On the time of corner kicks in soccer: an analysis of event history data

K. Ken Peng, X. Joan Hu and Tim B. Swartz *

Abstract

To understand the patterns of times to corner kicks in soccer and how they are associated with a few important factors, we analyze the corner kick records from the 2019 regular season of the Chinese Super League. This paper is particularly concerned with the elapsed time to a corner kick from a natural starting point. We overcome two challenges arising from such time-to-event analyses, which have not been discussed in the sports analytics literature. The first is that observations of times to corner kicks are subject to right-censoring. A given soccer starting point rarely ends with a corner kick but the occurrence of a different terminal event. The second issue is the mixture feature of short and typical gap times to the next corner kick from a particular one. There is often a subsequent corner kick quickly following a corner kick. The conventional event time models are thus inappropriate for formulating distributions of corner kick times. Our analysis reveals how the timing of corner kicks is associated with the factors of first versus second half of the game, home versus away team, score differential, betting odds prior to the game, and red card differential. We present applications of the developed statistical model for prediction to support tactics and sports betting.

Keywords: EM algorithm, factor effects, mixture distributions, predictive model, right-censored event times.

*K. Ken Peng is a PhD candidate, X. Joan Hu, and Tim B. Swartz are Professors of Statistics in Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Corresponding author: Tim B. Swartz (tim@stat.sfu.ca). The authors thank two anonymous reviewers whose commments helped improved the paper.

1 INTRODUCTION

In association football (i.e., soccer), corner kicks are of great interest since they can lead to goals. The fact that corner kicks are set pieces allows that attacking teams have time to prepare and strategize on the execution of corner kicks (e.g., depth and direction of the kick). Conversely, opposition teams also strategize on how to best defend corner kicks (e.g., man-to-man marking versus zonal marking). Because of the ability to strategize on corner kicks, it is tactically helpful to better understand the nature of corner kicks.

There has been a considerable literature on corner kicks. For example, **?** consider 124 matches and investigate the characteristics of corner kicks including the types of corner kicks, the efficacy with which they lead to shots and goals, and the identification of variables that are related to corner kick success. **?** provide team strength profiles using various performance variables including shots resulting from corner kicks. **?** analyze the evolution of corner kicks in terms of execution and tactics based on 192 matches from three successive World Cups. **?** survey the use of statistical methods across major sports.

This paper investigates the time elapsed since a natural starting point until a corner kick using event history data analysis. The major area of applications of event history data analysis are in biomedical research where many such analysis methods have been developed to accommodate different scenarios. In soccer, previous research has delved into analyzing time to events through survival analysis. For example, **?** and **?** utilized Cox-type models to explore the correlation between the timing of the first goal and that of the second goal during a match. **?** employed the inverse Gaussian hazard model to examine the significant determinants influencing the time to the first player substitutions within each team. A feature of time-to-event data is that not necessarily all events occur before the end of the study. In the corner kick application, a match or its first half does not typically end with a corner kick; in this case, the time to a potential corner kick must be longer than the elapsed time since the relevant starting point. This situation is referred to as right-censoring. **?**, for example, provide a comprehensive summary of the basic knowledge, and how commonly used methods for event time data analysis handle right-censored observations.

The second complication in the analysis of corner kicks is that there is a nonignorable probability of a subsequent corner kick shortly following a corner kick. This is often the consequence of the opposite team defending the corner kick. The conventional event time distributions are not suitable to describe the timing of corner kicks in general since they do not accommodate the aforementioned clustering feature. We employ a mixture

distribution model to formulate the phenomenon; see, for example, **?** for a review of the theory and methodological developments of finite mixture models. There are other temporal phenomena which may exhibit similar clustering features. For example, the time until an earthquake occurs is known to have clustering patterns due to occurrence of aftershocks (**?**). For another example, the recurrence of particular diseases such as seizures of people with epilepsy may result in clustering (**?**). We anticipate that the approach developed in this paper may be applied in other scientific domains.

