StatsBomb Conference 2022

“Estimated Player Impact” (EPI): Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models

Paper Track

Tahmeed Tureen, MS <tureen@umich.edu>
Sigrid Olthof, PhD <S.B.Olthof@ljmu.ac.uk>

1 Introduction

In recent years, football clubs have started to leverage advanced metrics and analytics for performance analysis, opposition scouting, and player recruitment [1, 2]. The expected goals (xG) metric has been one of the main drivers of this analytics revolution in football [1-4]. The models developed for xG enable analysts to retrospectively assess whether a team had under or over-performed in a match by comparing their goals with their goal-scoring opportunities [2, 5]. Now-a-days, even football broadcasters provide xG as a statistic in their tactical and match analyses [5]. Pundits can also use these metrics to add a statistical layer of analysis to their discussions when comparing teams and players over the course of a season [5]. An xG model is a probabilistic model that assigns a score between zero and one to any observed shot in a football match [3, 5]. The philosophy of the model is that: given a set of shot characteristics or predictors, the model estimates a probability score for the observed shot. This score, therefore, represents the estimated probability of a shot being converted to a goal based on a set of characteristics related to the shot [2, 3, 6].

xG models can vary due to differences in the set of shot characteristics used in development [3, 5, 7]. However, these models are typically developed using event-level football data like StatsBomb’s event data [7]. Event data is collected by tracking players over the course of a football match and logging their actions such as shots, passes, and tackles. For an xG model, the characteristics related to the **shot event** are then included as predictors in a probabilistic model [3, 5, 7]. Commonly used predictors are shot location, distance to goal, shot angle, type of play, body part used to shoot, shot type, and shot technique [5]. Newer models have also started to incorporate predictors beyond shot characteristics, such as goalkeeper location, defender location, received pass type, and whether the shot-taker is under pressure [5-7]. xG models are still regularly fine-tuned and new approaches continue to be proposed in football analytics [7].

An interesting trait of xG models is that they do not account for the players who take the shots. In statistical terms, there is no player predictor in the model. The absence of the player in xG models seems contrary to the nature of football, where a player's skill influences the success of shot-conversion. It is reasonable to assume that a striker would have a higher chance of scoring a goal as opposed to a center back. Moreover, not all strikers have the same goal-scoring capabilities. However, in an xG model, two separate shots that have the same measures for the model predictors will be assigned the exact same xG regardless of who is taking the shot. This is arguably a limitation of current xG models. The inclusion of players in xG models along with their shot characteristics can provide information about the skills of the shot-taker as well. Rather than a general xG measure, an individualized approach can estimate the player's effect on goal-scoring which represents their contribution to the success of a goal.

All event-level football data have repeated measures for multiple players in the dataset. Since players make multiple passes, engage in a number of tackles, and take several shots over the course of a match, they can show up more than once in the data. All of these events are logged under the players' names. As a result, any given player can be repeated in the dataset. In statistical terms, the events are nested under a player hierarchy in the data. This hierarchy is illustrated in Figure 1.

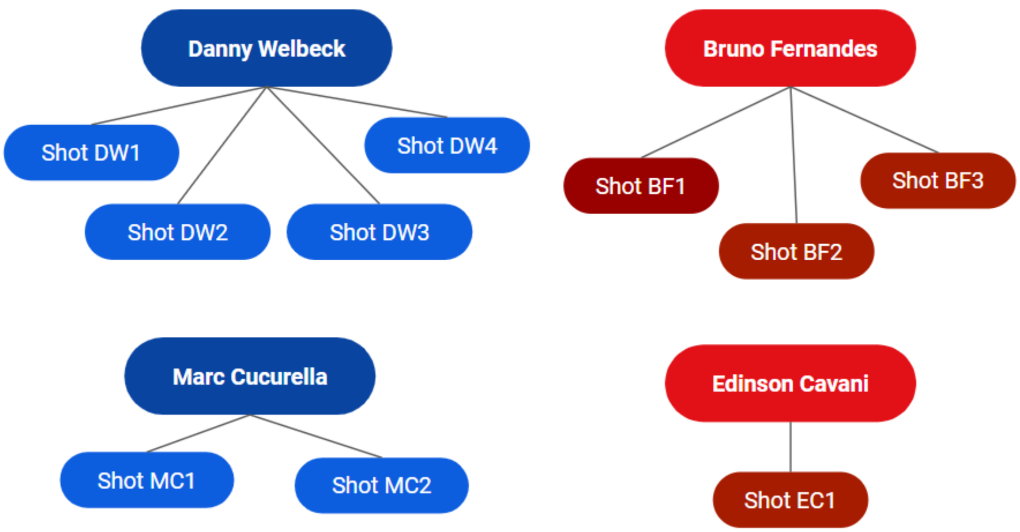


Figure 1. This diagram illustrates the hierarchical structure of event-level data for four different players collected from Brighton's 4-0 win over Manchester United on May 7th, 2022. In this match, Danny Welbeck (DW) took 4 shots, Marc Cucurella (MC) took 2 shots, Bruno Fernandes (BF) took 3 shots, and Edinson Cavani (EC) only took 1 shot.

This type of data hierarchy implies that the events that are associated with any specific player, are statistically correlated with each other. Therefore, this repeated measures structure violates one of the fundamental assumptions of most statistical and machine learning models: all observations are independent of each other. Due to this type of data hierarchy, the observations from the same player in football event data are not independent of each other. Popular probabilistic models used in football analytics like logistic regression and tree-based boosting models do not account for this within-player correlation. However, the application of these models to repeated measures can lead to improper inferences [8]. From a modeling framework, a player indicator could be included as a predictor to the aforementioned models. However, a logistic regression requires one player to be the reference (dummy) level and a tree-based model would have as many sparse predictors as players in the dataset. As a result, neither model has the ability to estimate “an effect” of a player on a football action. In football, a key assumption is that players have different skill sets and are different from each other. Any model that does not account for the shot-taker, will inherently ignore this assumption and treat every player as the same in the dataset. As a consequence, the current xG models estimate the same xG for an identical shot (all predictors are identical). Whereas, a player-adjusted xG model can provide an xG estimate for each individual player.

It is therefore important to use hierarchical statistical models that account for the repeated measures in event data and also inherently address the players in its model framework. These models are called hierarchical or multilevel models because (i) the model is fit on data that has a hierarchy and (ii) the model itself has a hierarchy in terms of its parameters [8]. Hierarchical models have shown to be beneficial in psychology [9], clinical trials [10], ecology [11], and more recently in basketball analytics [12]. The particular subset of hierarchical models used in this research are called Generalized Linear Mixed Models (GLMM). Due to the linear nature of these models, they can also provide interpretable predictor effects on xG. This type of analysis is not possible for tree-based models as they are black-box models [3]. As for logistic regression, the standard errors for its predictor effects are biased because it does not account for the within-player correlation [8].

For the purposes of this research, a player-adjusted xG model will be introduced and applied to the men’s and women’s game using StatsBomb event-level data. Additionally, the concept of player-adjusted football models can be applied to almost all of the events in StatsBomb’s data and used to refine other advanced metrics such as “xA” [13], “xT” [14], and “xGChain” [15]. Including player information directly into football analytics models has the potential to improve player analysis, scouting, and recruitment for football clubs and football researchers.

2 Research Aims

2.1 Data

Event-level data as provided by StatsBomb is used in this research. The datasets contain events that occured in 580 matches from the 2020/21 and 2021/22 seasons of the English Premier League (EPL) and 326 matches from the 2018/19, 2019/20, and 2020/21 seasons of the Women's Super League (WSL) in England. The events in the data refer to football actions such as shots, passes, and dribbles. Each event is labeled with different characteristics of a particular football action, including the player, their position on the field at the time of the event (in x-y coordinates) and granular information about the action itself. Table 1 presents a subset of what the event data looks like for shots. In this example, Mohamed Salah has three shots, Mason Mount has two shots, and Mateo Kovacic has one from the same match. The shot characteristics, however, can differ for each shot. This is illustrated with a detailed hierarchical diagram for one of these players in Figure 2.

Table 1. Example of StatsBomb event data for shots only. The events and shot characteristics are directly pulled from the real data. The Player Location is stored as x, y coordinates on a football pitch as determined by StatsBomb. The Event and Match IDs were re-defined for illustrative purposes.

