Beyond Expected Goals

Conference Paper · March 2018

CITATIONS
47

READS
12,003

1 author:

William Spearman
Liverpool Football Club
147 PUBLICATIONS **15,763** CITATIONS

SEE PROFILE

MIT SLOAN SPORTS ANALYTICS CONFERENCE
presented by ESPN

Beyond Expected Goals

William Spearman
Hudl
william.spearman@hudl.com

Abstract

Many models have been constructed to quantify the quality of shots in soccer. In this paper, we evaluate the quality of off-ball positioning, preceding shots, that could lead to goals. For example, consider a tall unmarked center forward positioned at the far post during a corner kick. Sometimes the cross comes in and the center forward heads it in effortlessly, other times the cross flies over his head. Another example is of a winger, played onside, while making a run in past the defensive line. Sometimes the through-ball arrives; other times the winger must break off their run because a teammate has failed to deliver a timely pass. In both circumstances, the attacking player has created an opportunity even if they never received the ball. In this paper, we construct a probabilistic physics-based model that uses spatiotemporal player tracking data to quantify such *off-ball scoring opportunities* (OBSO). This model can be used to highlight which, if any, players are likely to score at any point during the match and where on the pitch their scoring is likely to come from. We show how this model can be used in three key ways: 1) to identify and analyze important opportunities during a match 2) to assist opposition analysis by highlighting the regions of the pitch where specific players or teams are more likely to create off-ball scoring opportunities 3) to automate talent identification by finding the players across an entire league that are most proficient at creating off-ball scoring opportunities.

1. Introduction

How can we quantify the value of a player standing, unmarked, at the far post waiting for a cross to come in that never arrives? Alternatively, how can we quantify the value of a striker who has positioned himself in space near the penalty spot but gets nothing out of it because the winger fails to deliver the square pass? The work we present in this paper endeavors to answer these and other questions related to the effect of off-ball positioning on the probability of scoring.

Compared to many other sports, scoring in soccer is a rare phenomenon. Because of this, various statistics are used as a proxy for the performance of a team throughout the course of match. Simple counting statistics such as the number of shots taken, the number of crosses, or the number of passes into the box are used to judge team performance. In 2012 a new metric, expected goals [1], was developed by Sam Green to quantify the probability of a shot resulting in a score. Various non-shot expected goal extensions [2] [3] [4], have been proposed that use non-shot events to quantify the likelihood that a given sequence of play will result in a score.

With the proliferation of spatiotemporal tracking data, exciting new ways of measuring the probability of scoring have been developed by Patrick Lucey et. al. [5], Daniel Link et. al. [6], and others. The research in [5] quantifies the probability that a shot will result in a score using strategic features from the 10-seconds of play before the shot; the *dangerosity* metric introduced in [6]

represents an innovative heuristic for quantifying the scoring danger posed by a scenario with explicit modeling of important factors such defensive pressure and the number of attackers and defenders in key regions in front of the goal.

In this paper, we attempt to build a model that represents the *probability* that a player not currently in possession of the ball will score based only on the *instantaneous* game state. We term this the *off-ball scoring opportunity* (*OBSO*).[1] There are three aspects about our approach that we would like to highlight:

- **Off-Ball Chance Creation:** Players can be given credit for creating space in a scoring location even if the ball is never delivered to them.
- **Understandable Modeling:** Each component of the model has real soccer meaning and answers a specific soccer question such as: "where will the next pass go?" or "which regions of the pitch are controlled by the attacking team?". Many of these model components could form the basis for additional analysis with tracking data to answer questions about soccer phenomena apart from scoring.
- **Scoring Predictiveness:** As with a traditional expected goals model, the *OBSO* can be integrated to yield the total expect score. This allows us to compute the expected scoring production for a player or team or during the course of a match. In addition, due to the spatial aspect of the model, we can also predict *where* on the pitch and *when* during the match goals are most likely to have occurred.

At the core of our model is an extension to the *pitch control field* first developed in [7] that quantifies the *potential* regions of control at some time in the near future. The *potential pitch control field* and derived models provide a new framework for interrogating the tracking data to answer questions about the short-term evolution of the game-state.

2. Data

For this analysis we use soccer match data produced by Hudl. The data spans 58 matches played between teams from a 14-team professional soccer league during the 2017-2018 season. Our dataset includes both event data and spatiotemporal tracking data.

The event data comprises the on-ball actions that were performed during the match. For each event, the following information is known: 1) the match time at which the event occurred 2) the on-ball player 3) the type of event (e.g. pass, shot, goal, etc.). The tracking data represents the location of every player on the pitch with a temporal frequency of 25 Hz and the corresponding match time for each tracking frame is specified.

2.1. Data Preprocessing

Our goal is to produce an analysis that is relatively insensitive to data source, data quality, and requires minimal data processing and cleaning. Accordingly, we perform only two steps when processing the data: 1) tracking data smoothing 2) naïve data synchronization.

[1] For this paper, we will often refer to *off-ball scoring opportunity* as *opportunity*. If the word *opportunity* is used in a more general context, it will not be italicized.

To smooth the tracking data, we perform a least-squares smoothing [8] that reduces noise and computes the instantaneous values of the higher order derivatives: velocity and acceleration. As a convention, we rotate the data by 180° when needed to ensure that the home team is always attacking from left to right. When data from both home and away teams is displayed, the home team is shown in red while the away team is shown in blue.

The moment data is synchronized with the tracking data by matching the event timestamp with the match time of the corresponding tracking frame. This allows us to take snapshots of the match for each event that let us know who is interacting with the ball, how they are interacting with the ball, and the locations of all the other players on the pitch at that point during the match. The event data timestamps have some level of stochastic noise which leads to a temporal uncertainty on the order of three seconds. This can lead to moderate errors in the positions of the players. To minimize the effect of these errors, we choose a Bayesian approach to fitting our model (see Section 3.5 for a full description of our parameter estimation procedure) and select strong priors based on independently measurable/known information.

2.2. Data Anonymization

The data used for this analysis has been anonymized and the teams are referred to by a letter from A-N. Although only 58 games are used for the analysis, mid-season league standing of the teams are used to describe the team using a single letter. Thus, the team with the best mid-season record is Team A while the team with the worst mid-season record is Team N. In addition, player jersey numbers visible in certain visualizations have also been randomized and do not correspond to the true jersey number of the player.

