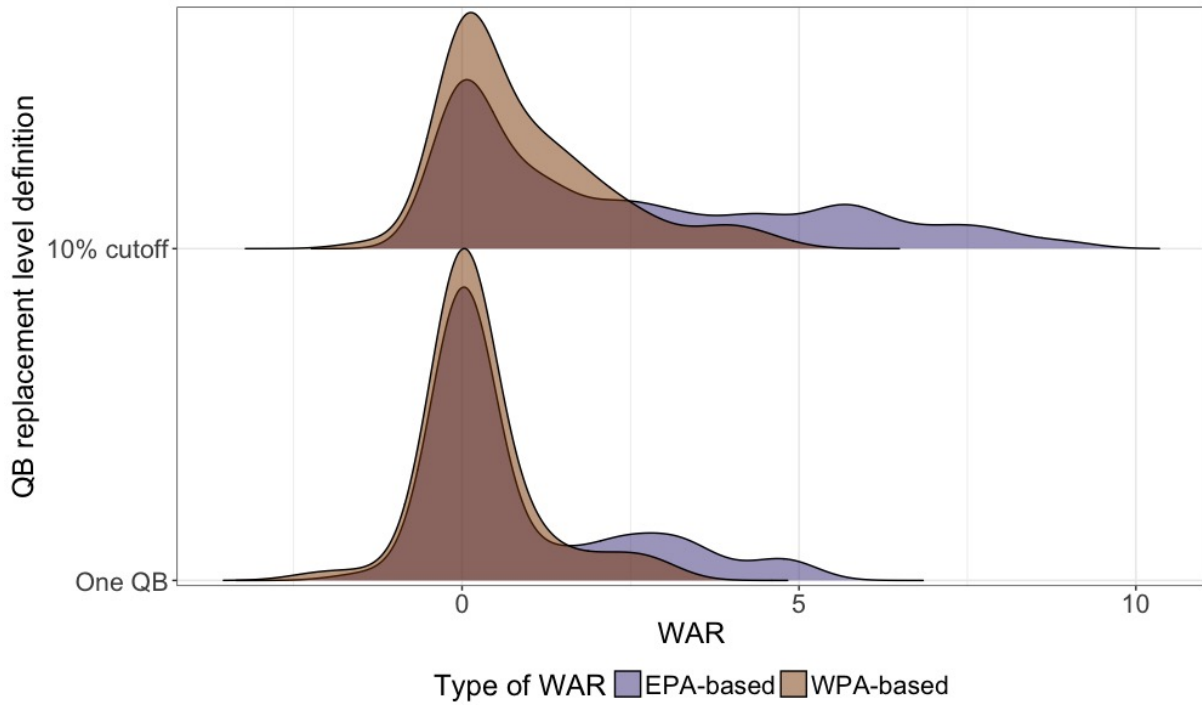
Figure 11: Distribution of QB *WAR* in 2017 season by type and replacement level definition.

air, such as Jameis Winston, as compared to short but accurate passers such as Case Keenum. We can also see where the replacement level QB estimates place for context. For RBs, we add together their iPA_{air} and iPA_{yac} estimates to summarize their individual receiving effect and plot this against their iPA_{rush} estimates. This provides a separation between RBs that provide value as receivers versus those who provide positive value primarily from rushing, such as Ezekiel Elliott. New Orleans Saints RB Alvin Kamara stands out from the rest of the league's RBs, providing elite value in both areas.

Using the drive resampling approach outlined in Section 4.4, we can compare the variability in player performance based on 1000 simulated seasons. Figure 16 compares the simulation distributions of the three types of *WAR* values (WAR_{air}, WAR_{yac}, WAR_{rush}) for selected QBs in the 2017 NFL season, with a reference line at 0 for replacement-level. We can clearly see that the variability associated with player performance is not constant, which is not suprising given the construction of the resampling at the drive level. However, we can see some interesting features of QB performances, such as how Seattle Seahawks QB Russell Wilson's three types of *WAR* distributions are overlapping significantly, emphasizing his versatility. Additionally, New England Patriots QB Tom Brady displays large positive WAR_{air} and WAR_{yac} values, but a clearly negative WAR_{rush} value. Finally, Joe Flacco's 2017 performance was at or below replacement level in the vast majority of simulations across all three types of *WAR*, indicating that he is not elite.

Figure 17 displays the simulation distributions for the top ten RBs during the 2017 NFL season, as ranked by their average total *WAR* across all simulations. Relative to the *WAR* values for QBs in Figure 17, the best RBs in the league are providing limited value to their teams.
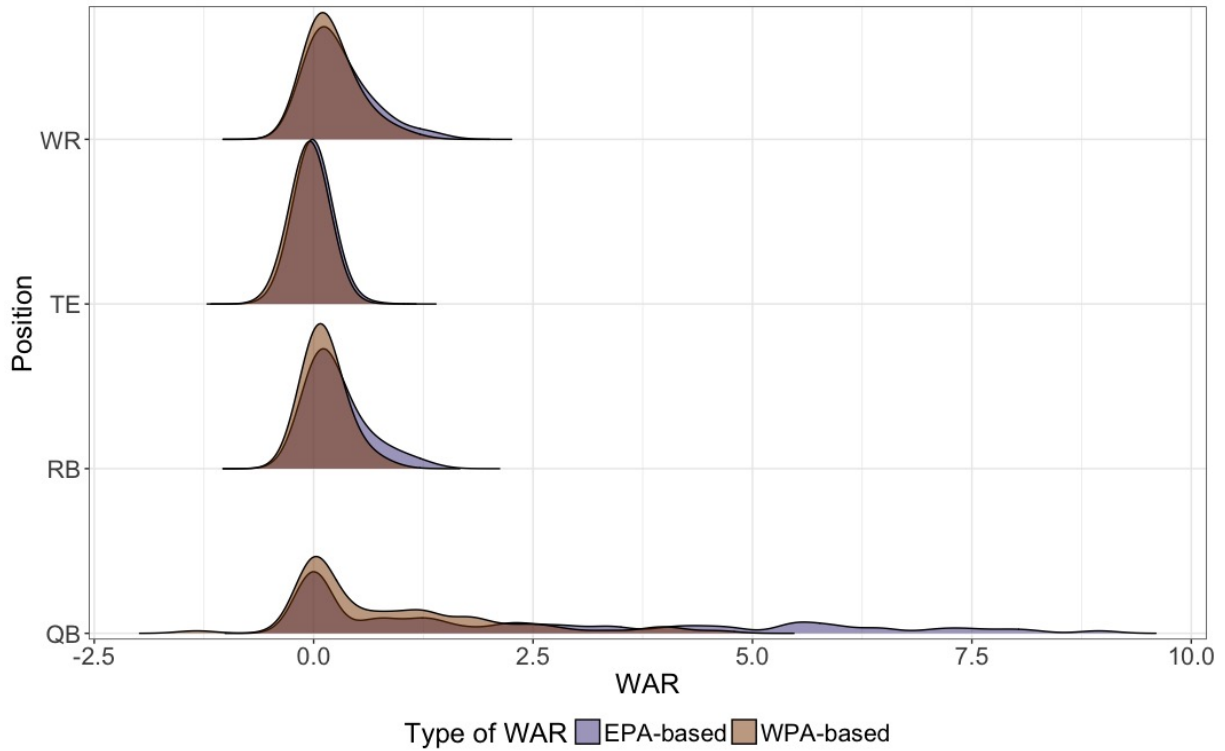
Figure 12: Distribution of *WAR* in 2017 season by type and position (ten percent cutoff used for replacement level QBs).