Event ID	Match ID	Player	Player Location (x,y)	Body Part	..	Shot Outcome
abc1	CHELIV2	M. Salah	(102.7, 39.9)	Right Foot	..	No
abc2	CHELIV2	M. Salah	(115.8, 51.5)	Left Foot	..	Yes
abc3	CHELIV2	M. Mount	(109.7, 51.8)	Right Foot	..	No
abc4	CHELIV2	M. Salah	(90.7, 30.8)	Left Foot	..	No
abc5	CHELIV2	M Kovacic	(97.6, 37.2)	Right Foot	..	Yes
abc6	CHELIV2	M. Mount	(94.5, 30.0)	Right Foot	..	No
..	..	..	..	..	..	..
xyz100	BRIMUN2	P. Gross	(117.8, 36.8)	Left Foot	..	Yes
xyz111	BRIMUN2	P. Gross	(101.7, 41.6)	Right Foot	..	No
xyz112	BRIMUN2	L. Trossard	(118.2, 39.7)	Other	..	Yes

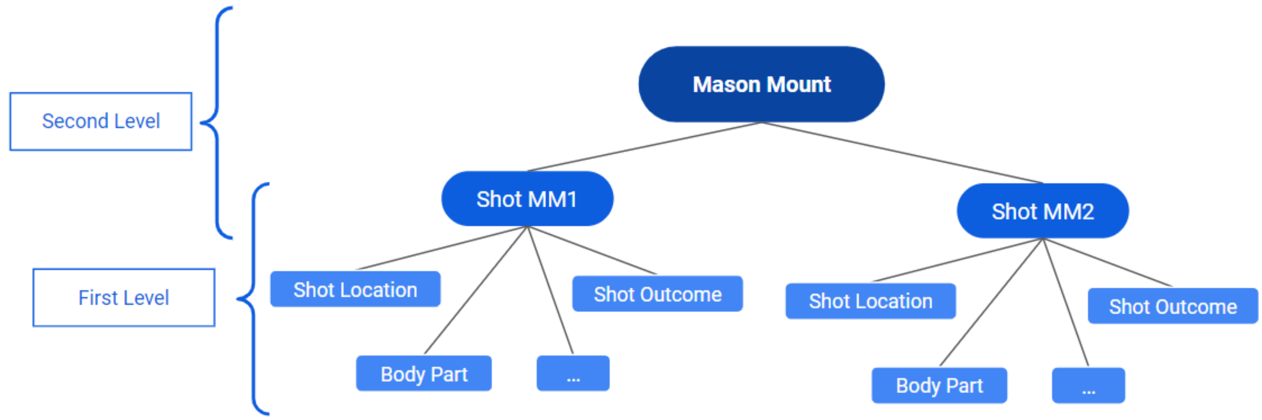


Figure 2. This hierarchy diagram is derived from Mason Mount's (MM) shot-data in Table 1. The first or lower level of the data hierarchy represents the characteristics of every shot. The second or higher level of the hierarchy represents the player who is repeated in the dataset.

2.2 Motivation

A reasonable football assumption in this example is that Mount has his own unique shot technique. Therefore, it is reasonable to claim that the shots from Mount are inherently correlated with each other because they are taken by the same player. Whereas, the shots taken by Salah are independent from Mount's shots but are correlated with the other shots from Salah. This aligns with statistical principles, because the data is collected from the same players over a period of time which introduces within-player correlations (event-level data is longitudinal data). As a result, every shot is essentially grouped or nested under a player (see Figure 2). Therefore, models need to be fed information about the hierarchical structure, otherwise it may lead to biased inferences (see Appendix A1 for an example).

2.3 Aims

This research aims to: (i) Develop an xG model that can estimate the goal-scoring probability of each shot in a football match, (ii) Estimate the effects of each player on xG and calculate player-specific xG values for each shot, and (iii) Draw generalizable inferences on how different shot characteristics affect shot-conversion. The focal point of this research is the quantification of player impacts on xG. Therefore, a Generalized Linear Mixed Model (GLMM) is proposed to address the hierarchical nature and repeated measures in football event data. These models inherently account for within-player correlations and provide results for the three research aims. To apply this proof of concept to both men's and women's football, the analyses are stratified for the EPL and the WSL.

3 Methodology

3.1 Probabilistic Models for Expected Goals

xG models are typically probabilistic models [3, 5, 6, 16]. The output of an xG model is interpreted as the estimated probability that a shot becomes a goal [3, 6, 16]. However, the goal variable in football data is binary (i.e. Yes = Goal, No = No Goal). Therefore, xG models use mathematical transformations to treat the target on a probability scale during model training. These models are also trained based on a set of input predictors. For this paper, separate xG models are developed for the EPL and the WSL.

3.2 Data Engineering

Prior to the analysis, multiple data processing and filtration steps are taken. The StatsBomb data as described in Section 2.1 is used for the development of the xG models. First, the event data is filtered for shots. The target or response variable is whether an open play shot results in a goal. Penalty and direct free-kicks are therefore dropped from the analysis because they represent a direct shot from a dead-ball situation.

In addition to the existing variables in the data, new variables are created to provide additional information into the models. These new variables contain information about the player's preferred foot, the positions of the shottaker relative to the goal, the position of the goalkeeper relative to the goal, and shot angles. For each player, their preferred footedness is determined by comparing the number of passes they take with each foot in the data [17]. The distances between the shottaker/goalkeeper and goal are derived using their location data. A shot triangle (cone) is drawn using the goal posts and the shot-taker's location as vertices. The angle associated with the shot-taker vertex is calculated using the cosine rule. Whether a goalkeeper is in the shot triangle is also determined using spatial analysis [6, 7]. The newly derived distance and angle variables contain information about the x-y locations of a shot; therefore the x-y coordinates are dropped from the model to avoid multicollinearity issues.

All continuous predictors are standardized by centering on the mean and scaling by their standard deviation before model training. The multi-class predictors are dummy-encoded such that one of the classes is the reference level. The reference level for each predictor can be viewed in Table 4. The binary predictors are one-hot encoded and the count variables are treated as is.

3.3 Model Predictors

The predictor variables are selected and derived from the data based on literature review as well as domain expertise. Certain predictors are not included in the final model to avoid multicollinearity issues. The predictors used in the models are as follows:

(i) Four continuous predictors: shot-taker's distance to goal [16, 18], goalkeeper's distance to goal [6], angle of shot triangle/cone [7, 16, 18], and shot impact height [19]

(ii) Six binary predictors: first time shot, goalkeeper's presence in shot triangle [6, 7], one on one shot, open goal shot, side of pitch from shot-taker's perspective, and shot-taker under pressure [6]

(iii) Two multi-class predictors: body part and shot technique [6, 16, 18]

(iv) One count predictor: number of defenders in shot triangle [6, 7].

3.4 Analytical Sample

The EPL and WSL datasets are filtered for open play shots. Any observation that has missing values for a model predictor is dropped. The WSL data in this research does not have measures for the "shot impact height" predictor and is therefore dropped from the WSL analysis. The EPL analytical sample consists of 13,938 shots from 553 unique players and the WSL sample consists of 7,928 shots from 327 unique players. It is important to note that there is an imbalance in shot distribution across players. In other words, not all of the players have the same number of shots.

3.5 Generalized Linear Mixed Model (GLMM)

The proposed models in this research are an extension from the foundations of non-hierarchical xG models. Therefore, it is of interest to first discuss what types of statistical models are currently used for xG development then unpack the proposed models mathematically.

GLM: A popular model for binary target variables is logistic regression which is a generalized linear model (GLM) [3, 16]. This model linearly learns the relationships between a set of predictors and the target. A trained model can then be used to calculate xG probabilities based on its predictors. Equation 1 illustrates the framework of a GLM:

$$g(\pi_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p}$$

Equation 1

In this equation, the i refers to the i-th shot in the dataset. The g() represents the logit link function and the π_{ij} term is the odds of a goal (i.e. $\frac{Pr[Y_i=1]}{Pr[Y_i=0]}$; Pr[] is probability) for the ith shot. The choice of the logit function is what makes this GLM a logistic regression. The x_i terms are the model predictors and the associated β terms represent the predictor effects. The β_0 represents the model intercept or baseline (i.e. when the predictor values are all zero). The GLM is not fed any information about the players as seen in Equation 1. This implies that the model fails to adjust for the within-player correlation. A categorical predictor could be included, however, that would require specifying one player as the reference level.

XGBoost: Another popular model in football analytics is the extreme gradient boosting (XGBoost) algorithm [3, 16]. It learns the relationships between a set of predictors and the target using gradient boosting decision trees [3]. XGBoost can also learn non-linear relationships between the variables since it is not a linear model like the GLM. However, this approach also does not account for the hierarchy in event data. The players could potentially be included as one hot-encoded predictors in the model. However, that would lead to several ungeneralizable features in the tree model (high cardinality issue).