2.3. Reproducibility

In an effort to maximize analysis reproducibility, modeling decisions are made to minimize the impact of data inaccuracies that may vary between providers of spatiotemporal tracking and event data. Ball data is not needed as on ball-events can be used to identify which player is currently touching the ball. The minimum event data required is: 1) when the event occurred 2) who touched the ball during the event and 3) whether the event was a goal, a shot, or another on-ball event. Although the tracking data used for this analysis was collected at 25 Hz, it is only necessary to have a snapshot of the positions and velocities of the players for each on-ball event. This reduces the required frames from ~8,000,000 (25 Hz for 90 minutes) to a more manageable ~1000 (one per event).

3. Methods

To compute the probability that an attacking team will score with the next on-ball event, we must compute the probability of a pass to each of the other 10 players and the probability that the receiving player will shoot and score. Of course, some of these probabilities are bound to be very small, e.g. the probability of the keeper scoring is almost always zero even though the back pass to the keeper is likely. Although conceptually simple, it is difficult to compute the above transition probabilities without spatial information. To simplify, we rephrase the problem as follows: what is the probability that the attacking team successfully passes to each point on the pitch and scores from that location. This can be conceptualized as a three step process.

- **Transition:** Probability that the next on-ball event occurs at an arbitrary point, r on the pitch. Denoted by: T_r.
- **Control:** Probability that a ball at point r will be controlled by the passing team. Denoted by C_r.
- **Score:** Probability of scoring from point r on the pitch assuming the next on-ball event occurs there.[2] Denoted by G_r.

The total probability of scoring with the next on-ball event for the attacking team can therefore be described by the summation in Equation 1:

$$P(G|D) = \sum_{r \in \mathbb{R} \times \mathbb{R}} P(G_r \cap C_r \cap T_r | D) \quad (1)$$

In this equation, D represents the instantaneous state of the game (player positions and velocities). The conjoined probability can be decomposed into a series of conditional probabilities as follows:

$$P(G|D) = \sum_{r \in \mathbb{R} \times \mathbb{R}} P(S_r|C_r, T_r, D) P(C_r|T_r, D) P(T_r|D) \quad (2)$$

In the next sections, we will address each term and the simplifications we use to compute them.

3.1. Control Model: The Potential Pitch Control Field

The most difficult expression in the decomposed conditional probabilities presented in Equation 2 is the second term: $P(C_r|T_r, D)$, the probability that the attacking team will control the ball at point r assuming the next on-ball event occurs there. We introduce a new model, the *potential pitch control field (PPCF)*. This concept borrows heavily from the *pitch control field* developed in [7]. Both models assume that while in proximity to the ball, a player's ability to make a controlled touch on the ball can be treated as a Poisson point process. The longer a player is near the ball without another player interfering, the more likely it becomes that they are able to make a controlled touch on the ball. As in [7], we seek to quantify the probability of control for each player at each location on the pitch. Instead of assuming the ball were to arrive at the destination instantaneously, we compute the time it would take the ball to arrive using aerodynamic drag. This means that more distant passes will give defenders and attackers more time to converge on the specified position. Furthermore, we introduce a decision making aspect to the model whereby the pass is assumed to travel at the most advantageous physically realistic time for the attackers. Before considering the full details of this model, an illustrative example of the *PPCF* is shown in Figure 1 below.

[2] Note: this does not assume that a shot is taken, so this term encompasses the probability that the on-ball event is a shot and the probability of that shot being successful.

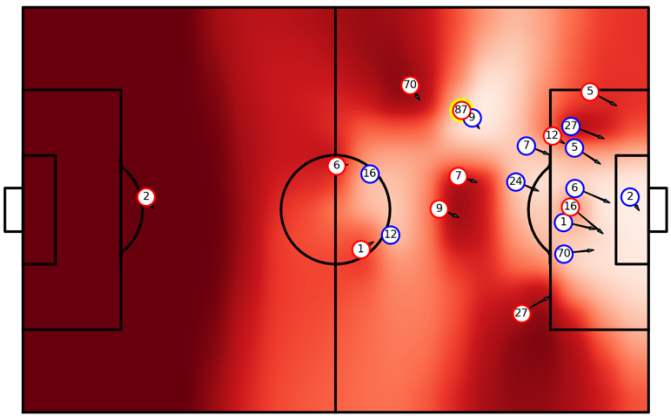


Figure 1. The example of the potential pitch control field (PPCF) is displayed above. Note that in reality, the PPCF is a three dimensional with the third dimension representing the player. In this example, it has been integrated over all players on the Red team. Red #87 controls the ball as highlighted in yellow. Dark red indicates strong control by the red attacking team (corresponding to a PPCF value of 1) while white indicates control by the blue team (corresponding to a PPCF value of 0). Players are represented by colored circles with the anonymized jersey number in the center. Player velocities are represented by vectors emanating from the corresponding player circle.

In Figure 1, we can identify a few important features of the *PPCF*. Notice how Red #9 and Red #7 are moving forward at speed. This opens up the space behind them which they no longer control due to the time it would take them to turn around. Red #5 is offside and thus is unable to legally influence play. Blue #27 is running towards the goal which opens up space left of the goal for Red #12 to control. The midfield region of the pitch is less dark indicating the potential for a counter attack by Blue #16 and Blue #12 if a turnover were to occur.

The differential equation used to compute the control probability for each player at a specified location, r, at time t is:

$$\frac{dPPCF_j}{dT}\left(t,\vec{r},T|s,\lambda_j\right) = \left(1-\sum_k PPCF_k\left(t,\vec{r},T|s,\lambda_j\right)\right) f_j(t,\vec{r},T|s)\lambda_j \qquad (3)$$

Where $f_j(t,\vec{r},T|s)$ represents the probability that player j at time t can reach location r within some time T. This term is addressed in more detail in Section 3.1.1. We incorporate time of flight of the ball by setting $PPCF_i(t,\vec{r},T|s,\lambda_i) = 0$ when T is less than the time of flight of the ball at location r. These flight times are discussed in more detail in Section 3.1.2. The parameter λ_i is the rate of control and represents the inverse of the mean time it would take a player to make a controlled touch on the ball. This parameter is discussed further in Section 3.1.3.

Integrating Equation 2 over T from 0 to ∞, we build a per-player probability of control. The resultant *PPCF* is three dimensions with two spatial dimensions on the pitch and a third dimension for the players. Generally, when visualizing the *PPCF*, we integrate over the players of the attacking team.