This is in agreement with the recent trend of NFL teams, who have been paying QBs increasing salaries but compensating RBs less (Morris, 2017). Two of the top RBs in the 2017 were rookies Alvin Kamara and Kareem Hunt, resulting into discussion of which player deserved to be the NFL's rookie of the year. Similar to Baumer et al. (2015) we address this question using our simulation approach and display the joint distribution of the two player's 2017 performances in Figure 18. In nearly 71% of the simulated seasons, Kamara leads Hunt in *WAR* providing us with reasonable certainty in Kamara providing more value to his team than Hunt in his rookie season. It should not come as a surprise that there is correlation between the player performances as each simulation consists of fitting the various multilevel models resulting in new estimates for the group averages, individual player intercepts as well as the replacement level performance.

Additionally, we examine the consistency of the *WPA*-based *WAR* from season-to-season based on the autocorrelation within players between their 2016 and 2017 seasons (excluding replacement level) and compare this to other commonly used statistics for QBs and RBs. Seen in Table 6, our estimates for QB *WAR* displayed higher correlations than both the commonly used Passer Rating statistic as well as Pro-Football-Reference.com's Adjusted Net Yards per Passing Attempt (ANY/A), which includes yards lost from sacks (Pro-Football-Reference, 2018). We also see higher correlations for RB *WAR* as compared to Brian Burke's Success Rate (percentage of rush attempts with *EPA* greater than zero) and rushing yards per attempt. Future work should consider a proper review and assessment of football statistics accounting for the number of attempts needed for determing the reliability of a statistic as well as accounting for when a player
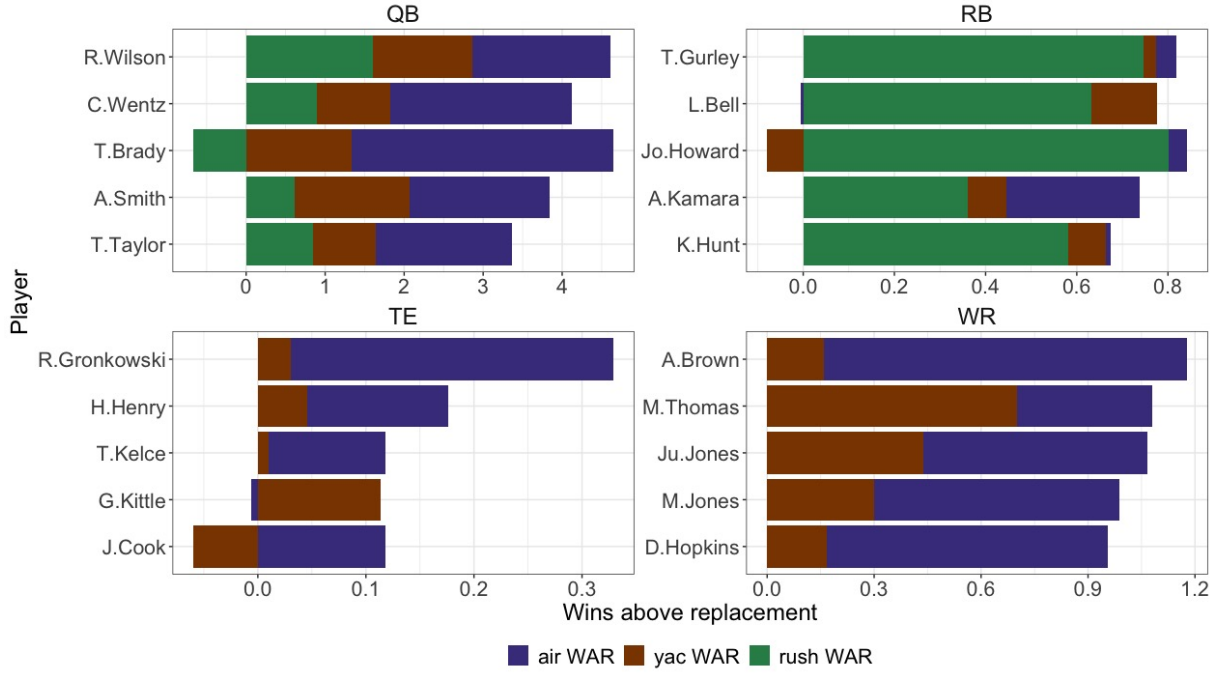
30