GLMM: To appropriately adjust for the players in the data, a new parameter for the players can be incorporated to the GLM framework to create a Generalized Linear Mixed Model (GLMM). By doing so, the GLM framework is then extended from a single-level to a multilevel framework (Equation 2):

$$g(\pi_{ij}) = \beta_0 + \beta_1 x_{ij,1} + \beta_2 x_{ij,2} + \ldots + \beta_p x_{ij,p} + \delta_j$$

Equation 2

This framework is almost identical to the GLM except for the j index and the δ_j term. The j index refers to the j-th player in the data. The δ_j term is a statistical parameter that represents the random effect associated with the j-th player. As the model is trained, it computes the values of the parameters in the equation including the δ_j. As a result, each player in the dataset will have their own estimated δ_j measure. This implies that each player will have their own unique intercept or baseline for a shot that they take. This player-specific baseline can be derived by calculating the equation, $\beta_0 + \delta_j$, once the parameters are estimated. The δ_j parameter can be included in the framework because the data has a hierarchy. If the data lacked a hierarchy, a GLM framework would suffice for analysis [8]. The particular GLMM framework used here is a mixed model with fixed effects for all predictors and a random intercept for players. The GLMM learns patterns from the input data and estimates values of the $\beta's$ and δ_j, This model can also provide uncertainty measures for the predictor effects in the form of 95% Wald Confidence Intervals (CIs) and the standard deviations for the random effects. The **lme4** package (version 1.1-21) in **R** is used to develop the GLMMs.

3.6 Estimated Player Impact (EPI)

From a statistical perspective, the random effect can be interpreted as the baseline change on xG that is attributable to the players: this change is **unmeasured** by the rest of the predictors in the model. However, explained by football terminology, the measure can be used and interpreted as the "estimated player impact" (EHI) on shot conversion. The unique skills from each player can now be statistically estimated by deriving their δ_j which quantify the effects of players on xG. The δ_j is a continuous measure. This implies that certain players have positive effects on the xG while others have negative effects. A positive value for a player would imply that they increase the xG value in the estimation. Whereas, a negative value for a player would decrease the xG estimation. It is important to note, however, that because GLMMs are fit using a logit link, these effects actually have a multiplicative impact on the xG scale. The player impacts derived in the models are only comparable between the players in their respective analytical samples. This means that the EPL EPIs are not comparable to the WSL EPIs as the models were stratified.

---

4 xG Prediction Experiments

4.1 Experimental Context

The first step in the analysis pipeline is the development of the xG models using GLMM. However, before drawing insights from the GLMM, it is important to empirically compare it to competing models. In this section, the GLMM is compared to logistic regression (GLM) and XGBoost to demonstrate its predictive capabilities. The GLM follows the same framework as Equation 1 and no shrinkage estimators are applied. The GLM is built using the **stats** package (version 3.6.2) in R. Four different XGBoost models are trained and labeled as XGBoost10, XGBoost25, XGBoost50, and XGBoost100. The specifications of these XGBoost models are provided in Appendix A3. These models are generated using the **xgboost** package (version 1.4.1.1) in R.

4.2 Evaluation Criteria

The models are evaluated using train-test split experiments. Normalized variables in both the train and test sets were standardized using the mean and standard deviation calculated from the train set. This is done to avoid information leakage from the test set. For each experiment, there is a naive model that assigns the goal proportion (number of goals / number of shots) in the train set as the xG value for the test shots. This naive model serves as the baseline for each experiment. To evaluate the performance of each model, the Normalized Brier Score (NBS) is used for xG estimations [3]:

$$Brier\ Score\ (BS) = \frac{1}{N}\sum_{ij}^{N}\left(p_{ij} - y_{ij}\right)^2$$

Equation 3

$$Normalized\ Brier\ Score\ (NBS) = \frac{BS_{new\ model}}{BS_{naive\ model}}$$

Equation 4

The $\left(p_{ij} - y_{ij}\right)^2$ in equation 3 represents the squared prediction error between a model's probability estimate p_{ij} and the ground truth y_{ij}. The NBS is the ratio of a model's BS to the naive BS in each experiment. The NBS can range from zero to positive infinity. The smaller the NBS, the larger the improvement a model has relative to the baseline. If the NBS is greater than 1, then the model is performing worse than the baseline. If the NBS is zero, then the model is perfectly predicting every shot in the test set.

4.3 Evaluation Data (Train-Test Sets)

Analyses on four separate train-test splits are performed to evaluate the models. The first train-test is a random split from the EPL data. The training and test sets consist of 70% and 30% of the shots respectively. This random sampling does not take into account that there are two seasons worth of data. For the second split, the 2020/21 EPL data is used for the training set and the 2021/22 EPL data is used for the test set. As for the third split, shots from the 2018/19 and 2019/20 WSL season are used for training and the 2020/21 WSL shots are used for testing.

One of the limitations of the GLMM model is that if a player is not available in the training set their random effect will be zero in the testing. Therefore, xG values for new players in a test or future set do not include a player effect. This phenomenon is present in all three of the experiments. Therefore, a fourth split consisting of shots from players who played both seasons of the EPL is performed: the models are trained on data from the 2020/21 season and evaluated on the 2021/22 season. The breakdown of the train-test splits are shown in Table 2:

Table 2. Breakdown for the train-test splits for the prediction experiments

Train-Test Split	Train Set No. of Shots	Train Set No. of Goals (%)	Train Set No. of Players	Test Set No. of Shots	Test Set No. of Goals	Test Set No. of Players
#1	9757	940 (9.63%)	536	4181	460 (11.0%)	473
#2	6678	679 (10.2%)	410	7260	721 (9.93%)	422
#3	4820	512 (10.62)	248	3108	354 (11.39%)	221
#4	5657	591 (10.44%)	279	5697	576 (10.11%)	279

4.4 Experiment Results

The experiment results are summarized in Table 3. In the first experiment, "XGBoost 50" has the best predictive performance as indicated by the lowest NBS value. Whereas in the other three, both GLM and GLMM outperform the tree-based models. The type of dataset has an influence on the NBS values and can lead to different types of performance results. This behavior can be seen in these experiments (experiment #1 versus the rest). Overall, the GLMM is comparable to the XGBoost models and GLM using the current data and predictors.

Table 3. Normalized Brier Scores (NBS) for the six different models in the experiments across the four separate train-test splits. **Bold*** values indicate the best performances

Model	NBS Test #1	NBS Test #2	NBS Test #3	NBS Test #4
XGBoost 10	0.8707	0.8722	0.8667	0.8814
XGBoost 25	0.8677	0.8755	0.8701	0.8822
XGBoost 50	**0.8615***	0.8641	0.8554	0.8730
XGBoost 100	0.8686	0.8747	0.8685	0.8844
GLM	0.8684	**0.8615***	**0.8348***	**0.8571***
GLMM	0.8679	0.8616	0.8364	**0.8571***

5 Results

5.1 Descriptive Statistics

The descriptive statistics for the analytical sample are summarized in Table 4. The inferences drawn from the GLMM are contextualized by these numbers. In other words, the predictor and player effect estimates are generalizable only for the respective seasons in the data. This is important because the data here does not contain a player's full career.

Table 4. Summary statistics for the analytical samples used in the research for both the English Premier League (EPL) and Women's Super League (WSL). The values in parentheses represent the standard deviation for continuous variables and percentages for categorical variables. The EPL data is from 20/21 and 21/22 seasons. The WSL data is from 18/19, 19/20, and 20/21 seasons.