An important research issue in sports analytics is to identify important factors and evaluate their impact on the timing of corner kicks. By maximum likelihood estimation with right-censored event times, we explore how times to corner kicks are associated with conventional factors/exposures. The records of the Chinese Super League (CSL) matches in the 2019 season are used for illustration. The final model developed from the analysis is then employed to conduct a simulation study, to showcase its applications.

The rest of this paper is organized as follows. In Section 2, we describe the aforementioned CSL data, and discuss about the right-censoring mechanism and a potential mixture distribution of the recorded corner kicks. Section 3 focuses on the time elapsed until a corner kick occurs since a natural starting point. We frame the available corner kick data as right-censored event times and present procedures for estimating the model parameters. In Section 4, the EM algorithm is applied to calculate parameter estimates and then select important factors to fit the proposed model via the CSL data. Applications of the model are illustrated by simulation. We conclude this article with a short discussion in Section 5.

2 THE 2019 CSL GAMES

Despite not being one of the top five soccer leagues in the world, the CSL is a high level professional league that attracts international star players. For example, the former Belgian international Marouane Fellaini and former Brazilian international Paulinho both currently (2023) play in the the CSL. It is believed that there are similarities between the style of play in the CSL and other top soccer leagues. A sample of recent papers that have investigated soccer analytics problems using data from the CSL include **?**, **?**, **?**, **?**, and **?**.

The 2019 CSL season involved 16 teams where each team played every opponent twice, once at home and once away. Among the potential 240 matches, there were 7 missing match records which we suspect were due to issues with the video equipment. We assume that these matches were missing completely at random. This paper is concerned with the

"event" data from the match records, with particular attention to the corner kicks. Soccer event data typically include 46 different soccer-related occurrences such as corner kicks, tackles, and passes. These events were recorded along with auxiliary information such as time stamps and the players involved whenever an "event" takes place. The available CSL match data include 2314 corner kicks with an average of $2314/233 = 9.93$ corner kicks per match. The event information is manually recorded by technicians who view the match recordings. The data are more comprehensive than the standard box scores, which only provide the total number of corner kicks taken by each team during a match.

Figure 1 provides a snapshot of the corner kick event data: 20 out of the $240 - 7 = 233$ games are randomly selected and home teams' corner kicks are shown. The records of each team produces a point stochastic process, where the dots in the plot are corner kicks. One can observe the aforementioned phenomenon that a subsequent corner kick might immediately follow a corner kick. This can result in a clustering of multiple corner kicks, and thus the gap times between two such consecutive corner kicks are short. The stars in the plot represent the times when a team scored a goal; the triangles denote the ends of the first half. Both scoring and the end of the half can be viewed as "terminal events" because both events stop the game and lead to a restart at the middle of the field. This causes the timing of corner kicks to restart. The dashed line indicates when the match time reached 45 minutes. In a soccer game, the referee can add extra time onto the end of a half because of multiple pauses. To conduct our analysis, we define two types of corner kick times: Type 1 represents the duration from a non-corner kick starting event to a corner kick, while Type 2 represents the time elapsed from one corner kick to the next without interruption. Both types of corner kick times may be subject to right censoring by a terminal event following a corner kick. Further elaboration on the formulation of corner kick times will be presented in Section 3.1. In the remainder of this section, we focus on the available observations on times to corner kicks, which are right-censored event times, and formulate via mixture distribution models the short and long gap times.

2.1 Right-censored times to corner kicks

Right censoring happens in our application when a terminal event such as a goal or the ending of a half game occurs before the next corner kick. The event time of interest, the time to corner kick, is then only known to exceed the occurrence time of the terminal event. Ignoring the right censoring would cause underestimation of the true distribution of

the time to corner kick. The matches in our dataset had 4656 starting points, where 2314 and 2342 of them ended by recorded corner kicks and terminal events, respectively. That led to 2314 observed times to corner kicks, and 2342 right-censored corner kick times. It is clearly important to account for the right censoring in the analysis of the corner kick times with such a heavy right censoring. We note that this insight has not been recognized in the sports analytics literature. In the Section 1 of supplementary file, we provide analyses to demonstrate the importance of addressing the right censoring when analyzing corner kicks.