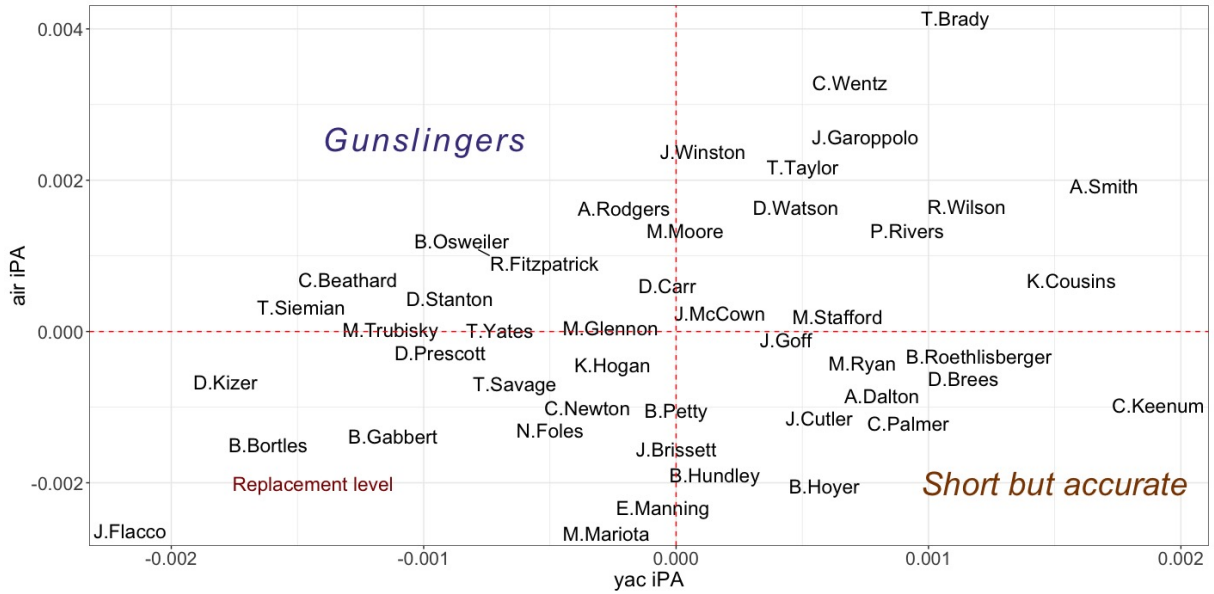Figure 13: Top five players in *WAR* by position for the 2017 season.

Figure 14: Estimates for QB efficiency from *iPA_air* against *iPA_yac* for the 2017 season.

changes teams (Yurko et al., 2017), and also apply the framework laid out by Franks et al. (2017).

Although it does not provide a measure for individual players' contributions, we can sum together the seven possible $tPA_{rush,side-gap}$ estimates for a team providing a proxy for their offensive line's overall efficiency in contributing to rushing plays. We can also look at individual side-gaps for specific teams to assess their offensive line's performance in particular areas. Fig-
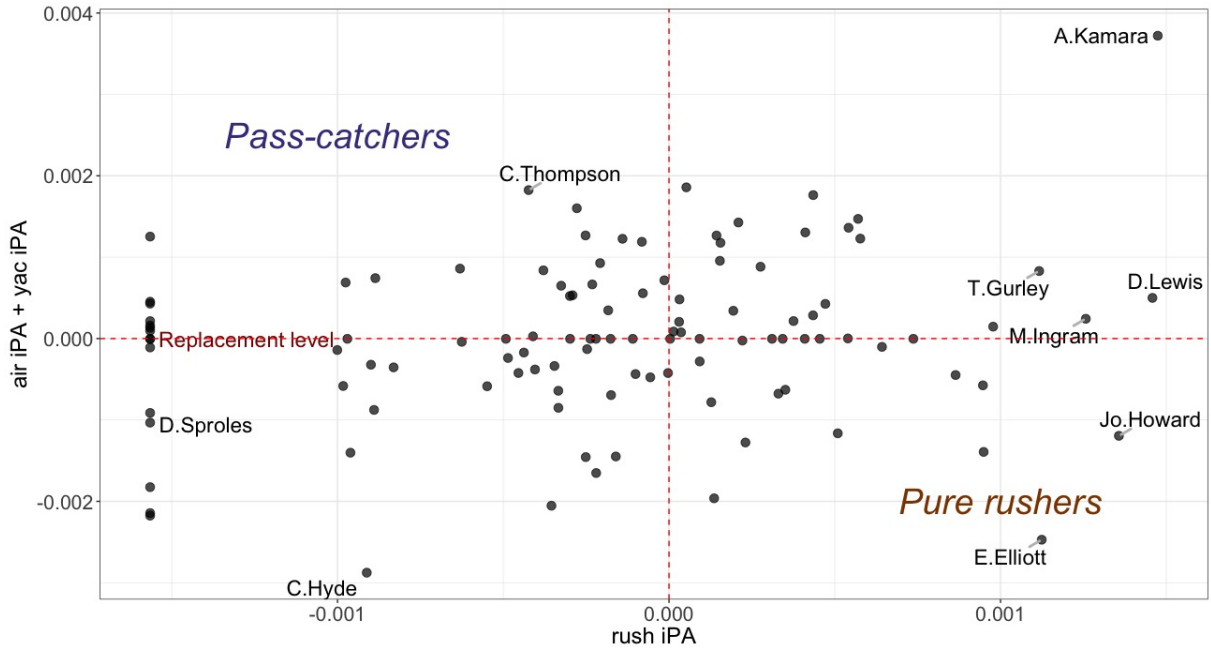
Figure 15: Estimates for RB efficiency from receiving ($iPA_{air} + iPA_{yac}$) against rushing (iPA_{rush}) for the 2017 season.
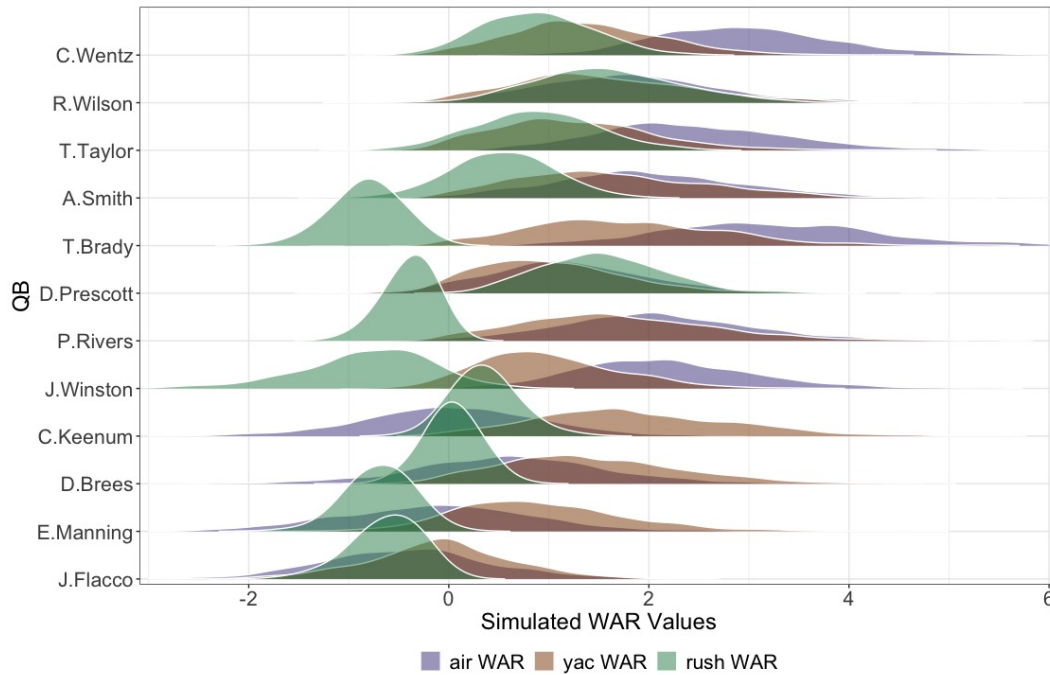
Figure 16: Simulation distributions of 2017 *WAR* by type for a selection of twelve QBs.