Characteristic	EPL	WSL
Matches [n]	580	326
Shots [n]	13,938	7,928
Goals [n]	1,400 (10.0%)	866 (10.9%)
Players [n]	553	327
Seasons [n]	2	3
Body Part		
Reference: Preferred Foot	8,905 (63.9%)	5,035 (63.5%)
Head	2,417 (17.3%)	1,267 (16.0%)
Other	43 (0.31%)	16 (0.21%)
Other Foot	2,573 (18.5%)	1,610 (20.3%)
Distance from Goal [SB units]	17.8 (7.56)	18.3 (8.81)
Distance of Goalkeeper from Goal [SB units]	3.39 (2.23)	2.97 (2.77)
First Time Shot	4,499 (32.3%)	2,251 (28.4%)
Goalkeeper in Shot Triangle	13,487 (96.8%)	7640 (96.4%)
Defenders in Shot Triangle [n]	1.08 (1.09)	1.13 (1.24)
One-on-One Shot	641 (4.60%)	382 (4.82%)
Open Goal Shot	146 (1.05%)	92 (1.16%)
Shot Angle [degrees]	26.6 (15.7)	27.5 (18.4)
Shot Impact Height [yards]	0.46 (0.80)	N/A
Shot Technique		
Reference: Normal	10,776 (77.3%)	6435 (81.2%)
Backheel	57 (0.41%)	39 (0.49%)
Diving Header	68 (0.49%)	20 (0.25%)
Half Volley	1914 (13.7%)	876 (11.04%)
Lob	116 (0.83%)	64 (0.81%)
Overhead Kick	59 (0.42%)	16 (0.20%)
Volley	948 (6.80%)	478 (6.03%)
Under Pressure	2,939 (21.1%)	1,607 (20.3%)

[SB units: "StatsBomb Pitch Units"]

5.2 Predictor Effects on xG

The predictor effects derived from the GLMM are summarized in Table 5. The effects in the table are exponentiated and the results are interpretable as odds ratios. Effects greater than one are interpreted as having a positive multiplicative effect on the odds of a shot conversion. Whereas, effects that are less than one have a negative multiplicative effect. These effects are adjusted for all of the predictors in the model. 95% confidence intervals (CI) are also provided for each of the predictor effects. A CI containing a value of one suggests that the relationship between the predictor and response is not statistically significant.

For the EPL, predictors such as Distance of Goalkeeper from Goal, Open Goal Shot, and Shot Angle have positive associations with xG, whereas taking a shot with the other foot (relative to preferred foot), Distance from Goal, number of defenders in Shot Triangle, Shot Impact Height, a diving header (relative to a normal shot) and Under Pressure have negative associations with xG. As for the WSL, predictors such as a headed shot (relative to preferred foot), Distance from Goal, Goalkeeper in Shot Triangle, number of defenders in Shot Triangle, and the shot being a volley or backheel (relative to a normal shot) have negative associations with xG. But, Distance of Goalkeeper to Goal, One-on-One shot, Shot Angle, and the lob shot (relative to normal) have positive associations with xG. Interpretation of the predictor effects vary across the different types of predictors (continuous, categorical, and count). To illustrate, the odds ratios are interpreted for shot angle, body part, and defenders in the shot triangle below.

Shot Angle: For a given EPL player, an increase of one standard deviation (15.7 degrees) from the average angle (26.6 degrees) is estimated to **increase** the xG value by 57% ([1.57 - 1.00] x 100 = 57%) while adjusting for other predictors.

Body Part (Other Foot): For a given EPL player taking a shot with their non-preferred foot, the shot xG is estimated to **decrease** by 18% ([0.82 - 1.00] x 100 = -18%) relative to taking it with their preferred foot while holding other predictors constant.

Defenders in Shot Triangle: For a given EPL player, the appearance of an additional defender in the shot triangle is estimated to **decrease** the xG value by 21% ([0.79 - 1.00] x 100 = -21%) while holding other predictors constant.

The other results can be interpreted similarly by following the examples above. It is important to note that these “effects” are interpreted as statistical associations and **not** causal effects. A visualization of these results is provided in Appendix A4.

Table 5. The odds ratios (predictor effects) and their respective 95% confidence intervals (CI) from the GLMM. **Bolded*** numbers represent odds ratios that demonstrate a statistically significant relationship with the target variable. The effects associated with scaled continuous predictors are interpreted on their respective standard deviation scale.

Characteristic	EPL Effect (95% CI)	WSL Effect (95% CI)
Body Part Reference: Preferred Foot Head Other Other Foot	 1.82 (0.75, 4.44) 0.85 (0.30, 2.36) **0.82 (0.70, 0.96)***	 **0.41 (0.31, 0.54)*** 0.20 (0.05, 0.85) 0.84 (0.68, 1.02)
Distance from Goal (Scaled)	**0.49 (0.43, 0.56)***	**0.49 (0.42, 0.58)***
Distance of Goalkeeper from Goal (Scaled)	**1.22 (1.15, 1.30)***	**1.12 (1.02, 1.22)***
First Time Shot	1.12 (0.96, 1.31)	1.21 (0.99, 1.47)
Goalkeeper in Shot Triangle	0.79 (0.60, 1.05)	**0.68 (0.48, 0.96)***
Defenders in Shot Triangle	**0.79 (0.74, 0.85)***	**0.77 (0.71, 0.83)***
One-on-One Shot	1.20 (0.94, 1.52)	**1.42 (1.06, 1.90)***
Open Goal Shot	**1.83 (1.16, 2.89)***	1.51 (0.88, 2.59)
Shot Angle (Scaled)	**1.57 (1.44, 1.72)***	**1.55 (1.39, 1.73)***
Shot Impact Height (Scaled)	**0.48 (0.35, 0.67)***	NA
Shot Side: Left	0.96 (0.85, 1.08)	0.86 (0.74, 1.01)
Shot Technique Reference: Normal Backheel Diving Header Half Volley Lob Overhead Kick Volley	 0.46 (0.21, 1.02) **2.50 (1.32, 4.74)*** 1.00 (0.80, 1.26) 1.65 (0.98, 2.77) 1.20 (0.38, 3.80) 0.86 (0.63, 1.18)	 **0.11 (0.02, 0.50)*** 2.88 (1.01, 8.19) 0.79 (0.61, 1.02) **8.27 (4.63, 14.77)** 0.67 (0.15, 3.08) **0.61 (0.44, 0.84)***
Under Pressure	**0.85 (0.72, 0.99)***	0.92 (0.75, 1.12)

5.3 Estimated Player Impact (EPI) Rankings

The EPI measures are derived from xG models and they represent players' influence on shot conversion. In this section, the players are ranked separately based on their position types: (i) Forward, (ii) Midfielder, (iii) Center Back, and (iv) Full Back/Wing Back. Some of the players can be categorized into multiple positions. The EPI standard deviations, StatsBomb xG (SB xG), number of goals, and shots from open play are also provided for additional context in Table 6. It is important to note that these rankings do not suggest that one player is better than the other. Rather it provides a data-driven measure of how much impact a player has on shot conversion. A player with a higher EPI suggests that their baseline xG value is higher while holding other predictors constant. For example, Ben Chilwell's baseline xG is 0.03 model units higher than Reece James' baseline xG (see $\beta_0 + \delta_j$ in Section 3.5) for a given shot. It is **not reasonable** to infer that Chilwell is a better player than James using solely this analysis.

The model estimates show that the top five forwards and midfielders generally have a greater EPI than center backs and wing backs. This is highlighted for both the EPL and WSL. The top five players across all positions have a positive impact on the shot conversion. All of these players have scored more goals than their total SB xG. Therefore, the EPIs can be seen as the players' impact contributing to their xG overperformance. Applications of EPI to football are demonstrated in case studies in Appendix A6.

Table 6. The estimated player impacts (EPI) and their standard deviations (SD) are derived from the generalized linear mixed models (GLMM) developed for the English Premier League (EPL) and Women's Super League (WSL). "SB xG" refers to the players' StatsBomb xG measure from open play shots.

EPL Forward	EPI (SD)	Goals (Shots)	SB xG	WSL Forward	EPI (SD)	Goals (Shots)	SB xG
Heung-Min Son	0.52 (0.16)	38 (151)	25.86	Ji So-Yun	0.12 (0.13)	13 (105)	7.03
Gareth Bale	0.27 (0.22)	10 (30)	5.22	Vivianne Miedema	0.12 (0.11)	54 (276)	45.3
Marcus Rashford	0.23 (0.20)	15 (95)	11.59	Caroline Weir	0.10 (0.13)	16 (100)	10.7
Ilkay Gundogan	0.22 (0.19)	20 (101)	15.74	Chloe Kelly	0.07 (0.13)	17 (113)	13.36
Riyad Mahrez	0.22 (0.19)	16 (105)	11.45	Caitlin Foord	0.06 (0.14)	10 (36)	8.20

EPL Midfielder	EPI (SD)	Goals (Shots)	SB xG	WSL Midfielder	EPI (SD)	Goals (Shots)	SB xG
James Maddison	0.35 (0.21)	11 (57)	4.48	Jordan Nobbs	0.10 (0.13)	18 (94)	12.8
Kevin De Bruyne	0.30 (0.19)	19 (138)	12.6	Kim Little	0.10 (0.14)	12 (49)	6.7