3.1.1. Time to Intercept

The probability that a player will be able to intercept the ball at a given location on the pitch within some time, T, is given by the term $f_j(t, \vec{r}, T|s)$ in Equation 3. To model this, we first compute the *expected* intercept time, $\tau_{exp}(t, \vec{r})$. This is done by finding the time it would take player j to reach location $\vec{r}$ from their start location $\vec{r}_j(t)$ with a starting velocity $\vec{v}_j(t)$ assuming they are able to accelerate with some constant acceleration a to a maximum speed v. The values of a and v should be chosen to correspond with the average sustainable speed and acceleration during a match. For this analysis, the values of 5 m/s and 7 m/s^2 are chosen respectively.[3] Numerous factors can lead to the *true* intercept time, $\tau_{true}(t, \vec{r})$, showing significant variance from this computed *expected* value, these include: tracking data inaccuracies, player facing, player awareness, tactical decision making, and other factors. To avoid the explicit modeling of these and other issues, we model the distribution of residuals, $\tau_{exp}(t, \vec{r}) - \tau_{true}(t, \vec{r})$, using the logistic function (we choose the logistic over the normal because of its heavier tails). This allows us to compute the probability that the player will be able intercept the ball using the cumulative distribution function of the logistic:

$$f_j(t, \vec{r}, T|s) = \left[1 + e^{-\pi \frac{T - \tau_{exp}(t,\vec{r})}{\sqrt{3}s}}\right]^{-1} \quad (4)$$

3.1.2. Time of Flight

To compute the time of flight of the ball, we simulate different trajectories using aerodynamic drag with a coefficient of drag that exhibits non-linear dependence on the speed of the ball as described in [9]. By varying the angle and speed of the ball, we can understand the minimum and maximum time it would take the ball to travel a certain distance.

When building the *PPCF*, we select the time of flight that most closely matches the arrival time of the nearest attacking player. This modeling choice advantages successful passing and helps to counteract the limitations of the decision model in the next section.

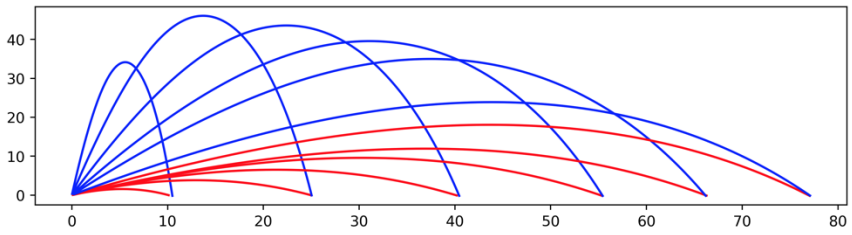


Figure 2. This figure illustrates possible ball trajectories (all units are in meters). Trajectories with a large flight time are shown in blue while trajectories with a smaller flight time are shown in red.

Although we do not explicitly model blocked passes, the control exerted by defenders through the *PPCF* will serve as a proxy for balls blocked near the receiver. Because long balls are likely to be lobbed, and therefore unable to be intercepted through the majority of their trajectory, this

[3] There is no reason that a and v cannot be set on a per-player basis allowing for more intelligent modeling of break-away situations with a fast attacking player, but for simplicity, we use fixed values of a and v.

simplification is largely justified. It bears mention that distant lobbed passes are more difficult to execute precisely than short ones. Although this is not modeled explicitly in the *PPCF*, it is treated by proxy in Section 3.2 where we utilize a normal distribution to disadvantage long-passing.

3.1.3. Control Rate

The parameter, λ_i is the control rate and has units of $1/\text{seconds}$. Higher values of λ_i indicate less time is required before the player is able to control the ball. In the context of computing opportunities, we would expect that this control rate might be different for attacking players and defending players. Attacking players want to make a precise controlled touch that results in a shot or continued possession while a defender will often be satisfied with heading the ball away or kicking it out of play. To account for this, we introduce a second parameter, κ, which scales the attacking control rate, λ. Thus, we can write the following expression for λ_i where A is the set of attacking players and B is the set of defending players.

$$\lambda_i = \begin{cases} \lambda & i \in A \\ \kappa\lambda & i \in B \end{cases} \tag{5}$$

Note that the parameter, λ_i is set to zero for attacking players when that attacking player is in an offside position.

3.2. Transition Model

The final term in in Equation 2 quantifies the likelihood that the next on-ball moment will occur at an arbitrary point $\vec{r}$. Figure 3 shows a histogram of the average displacement, $\Delta\vec{r} = \vec{r}_{i+1} - \vec{r}_i$, between subsequent on-ball events in data.

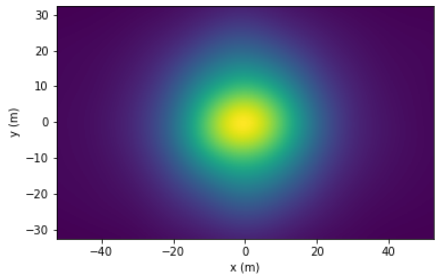


Figure 3. A 2D histogram of the relative location of subsequent ball touches, $\vec{r}_{i+1} - \vec{r}_i$. The standard deviation of the average displacement is 14 m.

Given that the ball is moved around the pitch through collisions with players (e.g. passes, headers, blocks, interceptions, etc.) it makes sense that the ball's motion will appear to be a form of two-dimensional Brownian motion and that the distribution of displacements between subsequent events will be normally distributed in aggregate.[4]

[4] Another way to conceptualize this is that players are more likely to attempt easy short passes and if they attempt a long pass, the resultant location will have higher variance due to the implicit angular variance present when passing.

Although, on average, the distribution of displacements for subsequent on-ball events may be normally distributed, we would expect there to be intelligent decision making on the part of passers. Passers are likely to select passes that are less likely to be intercepted. As we have already constructed a model that describes the probability that a pass to a given spatial location will be successful, we can superimpose these models and construct a decision probability density field using the following equation:

$$\mathrm{T}(t, \vec{r}|\sigma, \alpha) = \mathrm{N}(\vec{r}, \vec{r}_b(t), \sigma) \cdot \left[\sum_{k \in A} PPCF_k(t, \vec{r})\right]^{\alpha} \quad (6)$$

In this expression, A represents the set of all players from the team in possession, α is a model parameter used to scale the dependence of the decision conditional probability by the *PPCF* and N is a two-dimensional normal distribution.[5] The expression in Equation 6 is normalized to unity.