2.2 Mixture of long and short gap times

Another issue with the timing of corner kicks is that the waiting time to the next corner kick can be rather short when a corner kick is quickly followed by another corner kick. In the CSL data, the average waiting time of a corner kick is around 20 minutes, but there are around 300 corner kick waiting times less than one minute. That is, gap times between corner kicks are a mixture of two subgroups, namely the "long gap times" and the"short gap times".

We consider a mixture distribution model for gap times between consecutive corner kicks. The formulation includes the following advantages. Firstly, it captures the special feature of corner kicks that multiple corner kicks may happen within a short time period, and thus improves prediction based on the fitted model. In addition, it allows us to identify the important factors regarding the probability of a gap time to be long or short through evaluating the covariate effects. Note that we do not fully know within certain time after the starting point whether a gap time belongs to the short or long time subgroup.

In the next section, after introducing necessary notation for formulating corner kick times, we develop a mixture distribution model based on maximum likelihood estimation with right-censored event times.

3 MODEL FORMULATION AND ESTIMATION

3.1 Notation

Let T_k be the time of the kth event associated with a team in a game and Δ_k be the indicator for whether the kth event is a corner kick, where $\Delta_k = 0$ indicates the event is a terminal event such as scoring or the end of a half. Denote the time to the kth event since

the previous one by $Y_k = T_k - T_{k-1}$ for $k \geq 1$. Thus Δ_k indicates whether Y_k is an observed time to a corner kick from a starting point or the censoring time, and Δ_{k-1} tells whether Y_k is the gap time of two consecutive corner kicks or the time to the first corner kick after a game restart. We refer it to as a Type 1 time if Y_k is with $\Delta_k = 1$ and $\Delta_{k-1} = 0$, the time to the first corner kick from a terminal event; a Type 2 time if Y_k is with $\Delta_k = 1$ and $\Delta_{k-1} = 1$, the gap time of two consecutive corner kicks. See the graphical explanation in Figure 2.

Denote the observations on (T_k, Δ_k) and Y_k associated with team i at game j by (t_{ijk}, δ_{ijk}) and y_{ijk} for $k = 0, \ldots, n_{ij}$ with n_{ij} the number of the events, $j = 1, \ldots, J_i$, and $i = 1, \ldots, I$. Let X_l be a factor of interest for $l = 1, \ldots, L$. Denoted the corresponding observations on X_l by x_{lijk} for $l = 1, \ldots, 5$ associated with the kth event of team i at game j.

We further introduce η_k to indicate whether Y_k is a long gap time for $k = 1, \ldots$, the random variable η_k is in general not observable and only defined for observed Type 2 times to corner kicks. Let η_{ijk} be the indicator associated with Y_{ijk} when $\delta_{ijk} = \delta_{ij,(k-1)} = 1$.

3.2 Statistical modeling

The most popular model family in event time analysis is the Cox proportional hazards model. It assumes the conditional hazard function of an event time given covariates $\boldsymbol{Z} = \boldsymbol{z}$ is $h(t|\boldsymbol{z}, \boldsymbol{\theta}) = h_0(t)\exp(\boldsymbol{z}'\boldsymbol{\beta})$ for $t > 0$, where $h_0(\cdot)$ is the baseline hazard and $\exp(\boldsymbol{z}'\boldsymbol{\beta})$ describes the covariate effects (**?**). The corresponding conditional survivor function and density function are $S(t|\boldsymbol{z};\boldsymbol{\theta}) = \exp(-\int_0^t h(u|\boldsymbol{z};\boldsymbol{\theta})du)$, and $f(t|\boldsymbol{z};\boldsymbol{\theta}) = h(t|z;\boldsymbol{\theta})S(t|\boldsymbol{z};\boldsymbol{\theta})$, respectively.