Emile Smith Rowe	0.24 (0.21)	12 (53)	7.79	Fran Kirby	0.10 (0.13)	20 (111)	15.2
Bruno Fernandes	0.21 (0.18)	19 (175)	14.6	Georgia Stanway	0.08 (0.13)	21 (160)	17.2
Rodri	0.18 (0.21)	8 (76)	4.44	Keira Walsh	0.08 (0.14)	5 (30)	1.00

EPL Center Back	EPI	Goals (Shots)	SB xG	WSL Center Back	EPI (SD)	Goals (Shots)	SB xG
Kurt Zouma	0.19 (0.22)	6 (31)	2.12	Emma Mitchell	0.04 (0.14)	3 (16)	0.82
Thiago Silva	0.15 (0.23)	5 (31)	2.02	Maren Mjelde	0.03 (0.14)	3 (17)	1.53
Gabriel Magalhaes	0.15 (0.22)	7 (46)	3.73	Gemma Evans	0.03 (0.14)	2 (7)	0.45
Diego Llorente	0.10 (0.23)	4 (18)	2.11	Esme Morgan	0.02 (0.14)	2 (9)	0.56
Michael Keane	0.10 (0.23)	3 (16)	0.97	Aoife Mannion	0.02 (0.14)	1 (1)	0.02

EPL Wing Back	EPI	Goals (Shots)	SB xG	WSL Wing Back	EPI (SD)	Goals (Shots)	SB xG
Stuart Dallas	0.16 (0.21)	9 (79)	6.59	Caroline Weir	0.10 (0.13)	16 (100)	10.7
Ben Chilwell	0.13 (0.22)	6 (40)	3.60	Georgia Stanway	0.08 (0.13)	21 (160)	17.2
Reece James	0.10 (0.22)	6 (63)	3.70	Alisha Lehmann	0.06 (0.14)	10 (64)	7.87
Sergi Canos	0.09 (0.23)	3 (15)	1.40	Emma Mitchell	0.04 (0.14)	3 (16)	0.82
Matty Cash	0.08 (0.23)	3 (22)	1.37	Rachel Rowe	0.03 (0.14)	5 (36)	3.13

6 Discussion

The intricacies and inherent hierarchical structure of football event data is unpacked and discussed in detail in this paper. Due to the nature of football, observations are inherently nested under the player who creates the event. This data behavior has not been addressed previously in football analytics as most analyses seem to apply non-hierarchical models to the data [3, 5, 6, 16, 18]. In this research, a principled approach has been proposed that first assesses the data structure and then applies a statistically appropriate model for the data structure itself. The model presented in this research is a GLMM with just one random effect on the player. As a result, individual player's effects can be estimated. These estimates are then used in the calculation of a shot's xG that is adjusted for the player. Shots taken by Reece James and Ben Chillwell will now include the 0.10 and 0.13 values in their xG estimations respectively. These values represent their EPIs or contribution to the xG. Therefore, even if the shot characteristics are the same between them for any given shot, their xG values would still be different. So, the EPIs discussed in this paper can be interpreted as the influence each player has on shot conversion. It is interesting to also note that the standard deviations of the EPIs are not very small. Further research regarding variable selection for model development and larger sample sizes for each player could potentially lead to a decrease in the standard deviations.

This research generated xG models and EPIs for the EPL and WSL separately. The stratified analyses between the EPL and WSL provided interesting insights on shot characteristics that influence the xG in the two leagues. The distance of the shot-taker to the goal and number of defenders between the shot and the goal have a negative impact on xG in both EPL and WSL. A greater distance or more defenders result in a lower xG estimation. Alternatively, distance of the goalkeeper to the goal and the shot angle have a positive impact in both leagues. When the goalkeeper is farther from their goal or the shot is taken in front of the goal results in a higher xG estimation. These shot predictors seem to have a similar impact on xG regardless of the league. Furthermore, there were differences observed between the leagues. For example, the Lob shot seems to have a significant positive association with goal odds in the WSL. Whereas, this shot was deemed non-significant for the EPL by the GLMM. However, the WSL models were trained using older seasons and this pattern may no longer hold true in the latest seasons with the rapid development of the women's game [16, 20]. Similarly, being under pressure seemed to have a significant negative impact on EPL shot-takers, but there was not enough evidence in the WSL to suggest such an effect. These differences highlight the unique characteristics of men's and women's football and provide evidence that commercial xG models should be stratified for the men's and women's leagues. Further research regarding models for different leagues across Europe could potentially provide additional insights on the differences between European leagues. Scouting departments can then leverage these analyses to scout players from other leagues.

The results from the EPL and WSL analyses provide meaningful insights on the players' contributions to their team's success. A positive EPI suggests that a player increases the xG of a shot. This paper shares the top five rankings for EPI between the two leagues for several position groups. In general, Heung-Min Son was estimated to have the highest impact on goal conversion in the EPL. Over the years, he has regularly over-performed his player xG [21]. In the 2021/22 season, he took fewer shots than Mohamed Salah (86 vs. 123 shots, respectively) with whom he shared the golden boot (23 goals) [21, 22]. Son's high EPI suggests that he has a unique skill that enables him to score many goals from fewer shots as opposed to other players. This unique skill is left unmeasured by xG models that do not adjust for the players. In the WSL analysis, the EPI measures are generally lower than the EPL. This could be because the shot characteristics in the model explain the variability in xG more in the WSL. Another explanation could be that most players in the WSL are similar to each other in terms of shot conversion. However, the sample size of shots for the WSL is much lower than the EPL. Additional WSL data could lead to more differences between the players. It is interesting to note that Vivianne Miedema and Ji So-Yun have the same EPI but Miedema has scored more goals. Both So-Yun and Miedema have performed better than what was expected from their xG. However, Miedema took more shots and scored the most goals in the WSL data. Whereas, So-Yun scored from fewer shots and the model identifies her to have a similar impact on shot-conversion as Miedema.

The applications of this research revolve around xG. However, it should be noted that GLMMs can be applied to almost any continuous, count, and even multi-class target variables. As a result, these models can be applied to derive the player impacts for other advanced models such as "PSxG" [23], "xA" [13], and "xT" [14]. These player effects for different metrics can be used to scout multiple facets of a player rather than just their xG. Additionally, there are other ways to frame a GLMM with appropriate justifications. For example, some of the predictors in the model can be modeled using varying-slopes: estimating a predictor effect that changes for each player. This idea can be applied to the demo in Appendix A1. It can also be argued that there is yet another level of hierarchy in the data: players nested under their respective teams. Teams are repeated in the data and the GLMM can be further extended for team random effects. The GLMM is a frequentist approach to analyzing hierarchical data. There are also Bayesian alternatives that can be explored and may offer more modeling flexibility.

A final interesting discussion to note is that the objective of an xG model is not to build the **best predictive model**; but rather to calculate the probability of a goal based on pre-shot predictors. By changing one of the pre-shot predictor measures, an analyst can calculate the change in xG. The "Post-Shot xG" (PSxG) [23] or "Expected Goal-on-Target" (xGOT) [24] metrics theoretically suggest that adding post-shot characteristics can increase the predictive performance of an xG model. However, including post-shot characteristics would change the interpretation of an xG model to a PSxG model. In this paper, the modeling framework provided a principled and statistically-grounded approach to developing an xG model based on the data at hand rather than focusing on predictive performance. However, the paper also provided empirical experiments that demonstrate its predictive ability in comparison to its counterparts in Section 4. These results are not all too surprising because the predictors were engineered in a manner that allows linear relationships to be modeled. It is important, however, to note that the train-test splits do have low sample sizes. A different dataset could lead to different types of performance results. Additionally, incorporating more predictors can affect the performances of each of the experiment models. For example, a non-linear relationship between a new predictor and the target can hinder the linear models. An analyst could also technically keep fine-tuning an XGBoost until desirable test results are observed.

In conclusion, this paper applied hierarchical statistical models to StatsBomb event data to demonstrate how they can be leveraged for football analytics. The models can be used to develop xG models, adjust for players to estimate player xG for each shot, and finally to provide interpretable insights about the game of football. The analysis demonstrated in this research can enable football researchers to extend their work to also draw insights about the players in their data and their advanced metric models.

Acknowledgements

We would like to thank StatsBomb for providing the data and for giving us the platform to share our football research at the 2022 StatsBomb Conference in London, United Kingdom. We would also like to thank the following people for their feedback and support during the research and conference preparation process.

Matthew Zawistowski, Clinical Associate Professor of Biostatistics, Department of Biostatistics, University of Michigan

Will Morgan, Senior Data Scientist, StatsBomb, United Kingdom

Katie Slade, Events Manager, StatsBomb, United Kingdom

References

[1] C. Biermann, *Football Hackers: The Science and Art of a Data Revolution*. Blink Publishing, 2019. https://books.google.com/books?id=OjMWwQEACAAJ

[2] J. Tippett, *The Expected Goals Philosophy: A Game-Changing Way of Analysing Football*. Spain: Independently Published, 2019.