3.3. Score Model

The first conditional probability in Equation 2, $P(S_r|C_r, T_r, D)$, represents the likelihood of scoring from a location r assuming that the ball is successfully controlled at that location by the attacking team. To simplify our model, we ignore game state, D, for this term and base the model solely on distance to the goal. Our assumption is that defensive positioning will be proxied through use of the *PPCF* in the other conditional probability models.

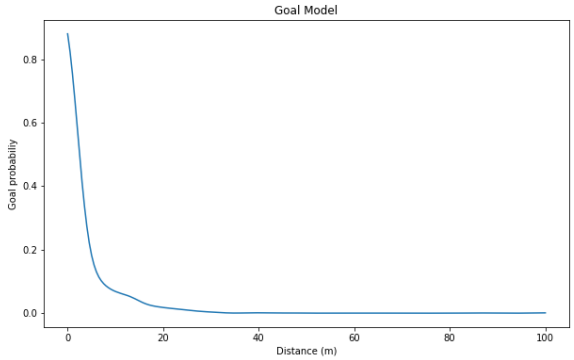


Figure 4. The probability of scoring plotted against the distance from goal. Gaussian kernel smoothing is used to ensure a continuous distribution.

Event data is used to compute the probability of scoring a goal conditioned on distance to the goal. The functional dependence seen in Figure 4 represents the average scoring chance given an on-ball event at a particular distance from goal. It is likely that this will not represent the true distribution due to an inherent shot selection bias. For example, we would expect that if unpressured, a player is more likely to shoot from and score at *20 m* than the average player making a ball-touch *20 m* from

[5] One term conspicuously absent from the next on-ball event probability density is a preference for events that move the ball closer to the target goal. In future variants of this analysis, we intend to incorporate such a term. For now, however, the failure of the model to incorporate this information means that opportunities that may appear as clear scoring chance to a domain expert may be under-estimated by the model.

the goal. To allow the data to account for this, we add a model parameter, β, that permits the shape of the above distribution to vary while maintaining the monotonic decreasing behavior we expect:

$$S(\vec{r}|\beta) = \left[S_d(|\vec{r} - \vec{r}_g|)\right]^{\beta} \quad (7)$$

In Equation 7, $\vec{r}_g$ is the location of the target goal and $S_d(x)$ represents the data-derived function displayed in Figure 4.

3.4. Combination

Using Equation 2, we combine the conditional probabilities to give a single, unified model which represents the posterior probability of scoring with the next on-ball event at a particular location. The constituent conditional probability and probability density maps are represented spatially in Figure 5. Note that Figure 5a and Figure 5b are probability maps while Figure 5c and Figure 5d are spatial probability densities which must be spatially integrated to be interpretable as a probability.

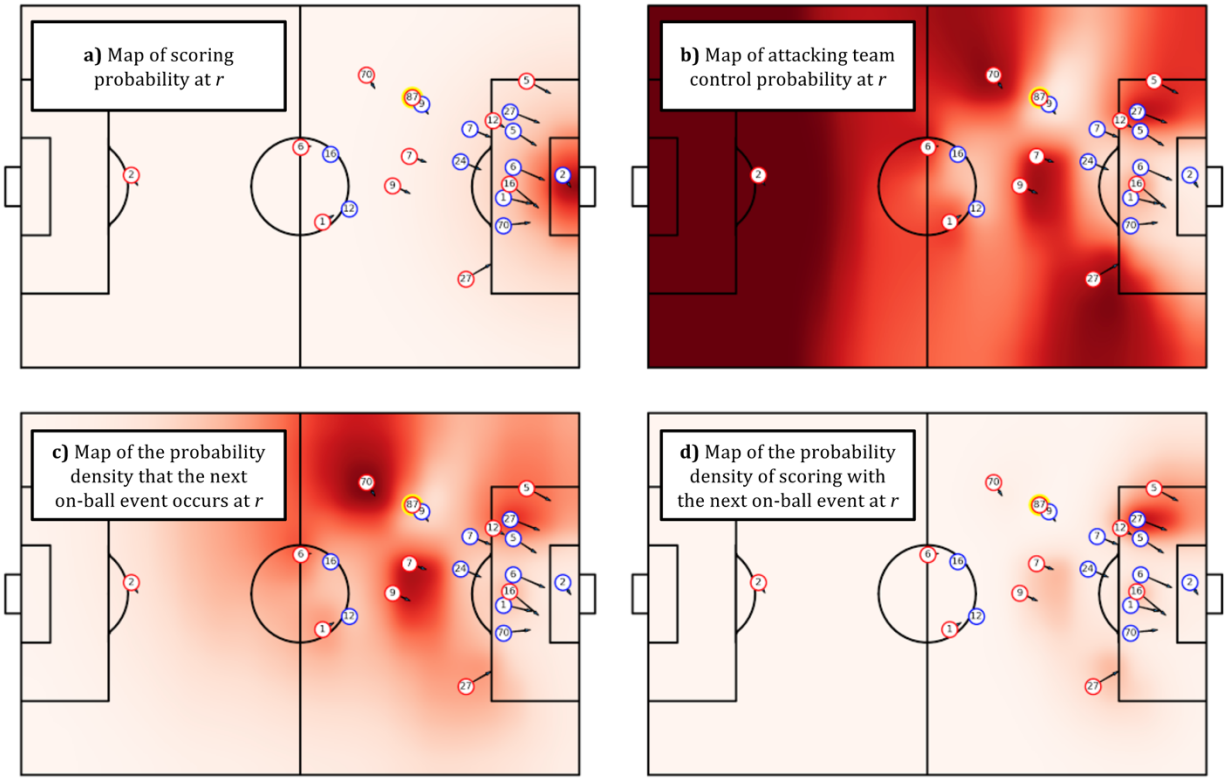


Figure 5. a) The scoring model (no spatial normalization, dark red corresponds to unity). b) The control model: probability of the attacking team controlling the ball at a given location assuming the next on-ball event occurs there (no spatial normalization, dark red corresponds to unity). c) Transition model: probability density of the location of the next on ball-event (normalized to unity) d) Off-ball scoring opportunity model: probability density of scoring with the next on-ball event at the specified location. Red #87 has the ball. Red #5 is offside and is not included in the computation. The integrated magnitude of the OBSO is 1.1%.