Specifying the baseline hazard $h_0(t)$ into a parametric form, the model is then a parametric proportional hazard model. This paper considers parametric proportional hazards models for each type of the corner kick times. We assume that a Type 1 corner kick time (i.e. Y_k with $\Delta_k = 1$ and $\Delta_{k-1} = 0$) follows a 2-parameter Weibull distribution conditional on the covariates $\boldsymbol{Z}$ with the conditional hazard function

$$h_1(y|\boldsymbol{Z} = \boldsymbol{z}; \boldsymbol{\beta}_1) = \gamma_1\lambda_1(\lambda_1 y)^{\gamma_1 - 1}e^{\boldsymbol{z}'\boldsymbol{\beta}_1}, \tag{1}$$

where $\boldsymbol{\theta}_1 = (\lambda_1, \gamma_1, \boldsymbol{\beta}_1)$ with $\boldsymbol{\beta}_1$ a vector with the same dimension as $\boldsymbol{z}$, and λ_1 and γ_1 are the Weibull rate and shape parameters. The covariate $\boldsymbol{Z}$ is a vector with components selected from the ones of $\boldsymbol{X} = (X_1, \ldots, X_5)'$, the five potential important factors listed in Table 1.

The variables are defined as follows.

- X_1 is an indicator variable denoting whether the corner kick occurred in the first or second half of the game.
- X_2 is an indicator variable indicating whether the home or away team executed the corner kick.
- X_3 is a numerical variable representing the current score difference of the team taking the corner kick, where a positive value indicates a lead and a negative value indicates a deficit.
- X_4 is a numerical variable reflecting the current difference in red cards, with $X_4 > 0$ denoting more players for the team taking the corner kick.
- X_5 is a numerical variable representing the European odds of the team executing the corner kick. These odds, fixed at the beginning of the game, are considered a proxy for team strength, encompassing factors such as home/away status, current form, injuries, player rotation, rest, and more, relevant to sports bettors.

The Weibull distribution was selected for our analysis based on a preliminary comparison of analyses using the exponential, Weibull, and lognormal distributions, as well as a non-parametric analysis by the Kaplan-Meier estimator. The consideration of long and short gap times between two consecutive corner kicks motivates us to assume Type 2 corner kick times follows a mixture distribution. Let $\eta_k = 1$ or 0 represent the gap time Y_k with $\Delta_k = 1$ and $\Delta_{k-1} = 1$ to be long or short, named by Type 2-L or Type 2-S, respectively. The conditional distributions of $[Y_k | \eta_k = 1, \boldsymbol{z}_L]$ and $[Y_k | \eta_k = 0, \boldsymbol{z}_S]$ are Weibull distributions with the following hazard functions:

$$h_2^L(y|\boldsymbol{z}_L;\boldsymbol{\beta}_2^L) = \gamma_2^L \lambda_2^L (\lambda_2^L y)^{\gamma_2^L - 1} e^{\boldsymbol{z}_L' \boldsymbol{\beta}_2^L}, \quad \text{and} \tag{2}$$

$$h_2^S(y|\boldsymbol{z}_S;\boldsymbol{\beta}_2^S) = \gamma_2^S \lambda_2^S (\lambda_2^S y)^{\gamma_2^S - 1} e^{\boldsymbol{z}_S' \boldsymbol{\beta}_2^S}. \tag{3}$$

Denote the parameters of Model 2-L (2) and Model 2-S (3) by $\boldsymbol{\theta}_2^L = (\lambda_2^L, \gamma_2^L, \boldsymbol{\beta}_2^L)$ and $\boldsymbol{\theta}_2^S = (\lambda_2^S, \gamma_2^S, \boldsymbol{\beta}_2^S)$, where $\boldsymbol{\beta}_2^L$ and $\boldsymbol{\beta}_2^S$ are vectors with the dimensions corresponding to $\boldsymbol{z}_L$ and $\boldsymbol{z}_S$. Further, denote the conditional density and survivor functions of Type 2-L and Type 2-S corner kick times by $f_2^L(\cdot|\boldsymbol{z}_L;\boldsymbol{\theta}_2^L)$ and $S_2^L(\cdot|\boldsymbol{z}_L;\boldsymbol{\theta}_2^L)$, and $f_2^S(\cdot|\boldsymbol{z}_S;\boldsymbol{\theta}_2^S)$, and $S_2^S(\cdot|\boldsymbol{z}_S;\boldsymbol{\theta}_2^S)$, respectively.