[3] T. Decroos, J. Davis. “Interpretable Prediction of Goals in Soccer,” presented at StatsBomb Innovation in Football Conference, 2019

[4] M. Van Roy, P. Robberechts, J. Davis, W.-C. Yang, and L. De Raedt, “Leaving Goals on the Pitch: Evaluating Decision Making in Soccer,” presented at the 15th MIT Sloan Sports Analytics Conference, April 08-09, 2021

[5] J. Whitmore, “What are expected goals (XG)?,” *The Analyst*, 07-Oct-2021. [Online]. Available: https://theanalyst.com/na/2021/07/what-are-expected-goals-xg/. [Accessed: 12-Sep-2022].

[6] G. Anzer and P. Bauer, “A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer), ” *Front. Sports Act. Living*, March 29, 2021

[7] D. Vatavani, “Upgrading Expected Goals,” *StatsBomb*, 16-May-2022. [Online]. Available: https://statsbomb.com/articles/soccer/upgrading-expected-goals/. [Accessed: 12-Sep-2022]

[8] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press, 2006.

[9] M. Mumper, “Multilevel modelling,” *American Psychological Association*. [Online]. Available: https://www.apa.org/science/about/psa/2017/01/multilevel-modelling. [Accessed: 12-Sep-2022].

[10] A. Field and D. Wright, “A Primer on Using Multilevel Models in Clinical and Experimental Psychopathology Research, ” *Journal of Experimental Psychopathology*, January, 2011

[11] X. Harrison, L. Donaldson, M. Correa-Cano, J. Evans, D. Fisher, C. Goodwin, B. Robinson, and D. Hodgson, and R. Inger, “A brief introduction to mixed effects modelling and multi-model inference in ecology,” *PeerJ,* May 23, 2021

[12] S. Olthof, T. Tureen, L. Tran, B. Brennan, B. Winograd, and R. Zernicke, “*Biomechanical Loads and Their Effects on Player Performance in NCAA D-I Male Basketball Games,” Frontiers in Sports and Active Living,* December 12, 2021

[13] T. Worville, “Expected assists in context,” *Stats Perform*, 28-Apr-2020. [Online]. Available: https://www.statsperform.com/resource/expected-assists-in-context-2/. [Accessed: 12-Sep-2022].

[14] K. Singh, “Introducing Expected Threat (xT),” [Online]. Available: https://karun.in/blog/expected-threat.html. [Accessed: 12-Sep-2022]

[15] T. Lawrence, “Introducing xGChain and xGBuildup,” *StatsBomb*, 30-August-2018. [Online]. Available: https://statsbomb.com/articles/soccer/introducing-xgchain-and-xgbuildup/. [Accessed: 12-Sep-2022]

[16] J. Davis and P. Robberechts, “How data availability affects the ability to learn good xG models,” 12-May-2020. [Online]. Available: https://dtai.cs.kuleuven.be/sports/blog/how-data-availability-affects-the-ability-to-learn-good-xg-models. [Accessed: 12-Sep-2022]

[17] J. Yorke, “Pass Footedness in the Premier League,” *StatsBomb*, 26-April-2019. [Online]. Available: https://statsbomb.com/news/statsbomb-release-expected-goals-with-shot-impact-height/. [Accessed: 12-Sep-2022]

[18] A. Fairchild, K. Pelechrinis, and M. Kokkodis, “Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality,” *Journal of Sports Analytics,* January 1, 2018

[19] T. Knutson, “StatsBomb Release Expected Goals with Shot Impact Height,” *StatsBomb*, 31-July-2020. [Online] Available: https://statsbomb.com/news/statsbomb-release-expected-goals-with-shot-impact-height/. [Accessed: 12-Sep-2022]

[20] “Physical Analysis of France 2019 shows increase in speed and intensity,” *FIFA*, 6-July-2020. Available: https://www.fifa.com/tournaments/womens/womensworldcup/france2019/news/physical-analysis-of-france-2019-shows-increase-in-speed-and-intensity. [Accessed: 12-Sep-2022]

[21] “Son Heung-min” *FBref***.** [Online] Available: https://fbref.com/en/players/92e7e919/Son-Heung-min. [Accessed: 12-Sep-2022]

[22] "Mohamed Salah" *FBref*. [Online] Available: https://fbref.com/en/players/e342ad68/Mohamed-Salah. [Accessed: 12-Sep-2022]

[23] M. Goodman, "A New Way to Measure Keepers' Shot Stopping: Post-Shot Expected Goals," *StatsBomb*, 12-November-2018. Available: https://statsbomb.com/articles/soccer/a-new-way-to-measure-keepers-shot-stopping-post-shot-expected-goals/. [Accessed: 12-Sep-2022]

[24] J. Whitmore, "Introducing Expected Goals On Target (xGOT)," *StatsPerform*, 24-September-2019. Available: https://statsbomb.com/articles/soccer/a-new-way-to-measure-keepers-shot-stopping-post-shot-expected-goals/. [Accessed: 12-Sep-2022]

Appendix

A1 Missed Inferences for Repeated Measures Data

Toy data is used here to demonstrate how models can miss certain insights or provide biased inferences if repeated measures are ignored. Figure A1 shows three separate plots using the same data points. The x-axis represents the age of eight players and the y-axis represents their goal scored for two separate seasons.

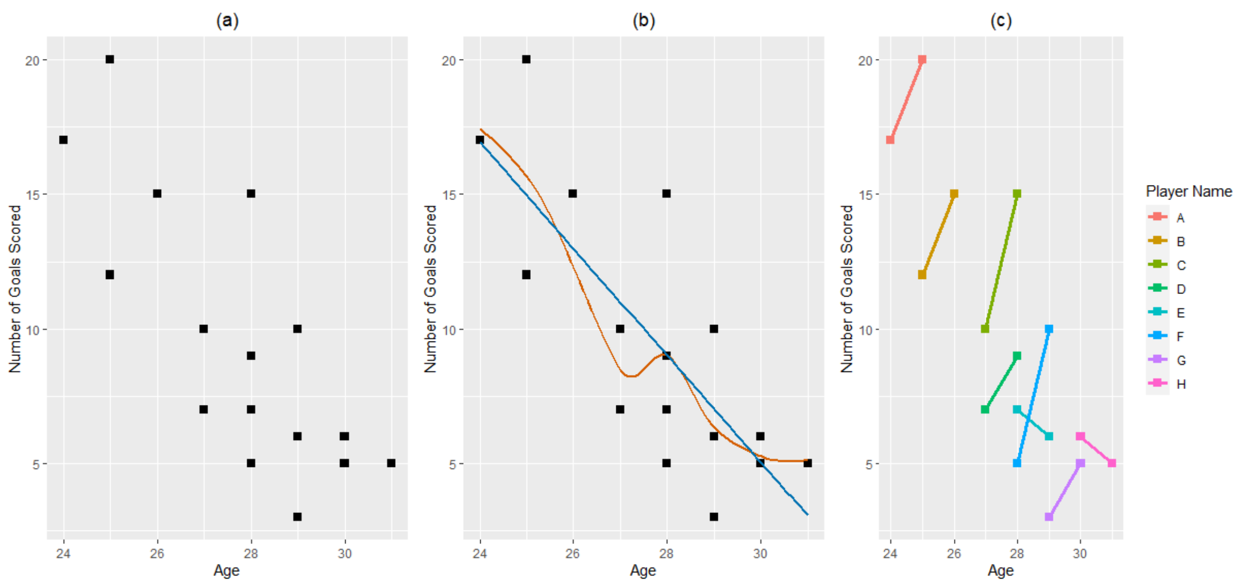


Figure A1. Illustration of how an analysis can miss out on additional insights when repeated measures are ignored. This data represents the number of goals scored by eight different players over a two year period.

In **plot A**, the analysis ignores information about the players' repeated measures. Through simple observation, it appears that the number of goals scored decreases as the players age. **Plot B** demonstrates how a non-hierarchical linear (blue) and a non-linear model (orange) would be applied to the data. In this analysis, the models would also suggest that there is a negative correlation between age and goals. However, by color-coding the data for players, **plot C** suggests a different interpretation. For players younger than 28 years, there seems to be a positive correlation. Whereas, players who are older than 28 do not all have the same directional patterns. Based on this exploratory analysis, we can see that the baseline is not the same across all players (as some are younger than others) and that age affects their goal scoring ability differently. The decreasing pattern in plot A is not generalizable to the players in this data because of player repetition. This example demonstrates how repeated measures can affect analysis if repeated measures are not accounted for. Additionally, the toy example deals with only one predictor; the more predictors in the analysis, the more convoluted these visualizations can get.