ure 19 displays the $tPA_{rush,side-gap}$ sum in 2017 against 2016 for each NFL team. The red lines provide indication to average performances in each year, so teams in the upper right quadrant performed above average overall in both years such as the Dallas Cowboys (DAL) which are
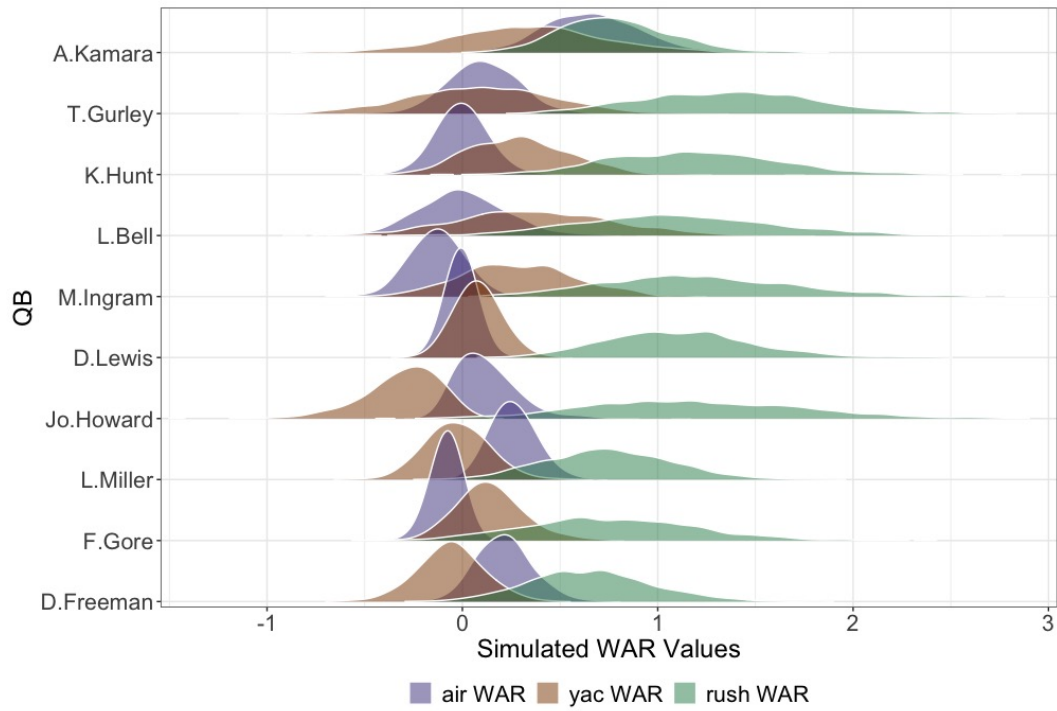
Figure 17: Simulation distributions of 2017 *WAR* value by type for top ten RBs.
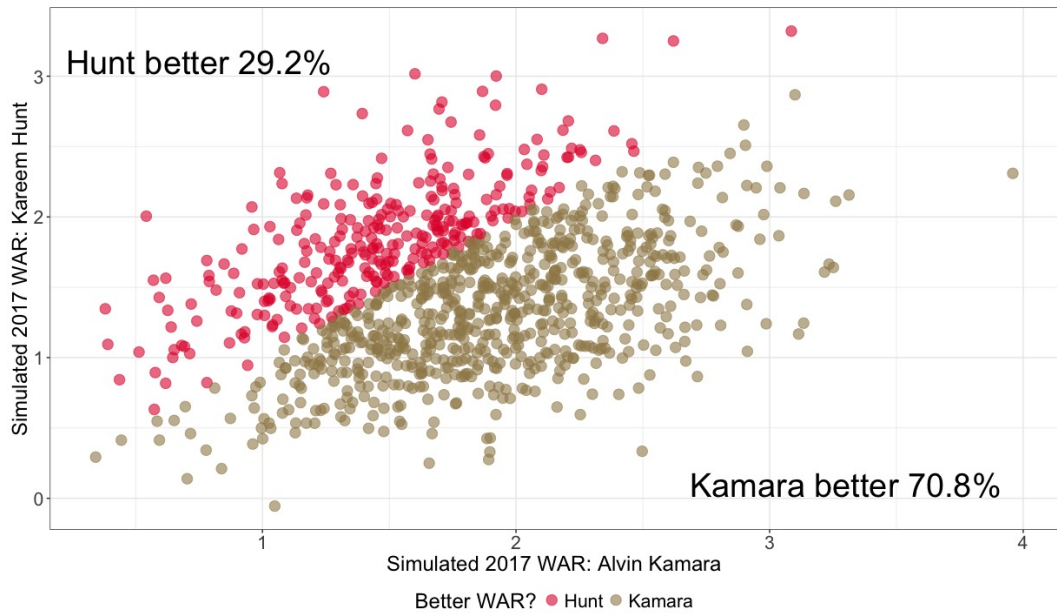
Figure 18: Joint distribution of *WAR* for Alvin Kamara vs. Kareem Hunt in 2017.

known to have one of the best offensive lines in football.