The indicator η_k is in general unobservable. We assume it follows a logistic regression model conditional on covariates $\boldsymbol{Z}_\eta$:

$$\pi(\boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha}) = P(\eta_k = 1|\boldsymbol{Z}_\eta = \boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha}) = \frac{1}{1+e^{-(\alpha_0+\boldsymbol{z}_\eta'\boldsymbol{\alpha})}}, \tag{4}$$

where the components of $\boldsymbol{\alpha}$ are the regression coefficients to the components of the selected factors. Apparently values of $\pi(\boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha})$ can be interesting in the corner kick application.

The models given in (2), (3) and (4) yield a mixture distribution model for Type 2 corner kick times with parameters $\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_2^L, \boldsymbol{\theta}_2^S, \alpha_0, \boldsymbol{\alpha})$. For $y > 0$ and $\boldsymbol{z}$ including all the components of $\boldsymbol{z}_L$, $\boldsymbol{z}_S$ and $\boldsymbol{z}_\eta$, its conditional survivor function and density function are

$$S_2(y|\boldsymbol{z};\boldsymbol{\theta}_2) = \pi(\boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha})S_2^L(y|\boldsymbol{z}_L;\boldsymbol{\theta}_2^L) + \left[1-\pi(\boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha})\right]S_2^S(y|\boldsymbol{z}_S;\boldsymbol{\theta}_2^S) \quad \text{and} \tag{5}$$

$$f_2(y|\boldsymbol{z};\boldsymbol{\theta}_2) = \pi(\boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha})f_2^L(y|\boldsymbol{z}_L;\boldsymbol{\theta}_2^L) + \left[1-\pi(\boldsymbol{z}_\eta;\alpha_0,\boldsymbol{\alpha})\right]f_2^S(y|\boldsymbol{z}_S;\boldsymbol{\theta}_2^S), \tag{6}$$

respectively. And the variables $\boldsymbol{z}_L$, $\boldsymbol{z}_S$ and $\boldsymbol{z}_\eta$ may not be the same.

3.3 Estimation procedures

Assume that a collection of independent and identically distributed realizations of $\{Y_k, \Delta_k, \boldsymbol{X}_k\}$ is available, denoted by $\left\{(y_{ijk}, \delta_{ijk}, \boldsymbol{x}_{ijk}) : k = 0,\ldots,n_{ij}, j = 1,\ldots,J_i, i = 1,\ldots,I\right\}$. We present procedures for estimating the parameters of the distribution models specified in Section 3.2 with the available data under the assumption of noninformative censoring conditional on the covariates.

Estimating the parameters in Model (1) is relatively straightforward. Following the notation introduced in Section 3.1, the likelihood function associated with the available data on Type 1 corner kick times is

$$L_1(\boldsymbol{\theta}_1) = \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{\{k:\delta_{ij,(k-1)}=0\}} f_1(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_1)^{\delta_{ijk}} S_1(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_1)^{1-\delta_{ijk}},$$

where $f_1(\cdot)$ and $S_1(\cdot)$ are the conditional density and survivor functions associated with Model (1). Maximizing the likelihood function with respect to $\boldsymbol{\theta}_1$ leads to the MLE (maximum likelihood estimator) $\hat{\boldsymbol{\theta}}_1$. Specifically, we may employ the *nlminb* function of the `R` software package to perform the optimization.

The likelihood function of the parameters in the conditional distribution of a Type 2 corner kick time based on the available data is then

$$L_2(\boldsymbol{\theta}_2) = \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{\{k:\delta_{ij,(k-1)}=1\}} f_2(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_2)^{\delta_{ijk}} S_2(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_2)^{1-\delta_{ijk}}, \tag{7}$$

where the conditional survivor function $S_2(\cdot)$ and density function $f_2(\cdot)$ are given in (5) and (6). Given the complexity of the mixture distribution, it is not straightforward to maximize the likelihood function in (7).