A2 XGBoost Parameters

The XGBoost models differ due to the number of boosting iterations and learning rates. All of these models are specified with a binary logistic objective function. Validation metrics, binary classification error and negative log-likelihood, were monitored during the model training. The remaining parameters are set to the default as specified by the xgboost function in the **xgboost** package (version 1.4.1.1) in R

Specification (Parameter)	XGBoost 10	XGBoost 25	XGBoost 50	XGBoost 100
Boosting Iterations (nrounds)	10	25	50	100
Learning Rate (eta)	0.30	0.10	0.10	0.10

A3 Visualization of Predictor Effects

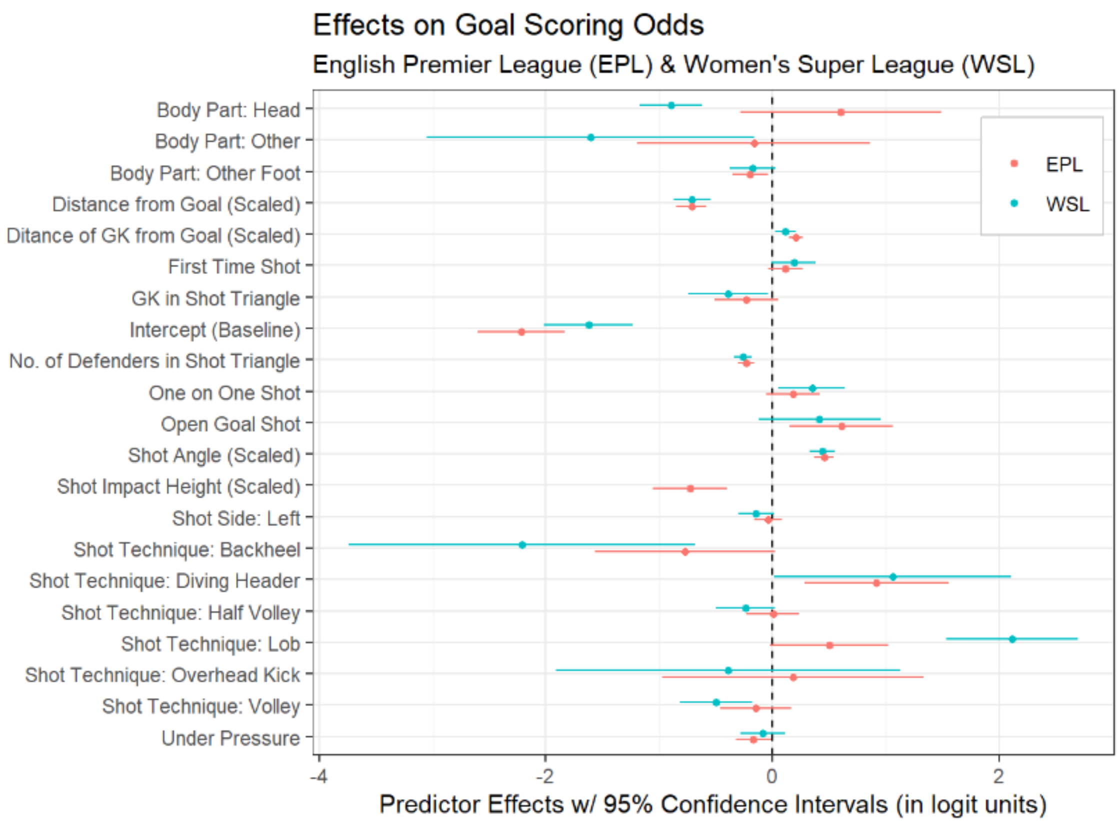


Figure A5. The log odds estimates and their respective 95% confidence intervals (CI) from the GLMM. CIs that **do not** cross the 0 vertical line are considered statistically significant results. Positive log odds represent a positive association between predictor and target. Negative log odds represent a negative association between predictor and target.

A4 Football Applications

Note: Video walkthroughs of these case studies can be accessed via the StatsBomb website. Please zoom in on the images for a more clear view of the numbers and visuals.

Study #1: Heung-Min Son and Riyad Mahrez

In this hypothetical case study, assume that analysts at Manchester City and Tottenham Hotspur can retrospectively assess the impact a new signing could have had on their previous matches. For example, the xG model is fed in Riyad Mahrez's EPI to back-estimate the xG for all of Son's shots at Tottenham. Similarly, Son's EPI is fed into the model to back-estimate the xG for Mahrez's shots at City.

1 Retrospective Analysis of Player Transfers

For demo purposes, assume that we are in a crazy parallel universe

- Manchester City acquired Heung-Min Son from Tottenham for an undisclosed fee plus Riyad Mahrez going the other way
- The analytics departments at both clubs can retrospectively assess how their new signings would have impacted their xG performances in the previous seasons

$$g(\pi_{ij}) = \beta_0 + \beta_1 x_{ij,1} + \beta_2 x_{ij,2} + \dots + \beta_p x_{ij,p} \boxed{+ \delta_j}$$

Equation 2

EPL Forward	EPI
Heung-Min Son	0.52
Riyad Mahrez	0.22

Conference 2022

1 Retrospective Analysis of Player Transfers

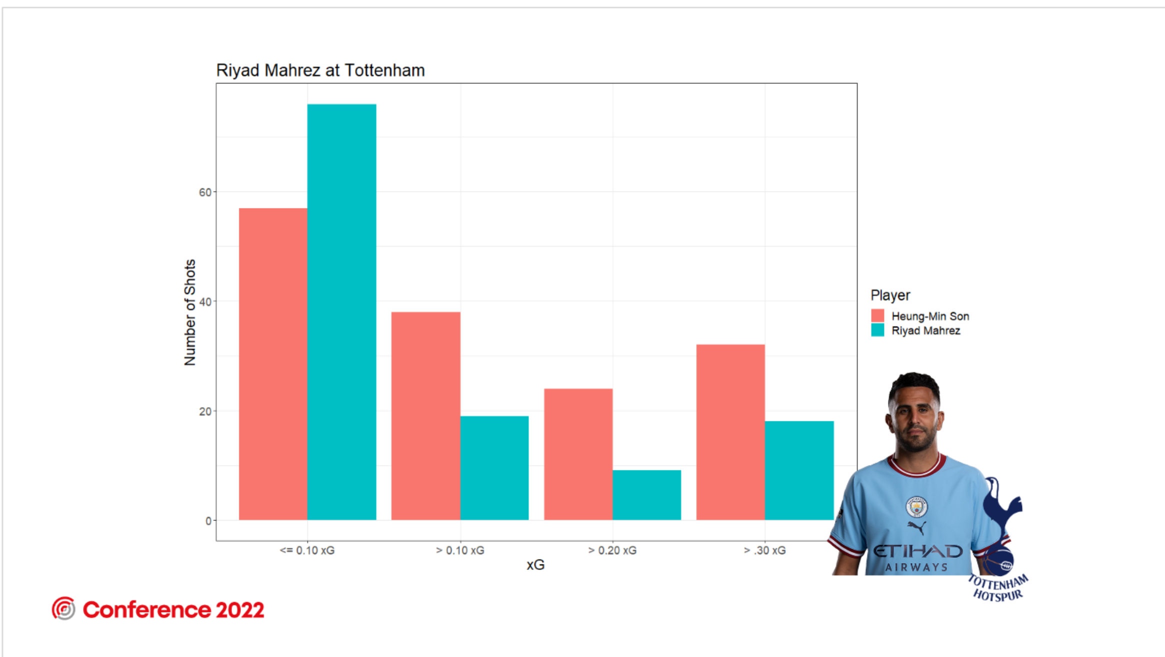
EPL Forward	Total GLMM xG	xG per Shot
Heung-Min Son	28.8	0.19
Riyad Mahrez	23.4	0.16

Number of Open-Play shots taken by **Son** for **Tottenham**: **151**

EPL Forward	Total GLMM xG	xG per Shot
Riyad Mahrez	12.11	0.12
Heung-Min Son	15.37	0.15

Number of Open-Play shots taken by **Mahrez** for **Manchester City**: **105**

Conference 2022



In this particular case, signing Mahrez would not bring much xG value to Tottenham if they were to lose Son in a transfer window. Whereas, Son would bring a general improvement in xG for City if they sold Mahrez. This same analysis can be applied to other players to evaluate potential transfers.