Table 6: Autocorrelation of QB statistics between 2016-17 seasons.

	WAR	Passer Rating	ANY/A
Autocorrelation	0.598	0.478	0.295

Table 7: Autocorrelation of RB statistics between 2016-17 seasons.

	WAR	Success Rate	Yards per Attempt
Autocorrelation	0.431	0.314	0.337
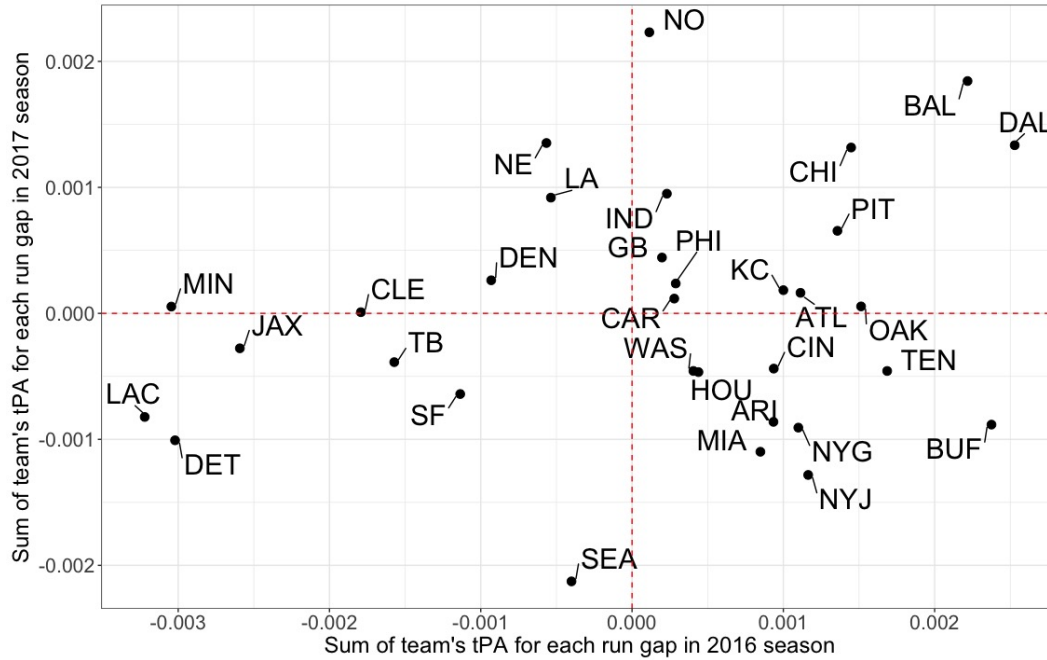
Figure 19: Team offensive line measures.

6 Discussion and Extensions

In this work, we have provided four major contributions to the statistical analysis of NFL football, in areas that can impact both on-field and player personnel decisions. These contributions are broken into three categories: software development and data, play evaluation, and player evaluation.

6.1 Data and Software Development

In the area of data access and software development, we provide an R package, nflscrapR, to provide easy access to publicly available NFL play-by-play data for researchers to use in their own analyses of the NFL. This package has already been used by researchers to further research

into NFL decision-making (Yam and Lopez, 2018).

6.2 Novel Statistical Methods for Play Evaluation

In the area of play evaluation, we make two contributions. First, we introduce a novel approach for estimating expected points using a multinomial logistic regression model. By using this classification approach, we more appropriately model the "next score" response variable, improving upon previous approaches. Additionally, our approach is fully reproducible, using only data provided publicly by the NFL. Second, we use a generalized additive model for estimating in-game win probability, incorporating the results of the expected points model as input. With these two play evaluation models, we can calculate measures such as expected points added and win probability added, which are commonly used to evaluate both plays and players.

With the notable exception of Lock and Nettleton (2014), researchers typically only vaguely discuss the methodology used for modeling expected points and/or win probability. Additionally, prior researchers in this area typically do not provide their specific expected points and win probability estimates publicly for other researchers to use and explore. Recently, Pro Football Focus used our approach for modeling expected points and found a clear relationship between their player grades and expected points added (Douglas and Eager, 2017). Importantly, in our work, all of these measures are included directly into the play-by-play data provided by `nflscrapR`, and our methodology is fully described in this paper. Moreover, all code used to build these expected points and win probability models is provided in `nflscrapR` and available on GitHub `https://github.com/ryurko/nflscrapR-models`. By taking these important steps, we ensure that all of our methods are fully reproducible, and we make it as easy as possible for researchers to use, explore, and improve upon our work.

6.3 Novel Statistical Methods for Player Evaluation

In the area of player evaluation, we introduce several metrics for evaluating players via our *nflWAR* framework. We use multilevel models to isolate offensive skill player contribution and estimate their individual wins above replacement. There are several key pieces of our *WAR* estimation that merit discussion.

First, estimates of *WAR* are given for several different areas of the game, including passing through the air, passing for yards after the catch, rushing, receiving through the air, and receiving yards after the catch. By compartmentalizing our estimates of player *WAR*, we are able to better characterize players and how they achieved success. For example, New Orleans Saints RB Alvin Kamara was unique in his success as both a rusher and a receiver in the 2017 NFL season, while other RBs like Los Angeles Rams RB Todd Gurley achieved most of their success as a rusher. Similarly, Seattle Seahawks QB Russell Wilson was unique in his success as a rusher as well as from passing through the air and for yards after the catch, with about equal *WAR* contributions in all three areas in the 2017 NFL season. This is in contrast to New England QB Tom Brady, who had tremendous success passing through the air and passing for yards after the catch, but provided negative *WAR* contributions as a rusher. We are also able to characterize players like Case Keenum, who in the 2017 NFL season performed very well as a passer for yards after the catch, but not as well as a passer through the air. While these findings may not surprise

knowledgeable football fans our framework also reveals the value of potentially overlooked skills such as the rushing ability of Tyrod Taylor and Dak Prescott, as seen in Figure 16. Their rushing value reflects not just their ability to scramble for positive value, but indicative of how they limit the damage done on sacks. Importantly, our player evaluation metrics are available for all skill position players, not just for QBs like previous approaches.