If the η_{ijk}'s are observable, the likelihood function based on the full data $\big\{(y_{ijk}, \delta_{ijk}, \boldsymbol{x}_{ijk}, \eta_{ijk}) : k = 0, \ldots, n_{ij}, j = 1, \ldots, J_i, i = 1, \ldots, I\big\}$ is

$$\begin{aligned} L_2^{full}(\boldsymbol{\theta}_2) = \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{\{k:\delta_{ij,(k-1)}=1\}} & [f_2^L(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_2^L)^{\delta_{ijk}} S_2^L(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_2^L)^{1-\delta_{ijk}}]^{\eta_{ijk}} \\ & [f_2^S(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_2^S)^{\delta_{ijk}} S_2^S(y_{ijk}|\boldsymbol{z}_{ijk};\boldsymbol{\theta}_2^S)^{1-\delta_{ijk}}]^{1-\eta_{ijk}} \pi(\boldsymbol{z}_{ijk};\alpha_0,\boldsymbol{\alpha})^{\eta_{ijk}} [1-\pi(\boldsymbol{z}_{ijk};\alpha_0,\boldsymbol{\alpha})]^{1-\eta_{ijk}}, \end{aligned} \tag{8}$$

and the parameter $\boldsymbol{\theta_2}$ can be estimated by maximizing the log-likelihood on the full data $l_2^{full}(\boldsymbol{\theta}_2) = log(L_2^{full}(\boldsymbol{\theta}_2))$.

In our application where the η_{ijk} values are unobservable, we adapt the commonly used EM (Expectation-Maximization) algorithm (?) for parameter estimation. This iterative procedure begins with an initial parameter estimate. In the E-step, the algorithm computes the expectation of the full data log-likelihood given the observed data using the current parameter estimate. Subsequently, the M-step maximizes the expected log-likelihood to update the parameter estimates. The procedure of EM algorithm in our application is shown as algorithm 1 below.

Using the method in **?**, we evaluate the observed Fisher information matrix:

$$E\big[-\frac{\partial^2 l_2^{full}(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2^2}|\text{observed data}, \hat{\boldsymbol{\theta}_2}\big] - Var\big[\frac{\partial l_2^{full}(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2}|\text{observed data}, \hat{\boldsymbol{\theta}_2}\big]. \tag{9}$$

The covariance matrix of the parameters is then estimated by the inverse of the observed Fisher information. All the computations were done in R software.

Algorithm 1 EM algorithm

Given the estimate at the dth iterations $\boldsymbol{\theta}_2^{(d)}$, with $\boldsymbol{\theta}_2^{(0)}$ for the initial value,

1: E-step. Calculate $Q(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^{(d)}) = E\left[l_2^{full}(\boldsymbol{\theta}_2)|\text{observed data}, \boldsymbol{\theta}_2^{(d)}\right]$ where the full data is given by $\left\{(y_{ijk}, \delta_{ijk}, \boldsymbol{x}_{ijk}, \eta_{ijk}) : k = 0, \ldots, n_{ij}, j = 1, \ldots, J_i, i = 1, \ldots, I\right\}$. η_{ijk} follows a Bernoulli distribution and $E\left[\eta_{ijk}|\text{observed data}, \boldsymbol{\theta}_2^{(d)}\right] = P(\eta_{ijk} = 1|\text{observed data}, \boldsymbol{\theta}_2^{(d)}) = \frac{P_1}{P_1+P_2}$ with

$$P_1 = \pi(\boldsymbol{z}_{ijk}; \alpha_0^{(d)}, \boldsymbol{\alpha}^{(d)}) f_2^L(y_{ijk}|\boldsymbol{z}_{ijk}; \boldsymbol{\theta}_2^{L,(d)})^{\delta_{ijk}} S_2^L(y_{ijk}|\boldsymbol{z}_{ijk}; \boldsymbol{\theta}_2^{L,(d)})^{1-\delta_{ijk}},$$
$$\text{and } P_2 = [1 - \pi(\boldsymbol{z}_{ijk}; \alpha_0^{(d)}, \boldsymbol{\alpha}^{(d)})] f_2^S(y_{ijk}|\boldsymbol{z}_{ijk}; \boldsymbol{\theta}_2^{S,(d)})^{\delta_{ijk}} S_2^S(y_{ijk}|\boldsymbol{z}_{ijk}; \boldsymbol{\theta}_2^{S,(d)})^{1-\delta_{ijk}}.$$

2: M-step. Update the estimate of $\boldsymbol{\theta}_2$ as $\boldsymbol{\theta}_2^{(d+1)} = argmax_{\boldsymbol{\theta_2}} \tilde{Q}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2^{(d)})$ using $nlminb$ function of the `R` software package.