Study #2: Scouting Emma Hayes's Shot Takers

Analysts can use these player adjusted models to perform opposition xG scouting. In this scenario, Manchester City WFC can assess the EPIs for their WSL rivals, Chelsea WFC, prior to a match. This allows analysts to identify which Chelsea players have the highest goal threat and provide complementary analyses for the coaching staff to consider. The analysts can also generate player-adjusted xG maps from previous City-Chelsea matches to evaluate how City's defense has performed historically against Chelsea.

2 Opposition Scouting

Assume you are an analyst at Manchester City WFC

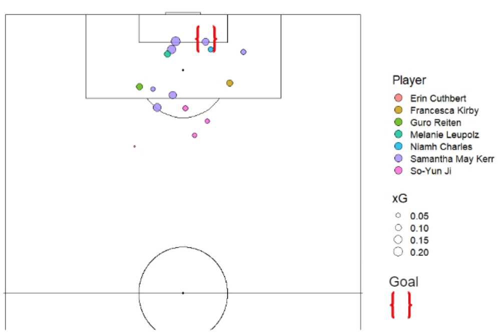
- You are tasked with the challenge of defending against Emma Hayes' Chelsea
- The Player Adjusted xG Models allows for player specific opposition xG scouting

Goal Threat

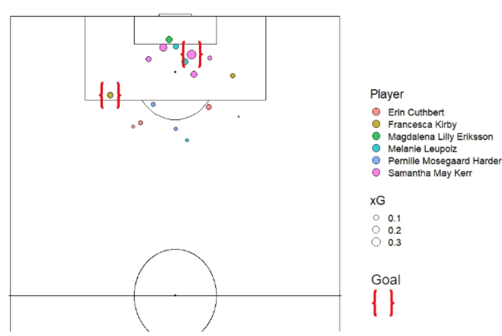
Chelsea Forwards	EPI	Goals (Shots)	GLMM xG
Ji So-Yun	0.12	13 (105)	7.1
Fran Kirby	0.10	20 (111)	15.8
Sam Kerr	0.04	22 (98)	19.7
Melanie Leupolz	0.04	4 (24)	2.2
Bethany England	0.03	29 (163)	27.6
Pernille Harder	0.03	8 (62)	6.6
Guro Reiten	-0.001	6 (55)	6
Erin Cuthbert	-0.02	9 (109)	10.1

Conference 2022

2 Opposition Scouting



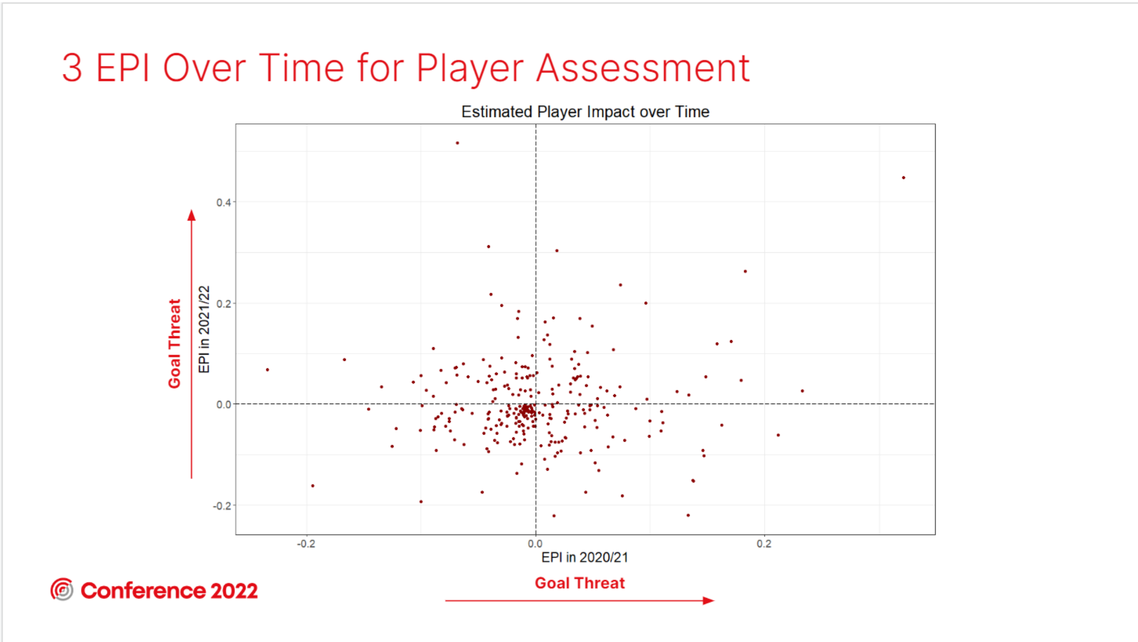
Man City 2 (1.96 xG) vs **Chelsea 2 (1.44 xG)**
1 Penalty

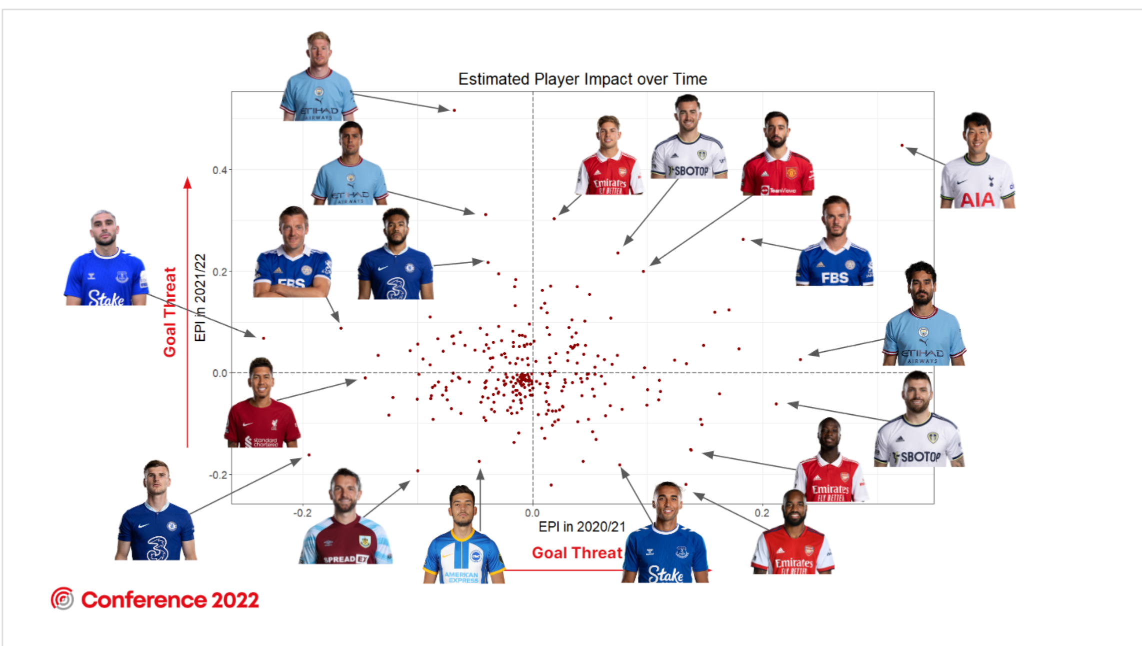


Chelsea 3 (2.07 xG) vs **Man City 1 (0.68 xG)**
1 Penalty 1 Penalty

Conference 2022

Study #3: Player Assessments over Time





Analysts can also develop stratified models for each season and track the differences in EPI over time for each of their players. This type of analysis enables clubs to evaluate their players over time and see if they have improved or declined in the previous seasons in terms of their EPI or contribution to xG. The graphs above (from the conference presentation) show x-y plots where each data point represents a player. The top right quadrant are players who had positive EPIs in both seasons (20/21 and 21/22). These are players who contribute positively to shot conversion. Heung-min Son is the stand out player once again. The bottom left quadrant are players who had negative EPIs in both seasons. These are players who have negatively contributed to their shot conversion. The bottom right quadrant are players who positively contributed in 20/21 but had a negative impact in 21/22. These players can be viewed as individuals who have declined in terms of xG impact over the course of these two seasons. The top left quadrant are players who negatively contributed to xG in the earlier season and positively contributed to the later season. These players can be interpreted as individuals who have improved their xG contribution. Clubs can use these analyses on top of other contextual information such as injuries, playing time, and international duties to assist in contract decisions regarding a player’s future.