Second, our multilevel modeling approach allows us to estimate *WAR* contributions for NFL offensive lines and their specific sides and gaps on rushing plays, providing the first statistical estimate of offensive line ability that also controls for factors such as RB ability, opposing defense, etc. We recognize, however, that this is not a perfect measure of offensive line performance for a few reasons. First, this does not necessarily capture individual linemen, as blocking can consist of players in motion and the involvement of other positions. Second, there is likely some selection bias that is not accounted for in the play-by-play data that could influence specific side-gap estimates. For example, a RB may cut back and find a hole on the left side of the line on a designed run to the right because there is nothing open on the right side, resulting in a play being scored as a run to the left. Because of this selection bias at the RB level – RBs are more often going to run towards holes and away from defenders – our team-side-gap estimates may be biased, especially for teams with particularly strong or weak areas of their line. This is an issue with the play-by-play data that likely cannot be remedied publicly until player-tracking data is made available by the NFL. Finally, we lack information about which specific offensive linemen are on the field or even involved in plays, preventing us from fitting player-specific terms in our multilevel model that would provide *WAR* estimates for individual offensive linemen. Researchers with access to this data can build this into our modeling framework with minimal issues. However, until more data becomes available, researchers can incorporate these measures with more nuanced approaches of measuring offensive line performance such as Alamar and Weinstein-Gould (2008) and Alamar and Goldner (2011).

Third, by adopting a resampling procedure similar to that of Baumer et al. (2015), we provide estimates of uncertainty on all *WAR* estimates. Our approach resamples at the drive-level, rather than resampling individual plays, to preserve the effects of any within-drive factors, such as play sequencing or play-calling tendencies.

Finally, our *WAR* models are fully reproducible, with all data coming directly from `nflscrapR`, and with all code provided on GitHub `https://github.com/ryurko/nflWAR`. Because we use parametric models, it is trivially easy to incorporate more information, such as information about which players are on the field, or information from player-tracking data. We encourage future researchers to expand and improve upon our models in future work.

6.4   The Road to WAR for Players at All Positions

One key benefit to our approach is that it can easily be augmented with the inclusion of additional data sources, e.g. player-tracking data or proprietary data collected by NFL teams. One important way in which our models can be augmented comes via the inclusion of data about which players are present on the field for each play.

Given this information, we can update our multilevel models from Section 4.1 by including additional positional groups. For example, for the non-QB rushing model, we can update the

model as follows:

$$\delta_{f,i} \sim Normal(\sum_k O^k_{rush,\nu_k[i]} + \sum_g D^g_{rush,\nu_g[i]} + \boldsymbol{P}_i \cdot \boldsymbol{\rho}, \ \sigma_{\delta_{rush}}) \text{ for } i = 1, \ldots, n \text{ plays,}$$

$$O^k_{rush,\nu_k} \sim Normal(\mu_{O^k_{rush}}, \ \sigma^2_{O^k_{rush}}),$$

$$D^g_{rush,\nu_g} \sim Normal(\mu_{D^g_{rush}}, \ \sigma^2_{D^g_{rush}}),$$

where O^k_{rush,ν_k} are the intercepts for offensive positions (indexed by k and varying according to their own model), D^g_{rush,ν_g} are the intercepts for defensive positions (indexed by g and varying according to their own model), and $\boldsymbol{P}_i$ and $\boldsymbol{\rho}$ are described as above. Similar updates can be made to the models representing QB rushing, passing through the air, and passing for yards after catch. After doing so, we can trivially calculate the individual points/probability above average for any player at any position following the approach outlined in Section 4.1.4. From there, we simply need adequate definitions for replacement level players at each of these positions, and we will have statistical *WAR* estimates for players of any position, including all offensive players and all defensive players.

The data necessary for employing this approach *does exist*, but it is not available publicly, and there are heavy restrictions on the uses of such data. For example, Sportradar has a data product for the NFL called "Participation Data", which specifies all players present on the field for all plays, with data provided from the NFL. This is stated directly: "Participation Data is complementary data collected by the NFL that indicates all 22 players on the field for every play of every game" (Sportradar, 2018).

However, Sportradar's data sharing agreement explicitly prohibits the use of this data in the creation of new metrics, even if only used privately, as detailed in clauses 1.6 and 14.2 of the agreement (Sportradar, 2017). When colleagues reached out to Sportradar for clarification on potential data availability, they were told that there is no data sharing agreement for academic use, and that even if one were to purchase these data products, no new statistics or evaluation methods could be developed using this data, as per their terms and conditions. It is not clear if the same restrictions would apply to NFL teams.

6.5 Extensions Relevant to NFL Teams

`nflscrapR` provides play-by-play data, including expected points and win probability results, dating back to 2009, and improvements are underway to extend this back even further. As such, we can calculate player *WAR* dating back to at least 2009. If teams are able to implement the framework discussed in Section 6.4, they would then have *WAR* estimates for players at all positions dating back almost a full decade. Teams that are able to do this could potentially gain substantial advantages in important areas of roster construction.

First, teams could more appropriately assess the contract values of players in free agency, similar to what is commonly done in baseball (Paine, 2015).

Second and perhaps most importantly, teams would be able to substantially improve their analysis for the NFL draft. Using an approach similar to that of Citrone and Ventura (2017), teams could substitute an objective measure of *WAR* in place of the more subjective measure of "approximate value" (AV) (Pro-Football-Reference, 2018), in order to project the future career

value in terms of *WAR* for all players available in the NFL draft. Additionally, teams employing this approach could create updated, *WAR*-based versions of the "draft pick value chart", first attributed to Jimmy Johnson and later improved by Meers (2011) and Citrone and Ventura (2017). In doing so, teams could more accurately assess the value of draft picks and potentially exploit their counterparts in trades involving draft picks.

Acknowledgements

References

Alamar, B. (2010): "Measuring risk in nfl playcalling," *Journal of Quantitative Analysis in Sports*, 6.

Alamar, B. and K. Goldner (2011): "The blindside project: Measuring the impact of individual offensive linemen," *CHANCE*, 24, 25–29.

Alamar, B. and J. Weinstein-Gould (2008): "Isolating the effect of individual linemen on the passing game in the national football league," *Journal of Quantitative Analysis in Sports*, 4.

Albert, J. (2006): "Pitching statistics, talent and luck, and the best strikeout seasons of all-time," *Journal of Quantitative Analysis in Sports*, 2.