We use the same example for which the *PPCF* was presented in Figure 1. In Figure 5c, we see a probability density map that can be integrated to determine which receivers are most likely. In this case, there is an obvious easy back pass to Red #70 and this player represents the most likely target with a less likely outlet pass to Red #7 or Red #9. It is also possible that a forward through ball will be attempted to Red #12 or a long cross to Red #27.

According to the *opportunity* model presented in Figure 5d, the most dangerous region is to the left of the goal where Red #12 is available to receive a pass in the space that Blue #27 is currently occupying but cannot control because of their speed towards the end line. A shot from distance is available if the ball is passed to Red #9 or Red #7 and the cross to Red #27 on the right side could also result in a score if properly executed. Overall, this is not a scenario that represents a high probability of scoring. The total *opportunity* of scoring with the next ball touch integrates to 1.1%.

3.5. Parameter Estimation

The probability given in Equation 2 allows us to write the total likelihood for each on-ball event in the training set as follows:

$$\mathcal{L}(T|\boldsymbol{\theta}) = \prod_{D \in T} \begin{cases} P(G|D,\boldsymbol{\theta}) & k = 1 \\ 1 - P(G|D,\boldsymbol{\theta}) & k = 0 \end{cases} \tag{8}$$

Where T represents the training set of events, k indicates whether the subsequent on-ball event is a goal $k = 1$, or not $k = 0$, and $\boldsymbol{\theta}$ represents the vector of model parameters. For fitting, a set of five games training games is reserved from the total of 58 and excluded from the rest of the analysis. Due to the small size of the training data and issues with event data synchronization discussed in Section 2.1, we use a Bayesian approach with normally distributed priors to estimate the maximum a posteriori probability (MAP) for each parameter. Detailed information about the model parameters comprising $\boldsymbol{\theta}$ and their MAPs is found in Table 1.

Parameter	MAP	Units	Description	Prior Parameters	Prior Selection
$\boldsymbol{s}$	0.54	Seconds	Temporal uncertainty on player-ball intercept time.	$\mu = 0.5, \sigma = 0.1$	Use value from fit in [7]
$\boldsymbol{\lambda}$	3.99	Hz	$1/\lambda$ is proportional to the average time it takes a player to control the ball.	$\mu = 4.2, \sigma = 2.0$	Use value from fit in [7]
$\boldsymbol{\kappa}$	1.72	None	Defensive advantage, scales control rate for defending players.	$\mu = 1.5, \sigma = 0.5$	We expect moderate defensive advantage
$\boldsymbol{\sigma}$	23.9	Meters	Related to the mean distance between on-ball events.	$\mu = 14, \sigma = 10$	Use mean distance computed in Figure 3
$\boldsymbol{\alpha}$	1.04	None	Preference for maintaining possession increases with α.	$\mu = 1, \sigma = 0.2$	Expect approximate proportionality with PPCF
$\boldsymbol{\beta}$	0.48	None	Small values improve chance of scoring further from the goal.	$\mu = 0.5, \sigma = 0.5$	Open space improves scoring probability[6].

Table 1. A table describing the model parameters, the prior used and the maximum a posteriori probability (MAP).

4. Validation

In constructing the *off-ball scoring opportunity* model, our objective has been to produce a leading indicator of scoring that is less stochastic than scoring itself and allows us to assign opportunity

[6] In this model, β can be thought of as fudge factor to ensure that the resultant model can be integrated to give expected scoring. In future iterations of the analysis, this parameter can be replaced by using an improved scoring probability model.

creation to players even if no shot or score results. In other words, we want to build a metric that is more predictive of future scoring than scoring itself is. To validate that we have achieved this, we compute the correlation between three leading indicators of future scoring: 1) *OBSO* 2) shots and 3) goals themselves and we compare them to the scoring in a subsequent match on a per-player basis. Looking at Table 2, we see that *OBSO* is more correlated with future goals than either shots or goals themselves are.

i\i+1	OBSO	Shots	Goals
OBSO	0.60	0.37	0.26
Shots	~	0.35	0.17
Goals	~	~	0.12

Table 2. The Pearson correlation coefficient (PCC) between the current (ith) and the subsequent (i+1th) game for players in the 53 game test set for three performance indicators: OBSO, shots, and goals.

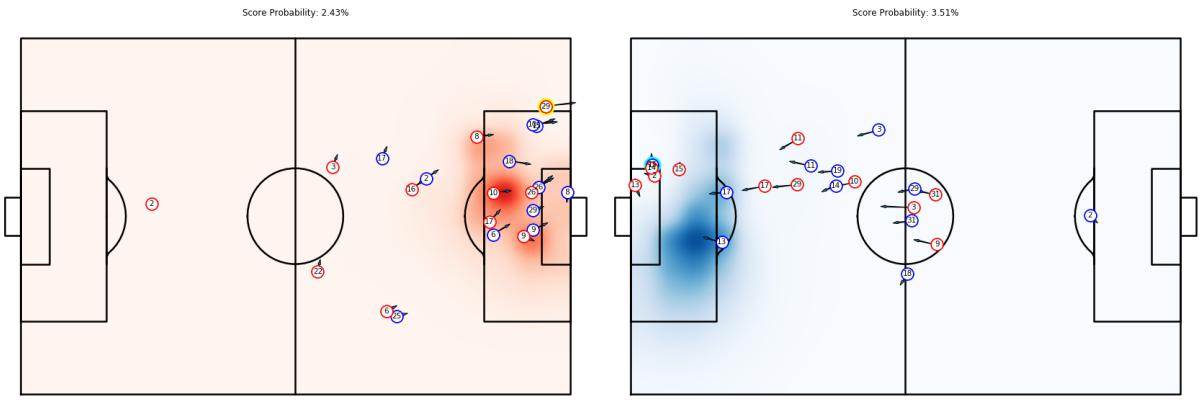
In addition, *OBSO* is correlated with both itself and with future shots. This indicates that *OBSO* can be used as a leading indicator of player performance that is less stochastic than goals or shots themselves are.

5. Applications

In this section, we use the *off-ball scoring opportunity (OBSO*) model to propose applications in four categories: 1) tactical moment analysis 2) match analysis 3) team performance and 4) player performance. Although we believe that the control model (*PPCF),* seen in Figure 5b, and the transition model, seen in Figure 5c, have many interesting applications apart from scoring, these remain outside the scope of this paper. The applications discussed in the subsequent sections focus on *off-ball scoring opportunities*.