Balreira, E. C., B. K. Miceli, and T. Tegtmeyer (2014): "An oracle method to predict nfl games," *Journal of Quantitative Analysis in Sports*, 10.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015): "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, 67, 1–48.

Baumer, B. and P. Badian-Pessot (2017): "Evaluation of batters and base runners," in J. Albert, M. E. Glickman, T. B. Swartz, and R. H. Koning, eds., *Handbook of Statistical Methods and Analyses in Sports*, Boca Raton, Florida: CRC Press, 1–37.

Baumer, B., S. Jensen, and G. Matthews (2015): "openwar: An open source system for evaluating overall player performance in major league baseball," *Journal of Quantitative Analysis in Sports*, 11.

Baumer, B. and G. Matthews (2017): *teamcolors: Color Palettes for Pro Sports Teams*, URL `http://github.com/beanumber/teamcolors`, r package version 0.0.1.9001.

Becker, A. and X. A. Sun (2016): "An analytical approach for fantasy football draft and lineup management," *Journal of Quantitative Analysis in Sports*, 12.

Burke, B. (2009): "Expected point values," URL `http://archive.advancedfootballanalytics.com/2009/12/expected-point-values.html`.

Burke, B. (2014): "Expected points and expected points added explained," URL `http://www.advancedfootballanalytics.com/index.php/home/stats/stats-explained/expected-points-and-epa-explained`.

Carroll, B., P. Palmer, J. Thorn, and D. Pietrusza (1988): *The Hidden Game of Football*, New York, New York: Total Sports, Inc.

Carter, V. and R. Machol (1971): "Operations research on football," *Operations Research*, 19, 541–544.

Causey, T. (2013): "Building a win probability model part 1," URL `http://thespread.us/building-a-win-probability-model-part-1.html`.

Causey, T. (2015): "Expected points part 1: Building a model and estimating uncertainty," URL `http://thespread.us/expected-points.html`.

Citrone, N. and S. L. Ventura (2017): "A statistical analysis of the nfl draft: Valuing draft picks and predicting future player success," Presented at the Joint Statistical Meetings.

Clark, T. K., A. W. Johnson, and A. J. Stimpson (2013): "Going for three: Predicting the likelihood of field goal success with logistic regression," *MIT Sloan Sports Analytics Conference*.

Dasarathy, B. V. (1991): *Nearest neighbor (NN) norms: NN pattern classification techniques*, Los Alamitos, CA: IEEE Computer Society Press.

Deshpande, S. K. and S. T. Jensen (2016): "Estimating an nba player's impact on his team's chances of winning," *Journal of Quantitative Analysis in Sports*, 12.

Douglas, B. and E. A. Eager (2017): "Examining expected points and their interactions with pff grades," *Pro Football Focus Research and Development Journal*, 1.

Drinen, D. (2013): "Approximate value: Methodolgy," URL `https://www.sports-reference.com/blog/approximate-value-methodology/`.

Eager, E. A., G. Chahrouri, and S. Palazzolo (2017): "Using pff grades to cluster quarterback performance," *Pro Football Focus Research and Development Journal*, 1.

Elmore, R. and P. DeWitt (2017): *ballr: Access to Current and Historical Basketball Data*, URL `https://CRAN.R-project.org/package=ballr`, r package version 0.1.1.

ESPN (2017): "Scoring formats," URL `http://games.espn.com/ffl/resources/help/content?name=scoring-formats`.

Franks, A. M., A. D'Amour, D. Cervone, and L. Bornn (2017): "Meta-analytics: tools for understanding the statistical properties of sports metrics," *Journal of Quantitative Analysis in Sports*, 4.

Gelman, A. and J. Hill (2007): *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge, United Kingdom: Cambridge University Press.

Goldner, K. (2012): "A markov model of football: Using stochastic processes to model a football drive," *Journal of Quantitative Analysis in Sports*, 8.

Goldner, K. (2017): "Situational success: Evaluating decision-making in football," in J. Albert, M. E. Glickman, T. B. Swartz, and R. H. Koning, eds., *Handbook of Statistical Methods and Analyses in Sports*, Boca Raton, Florida: CRC Press, 183–198.

Gramacy, R. B., M. A. Taddy, and S. T. Jensen (2012): "Estimating player contribution in hockey with regularized logistic regression," *Journal of Quantitative Analysis in Sports*, 9.

Grimshaw, S. D. and S. J. Burwell (2014): "Choosing the most popular nfl games in a local tv market," *Journal of Quantitative Analysis in Sports*, 10.

Horowitz, M., R. Yurko, and S. L. Ventura (2017): *nflscrapR: Compiling the NFL play-by-play API for easy use in R*, URL `https://github.com/maksimhorowitz/nflscrapR`, r package version 1.4.0.

James, B. (2017): "Judge and altuve," URL `https://www.billjamesonline.com/judge_and_altuve/`.

Jensen, J. A. and B. A. Turner (2014): "What if statisticians ran college football? a reconceptualization of the football bowl subdivision," *Journal of Quantitative Analysis in Sports*, 10.

Jensen, S., K. E. Shirley, and A. Wyner (2009): "Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball," *The Annals of Applied Statistics*, 3.

Judge, J., D. Brooks, and H. Pavlidis (2015a): "Moving beyond wowy: A mixed approach to measuring catcher framing," URL `https://www.baseballprospectus.com/news/article/25514/moving-beyond-wowy-a-mixed-approach-to-measuring-catcher-framing/`.

Judge, J., D. Turkenkopf, and H. Pavlidis (2015b): "Prospectus feature: Introducing deserved run average (dra) and all its friends," URL `https://www.baseballprospectus.com/news/article/26195/prospectus-feature-introducing-deserved-run-average-draand-all-its-friends/`.

Katz, S. and B. Burke (2017): "How is total qbr calculated? we explain our quarterback rating," URL `http://www.espn.com/blog/statsinfo/post/_/id/123701/how-is-total-qbr-calculated-we-explain-our-quarterback-rating`.

Kubatko, J., D. Oliver, K. Pelton, and D. T. Rosenbaum (2007): "A starting point for analyzing basketball statistics," *Journal of Quantitative Analysis in Sports*, 3.

Lahman, S. (1996 – 2017): *Lahman's Baseball Database*, URL `http://www.seanlahman.com/baseball-archive/statistics/`.

Lillibridge, M. (2013): "The anatomy of a 53-man roster in the nfl," URL `http://bleacherreport.com/articles/1640782-the-anatomy-of-a-53-man-roster-in-the-nfl`.

Lock, D. and D. Nettleton (2014): "Using random forests to estimate win probability before each play of an nfl game," *Journal of Quantitative Analysis in Sports*, 10.

Lopez, M. (2017): "All win probability models are wrong some are useful," URL `https://statsbylopez.com/2017/03/08/all-win-probability-models-are-wrong-some-are-useful/`.

Macdonald, B. (2011): "A regression-based adjusted plus-minus statistic for nhl players," *Journal of Quantitative Analysis in Sports*, 7.

Martin, R., L. Timmons, and M. Powell (2017): "A markov chain analysis of nfl overtime rules," *Journal of Sports Analytics*, Pre-print.

Meers, K. (2011): "How to value nfl draft picks," URL `https://harvardsportsanalysis.wordpress.com/2011/11/30/how-to-value-nfl-draft-picks/`.

Morris, B. (2015): "Kickers are forever," URL `https://fivethirtyeight.com/features/kickers-are-forever/`.

Morris, B. (2017): "Running backs are finally getting paid what theyre worth," URL `https://fivethirtyeight.com/features/running-backs-are-finally-getting-paid-what-theyre-worth/`.

Mulholland, J. and S. T. Jensen (2014): "Predicting the draft and career success of tight ends in the national football league," *Journal of Quantitative Analysis in Sports*, 10.

Oliver, D. (2011): "Guide to the total quarterback rating," URL `http://www.espn.com/nfl/story/_/id/6833215/explaining-statistics-total-quarterback-rating`.

Paine, N. (2015): "Bryce harper should have made $73 million more," URL `https://fivethirtyeight.com/features/bryce-harper-nl-mvp-mlb/`.

Pasteur, D. and K. Cunningham-Rhoads (2014): "An expectation-based metric for nfl field goal kickers," *Journal of Quantitative Analysis in Sports*, 10.

Pattani, A. (2012): "Expected points and epa explained," URL `http://www.espn.com/nfl/story/_/id/8379024/nfl-explaining-expected-points-metric`.

Phillips, C. (2018): *Scoring and Scouting: How We Know What We Know about Baseball*, Princeton, New Jersey: Forthcoming, Princeton University Press.

Piette, J. and S. Jensen (2012): "Estimating fielding ability in baseball players over time," *Journal of Quantitative Analysis in Sports*, 8.

Pro-Football-Reference (2018): "Football glossary and football statistics glossary," URL `https://www.pro-football-reference.com/about/glossary.htm`.

Quealy, K., T. Causey, and B. Burke (2017): "4th down bot: Live analysis of every n.f.l. 4th down," URL `http://nyt4thdownbot.com/`.

R Core Team (2017): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Ratcliffe, J. (2013): "The pff idp scoring system revisited," URL `https://www.profootballfocus.com/news/the-pff-idp-scoring-system-revisited`.

Romer, D. (2006): "Do firms maximize? evidence from professional football," *Journal of Political Economy*, 114.

Rosenbaum, D. T. (2004): "Measuring how nba players help their teams win," URL `http://www.82games.com/comm30.htm`.

Schatz, A. (2003): "Methods to our madness," URL `http://www.footballoutsiders.com/info/methods`.

Schoenfield, D. (2012): "What we talk about when we talk about war," URL `http://espn.go.com/blog/sweetspot/post/_/id/27050/what-we-talk-about-when-we-talk-about-war`.

Sievert, C. (2015): *pitchRx: Tools for Harnessing 'MLBAM' 'Gameday' Data and Visualizing 'pitchfx'*, URL `https://CRAN.R-project.org/package=pitchRx`, r package version 1.8.2.

Skiver, K. (2017): "Zebra technologies to bring live football tracking on nfl gamedays with rfid chips," URL `https://www.cbssports.com/nfl/news/zebra-technologies-to-bring-live-football-tracking-on-nfl-gamedays-with-rfid-chips/`.

Smith, D., S. Siwoff, and D. Weiss (1973): "Nfl's passer rating," URL `http://www.profootballhof.com/news/nfl-s-passer-rating`.

Snyder, K. and M. Lopez (2015): "Consistency, accuracy, and fairness: a study of discretionary penalties in the nfl," *Journal of Quantitative Analysis in Sports*, 11.

Sportradar (2017): *NFL Data Addendum*, URL `https://developer.sportradar.com/page/NFL_Addendum`, last updated 2017-08-11.

Sportradar (2018): *NFL Official API*, URL `https://developer.sportradar.com/files/indexFootball.html#player-participation`, version 2.

Tango, T. (2017): "War podcast," URL `http://tangotiger.com/index.php/site/comments/war-podcast`.

Tango, T., M. Lichtman, and A. Dolphin (2007): *The Book: Playing the Percentages in Baseball*, Washington, D.C: Potomac Book, Inc.

Thomas, A. and S. L. Ventura (2017): *nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R*, URL `https://CRAN.R-project.org/package=nhlscrapr`, r package version 1.8.1.

Thomas, A. C. and S. L. Ventura (2015): "The road to war," URL `http://blog.war-on-ice.com/index.html%3Fp=429.html`.

Thomas, A. C., S. L. Ventura, S. T. Jensen, and S. Ma (2013): "Competing process hazard function models for player ratings in ice hockey," *The Annals of Applied Statistics*, 7.

Wakefield, K. and A. Rivers (2012): "The effect of fan passion and official league sponsorship on brand metrics: A longitudinal study of official nfl sponsors and roo," *MIT Sloan Sports Analytics Conference*.

Yam, D. R. and M. J. Lopez (2018): "Quantifying the causal effects of conservative fourth down decision making in the national football league." URL `https://statsbylopez.files.wordpress.com/2018/01/quantifying-causal-effects.pdf`, under Review.

Yurko, R., S. Ventura, and M. Horowitz (2017): "Nfl player evaluation using expected points added with nflscrapr," Presented at the Great Lakes Sports Analytics Conference.

Zhou, E. and S. Ventura (2017): "Wins and point differential in the nfl," URL `https://www.cmusportsanalytics.com/wins-point-differential-nfl/`.