5.1. Tactical Moment Analysis

An important aspect of opposition analysis and post-match analysis for the analyst at a soccer club lies in identifying critical moments during the course of the match. This is a time consuming job and requires many hours of video review. The *opportunity* model presented in this paper both *quantifies* and *visualizes* scoring *opportunities*. The integrated magnitude of the *opportunities* can help in choosing which clips to watch while the *opportunity* maps themselves provide additional insight into the *opportunity*. Four examples are shown below.



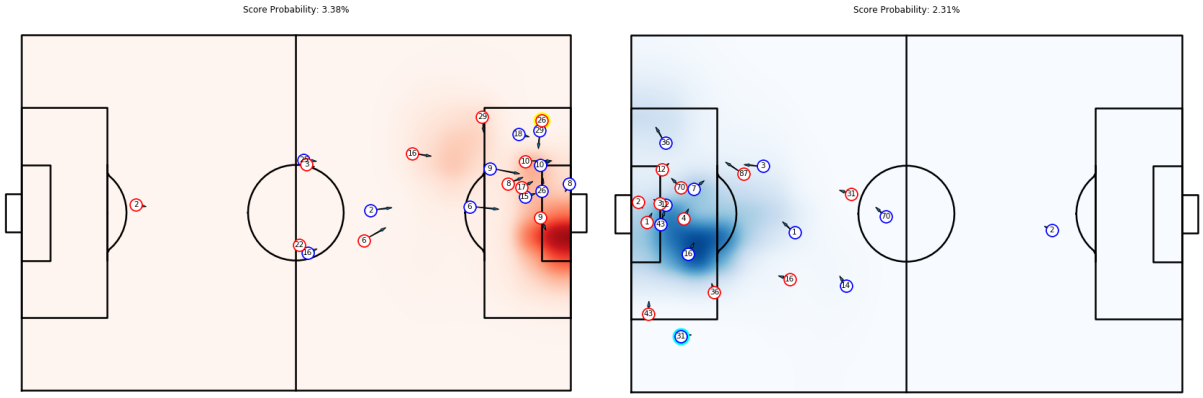


Figure 6. This figure shows selected opportunity maps. a) Upper left: Red #29 has the ball. The integrated scoring chance is 2.43%. b) Upper right: Blue #15, obscured by Red #14, has the ball, integrated scoring chance is 3.51%. c) Lower left: Red #26 has the ball. The integrated scoring chance is 3.38%. d) Lower right: Blue #31 has the ball. The integrated scoring chance is 2.31%.

In Figure 6a, Red #29 controls the ball and is running quickly towards the end line.[7] Meanwhile, Red #10 and Red #17 have created space at the top of the box and are open to receive a lobbed cross. There are other target options including Red #9 who is closer to the goal but because of the defensive coverage from Blue #9 and the distance between Red #29 and Red #9, there is a lower probability of scoring from this pass target.

In Figure 6b, Blue #15 (obscured by Red #14) has the ball and appears to have attracted the attention of the defense. This has left Blue #13 free to make a run into the box in excellent scoring position. The pass will be difficult in a way that is not fully captured by the model due to the proximity of the defenders on Red #14, but if the pass reaches Blue #13, it is an easy score.

In Figure 6c, Red #26 is on the left side of the box. There are a number of teammates near him inside the box, but the best scoring chance comes from the long cross to Red #9 who is wide open on the six-yard line near the far post.

In Figure 6d, Blue #31 has the ball and is relatively free to run into the box himself,[8] but Blue #16 is open and in excellent position to receive a pass and take a shot before being closed down by the defense. The distance from goal and the presence of the keeper limits the overall magnitude of this chance.

5.2. Match Analysis

By integrating the *opportunity* maps throughout the course of the match, we can get a sense of how the match played out spatially and temporally.

[7] One limitation of our model in its current form is that it does directly not account for the difficulty of crossing while running at speed. This has the effect of over-estimating the scoring chance from a scenario such as that in Figure 6a while underestimating the scoring chance for a passer who is stationary and under no defensive pressure.
[8] Note: the opportunity for the player in possession of the ball to score is not considered in the *OBSO* model since this is accounted for during the previous moment before the ball was passed to this player.

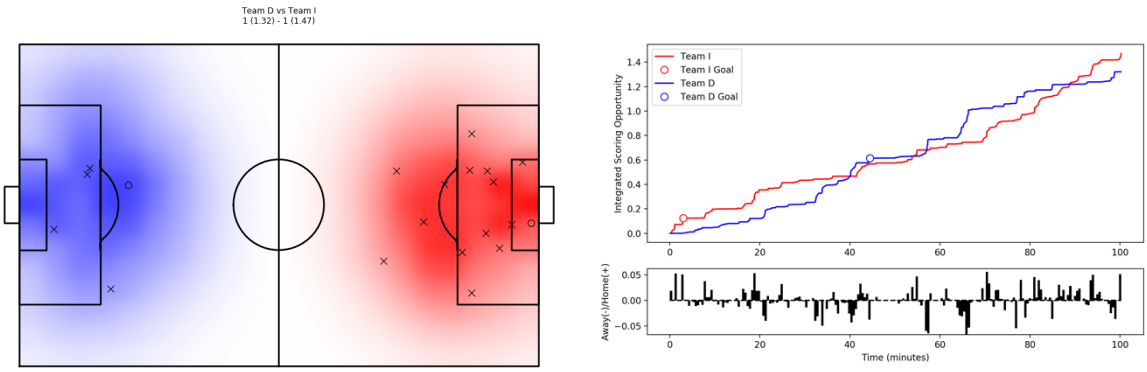


Figure 7. Match analysis figures for a game between Team I (home team in red) and Team D (away team in blue) with a final score line of 1-1. a) The left figure represents the time-integrated spatial opportunity map. A shot is denoted by an "x" and goal is denoted by an "o". b) The right figure demonstrates a space-integrated time varying opportunity map. The upper plot shows the integrated scoring opportunity versus time while the bottom plot shows the instantaneous scoring opportunity per team with positive values for the home team and negative values for the away team. Scores are denoted with an "o".

Figure 7 shows possible visualizations that can be used to represent the *opportunity* model during the course of a match. In this match, Team I dramatically outshot Team D which would generally lead to a mismatch in the number of expected goals in favor of Team I, however, their integrated scoring opportunities are remarkably similar, 1.32 versus 1.47, a similarity reflected in the final score line of 1-1. Notice how in the time-varying plot on the right, Team I creates many opportunities in the first 20 minutes resulting in one score. At the 30-minute mark, the momentum begins to shift to Team D culminating in a score right before the half.

5.3. Team Performance

As with expected goals, *off-ball scoring opportunity* can be used to identify trends at the team and at the player level. Over a large number of games, we would expect the average *OBSO* to regress to the average number of goals scored, but because *OBSO* does not include information about whether the *opportunities* were capitalized on, deviation from the average correlation between goals scored and *OBSO* can be used as a proxy to measure team decision making within and around the penalty area and of finishing skill.

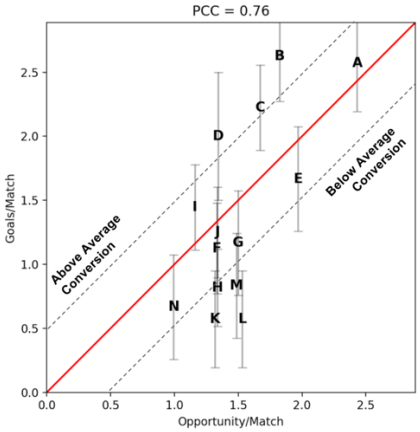


Figure 8. The average goals/match plotted versus the average opportunity/match for the 14 teams in the test set of 53 games. The Pearson correlation coefficient is found to be 0.76.

In Figure 8, we can see how each team's goals per match compares to their *opportunity* per match. As expected, the higher ranked teams tend to have higher goals scored and *opportunity* created per match. Interestingly, there is a cluster of lower ranked teams (Team H, Team K, Team L, and Team M) that all exhibit below average *opportunity* conversion. Apart from the element of randomness in scoring, there are three main reasons that could explain this worse than expected conversion rate: 1) inferior skill in passing or receiving 2) inferior awareness of opportunities and/or 3) inferior finishing.

5.4. Player Performance

As with our analysis of teams, we can highlight player performance and look for trends in player behavior. In Table 3, we can see information about the top 20 players ranked by their per-match *off-ball scoring opportunity*. Unsurprisingly, the majority are center forwards but a few who do not play that position are also in the top 20. Most interesting is the right back from Team A who is clearly an attacking threat as borne out by their scoring production and their mean *OBSO*/match.

Team	Rank	Position	Mean OBSO/Match	Mean Goals/Match	# Matches
Team E	**1**	Center Forward	0.44	0.00	1
Team A	**2**	Center Forward	0.34	0.20	5
Team A	**3**	Center Forward	0.34	0.29	7
Team E	**4**	Center Forward	0.32	0.17	6
Team J	**5**	Center Forward	0.31	0.00	2
Team C	**6**	Center Forward	0.29	1.00	5
Team C	**7**	Right Winger	0.28	0.67	9
Team B	**8**	Center Forward	0.28	0.50	8
Team B	**9**	Center Forward	0.27	0.86	7
Team G	**10**	Center Forward	0.26	0.17	6
Team A	**11**	Center Forward	0.26	0.14	7
Team A	**12**	Attacking Midfielder	0.25	0.43	7
Team C	**13**	Center Forward	0.25	0.56	9
Team L	**14**	Right Midfielder	0.24	0.20	5

Team H	**15**	Center Midfielder	0.24	0.00	1
Team E	**16**	Left Winger	0.24	0.67	6
Team H	**17**	Center Forward	0.24	0.14	7
Team D	**18**	Center Forward	0.23	1.00	4
Team M	**19**	Attacking Midfielder	0.23	0.00	6
Team A	**20**	Right Back	0.23	0.29	7

Table 3. This table provides information on the top 20 players ranked by their average off-ball scoring opportunity per match. Additionally, we show the position they play, which team they play for, the mean goals scored per match and the number of matches which they participated in (in our test set of 53 games).

Another interesting player is the midfielder on Team M (19th ranked). Despite playing for a weaker side and having no goals, it appears the midfielder is able to create substantial space in dangerous areas judging from their mean *OBSO*/match. It is possible that this player could become a scoring threat in the right situation.

Certain players have distinctive danger zones where most of their shots and goals come from. An example of this is the 20th ranked player in Table 3, the right back from Team A. Whether it results in a score or not, this player tends to create *opportunities* from distance on the right side of the pitch. Notice how the *opportunity* maps maintain a similar profile over the four games shown in Figure 9 even though the scoring and shooting output varies from match to match.

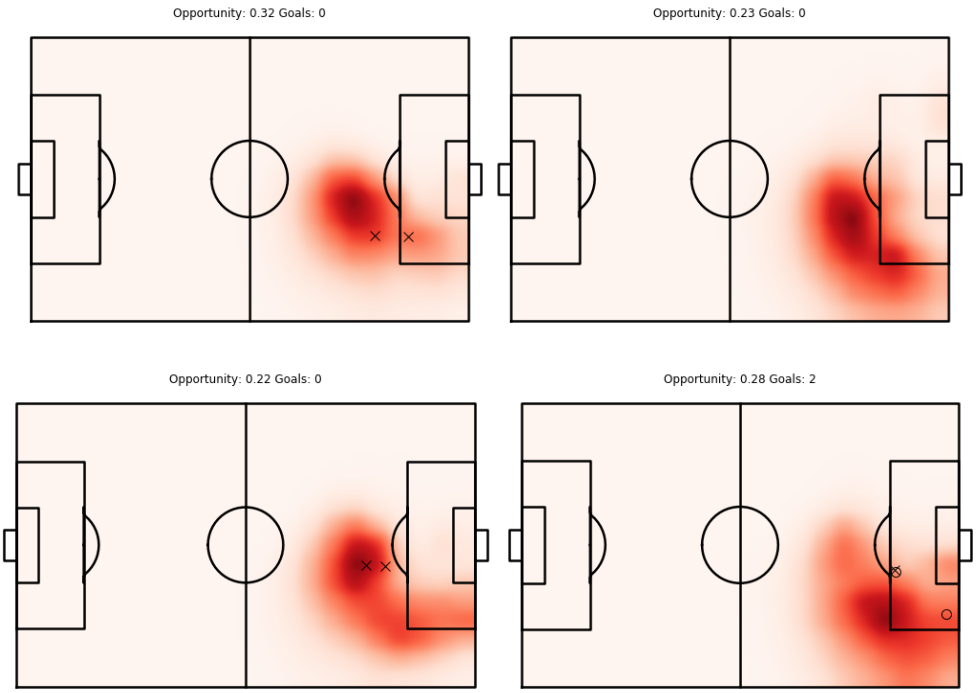


Figure 9. Four off-ball scoring opportunity maps for the right back from Team A (ranked #20 in Table 3). Each map represents a game. Integrated opportunity is shown above each plot. Shots are shown by an "x" and goals are shown by an "o".

Compare these to the scoring maps for a center forward found in Figure 10. This particular center forward appears to play a bit left of center creating space within the penalty area and inside the 6-yard line.

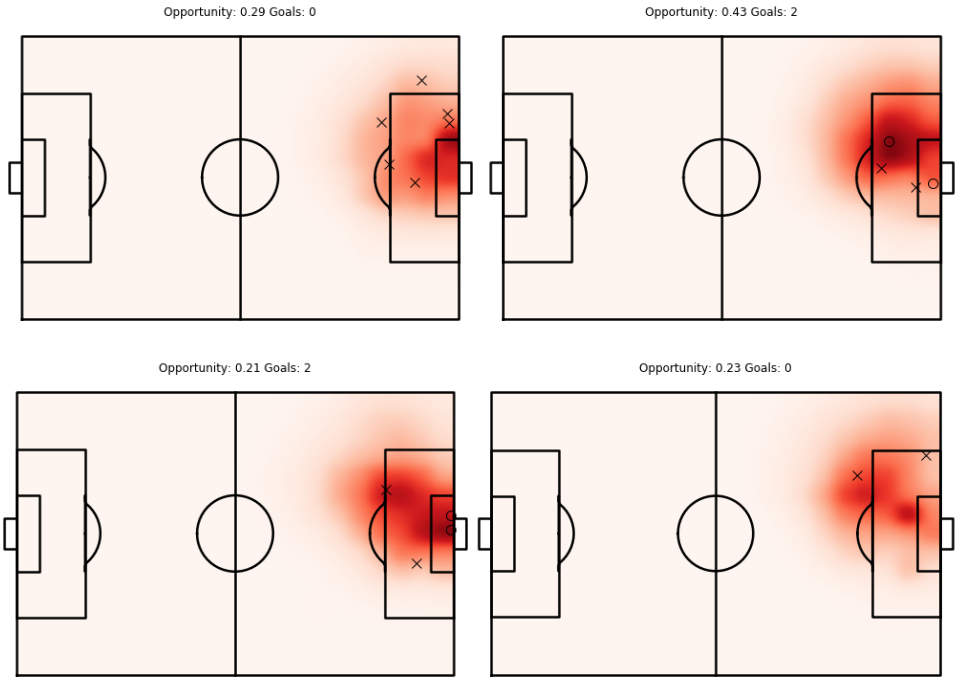


Figure 10. Four off-ball scoring opportunity maps for a center forward from Team C (ranked #6 in Table 3). Each map represents a game. Integrated opportunity is shown above each plot. Shots are shown by an "x" and goals are shown by an "o".

6. Conclusions and Discussion

The increasing prevalence of spatiotemporal tracking information in soccer has dramatically broadened the number of questions the analyst can ask the data. We have presented a new approach to using this tracking data to quantify *off-ball scoring opportunity (OBSO)*. This metric can be used as a leading indicator of future player scoring and there are many possible applications for the *opportunity* model. The *OBSO* can be used by club analysts to expedite the process of discovering key moments during post-match analysis. Player-specific opportunity maps can be used by opposition analysts to identify dangerous regions that may need extra defensive attention to mitigate the attacking threat posed by a specific opposing player in an upcoming match. For scouting, the *OBSO* can be used to identify talented players with good spatial awareness who could thrive if given the opportunity.

The model presented in this paper is not without its limitations and future work will be needed to refine the scoring model (Section 3.3) and the effect that defensive pressure and player speed has on the ability of a player to successfully deliver the ball to a teammate. Despite these limitations, we are excited by the use of spatial probability densities to predict future game state over short time scales and we believe that this approach and the constituent models that describe control (Section 3.1) and transition (Section 3.2) can serve as the basis for further soccer research.

Acknowledgements

I would like to give a special thanks to the Hudl Analysts and the Hudl Research & Development team who worked so tirelessly to produce the spatiotemporal tracking data and event data that made this research possible.

References

[1] S. Green, "Assessing The Performance of Premier League Goalscorers," 12 Apr 2012. [Online]. Available: http://www.optasportspro.com/about/optapro-blog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/. [Accessed 4 Dec 2017].

[2] W. Gurpinar-Morgan, "Valuing Possession," 25 Aug 2015. [Online]. Available: https://2plus2equals11.com/2015/08/25/valuing-possession/. [Accessed 4 Dec 2017].

[3] D. Altman, "OptaPro Forum: Beyond Shots," 9 Mar 2015. [Online]. Available: http://www.optasportspro.com/about/optapro-blog/posts/2015/film-optapro-forum-beyond-shots/. [Accessed 4 Dec 2017].

[4] J. Boice, "How Our Soccer Club Projections Work," 19 Jan 2017. [Online]. Available: https://fivethirtyeight.com/features/how-our-club-soccer-projections-work/. [Accessed 4 Dec 2017].

[5] P. Lucey, A. Bialkowski, M. Monfort, P. Carr and I. Matthews, "'Quality vs Quantity': Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data," in *MIT Sloan Sports Analytics Conference*, Boston, 2015.

[6] D. Link, S. Lang and P. Seidenschwarz, "Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data," *PLoS ONE,* vol. 11, no. 12, p. e0168768, 2016.

[7] W. Spearman, P. Pop, A. Basye, R. Hotovy and G. Dick, "Physics-Based Modeling of Pass Probabilities in Soccer," in *MIT Sloan Sports Analytics Conference*, Boston, 2017.

[8] A. Savitzky and M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry,* vol. 36, no. 8, pp. 1627-1639, 1964.

[9] T. Asai and K. Seo, "Aerodynamic drag of modern soccer balls," *SpringerPlus,* vol. 2, no. 1, p. 171, 2013.