Repeat E-step and M-step until the D-th iteration where the absolute difference of the log-likelihood value at $\boldsymbol{\theta}^{(D)}$ from it at $\boldsymbol{\theta}^{(D-1)}$ is smaller than 0.001.

4 ANALYSIS OF CSL DATA AND APPLICATIONS

We analyze the corner kick records of the CSL 2019 season games under the models presented in Section 3 and develop a predictive model for the time to the next corner kick. Two examples for applications of the model are then provided.

4.1 Model selection

4.1.1 Preliminary analysis

A practical issue of applying a proposed model to real soccer data is that the mechanism of corner kick waiting times may be different for home and away teams. A convenient way to distinguish the corner kicks by the home and away teams is to introduce an indicator of the home team as a covariate of the form $e^{\boldsymbol{z}'\boldsymbol{\beta}}$ and include it in the model. We started with this approach, considering $\boldsymbol{Z} = (X_1, X_2, ..., X_5)$. Under this setting, the home team and the away team only differ through the parameters related to the covariate X_2. We also provide a stratified analysis in which we repeat the same analysis separately for home and away teams.

In our initial analysis, we proceeded to estimate the models without incorporating covariates, in order to underscore the necessity of considering the mixture component. The results show that ignoring the mixture of Type 2-L and Type 2-S underestimates the long waiting times, and also fails to distinguish the short ones.

With the covariates, we first assume that all three sub-models Model-1 (1), Model 2-L (2), and Model 2-S (3) have different sets of parameters and estimate them by the procedure in Section 3.3. It seems reasonable to consider that Model-1 (1) is equivalent to Model 2-L (2). This is because when $\delta_{ij,(k-1)} = 1$ (indicating the most recent event was a corner kick) and $\eta_{ijk} = 1$ (indicating the previous corner kick does not lead to another one), the ball is immediately cleared, effectively resulting in a restart situation (i.e., Type 1). Under this assumption, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2^L$, which reduces the number of the parameters of the full model.

For a more detailed exposition of our model selection, please refer to Section 2 in the supplementary file.

4.1.2 A Predictive Model

Depending on the results from the preliminary analyses reported in the supplementary material, we concluded that Type 2-S time and the probability of a Type 2-S occurs may be not likely associated with any listed covariates. Therefore, we assume that there are no covariate effects on Type 2-S times, as well as the probability being the Type 2-S time. That further simplifies the model to reduce the number of parameters. As indicated in Section 4.1.1, assuming that Model-1 (1) and Model 2-L (2) are the same, we set $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2^L$. Therefore, the parameters need to be estimated are $(\boldsymbol{\theta}_2^L, \alpha_0, \lambda_2^L, \gamma_2^L)$. Having $\rho = \frac{1}{1+e^{-\alpha_0}}$ represents the probability of having Type 2-L corner kick next.

Table 2 presents the outcomes of our final model. Based on the numerical results, the probabilities of having Type 2-L next corner kick are similar for home and away teams with $\hat{\rho} = \frac{1}{1+e^{-\hat{\alpha}_0}} \approx 0.91$ with 95% CI $(0.90, 0.92)$. This suggests that the mixture assumption be essential. For the expected Type 2-S waiting time, we find it on average to be 0.62 (95% CI: (0.58, 0.66)) minutes for the home team and 0.57 (95% CI: (0.53, 0.61)) minutes for the away team. For Type 2-L and Type 1 times, the anticipated waiting time is on average 17.74 minutes (95% CI: (13.46, 22.00)) for the home team and 18.99 minutes (95% CI: (14.12, 23.86)) for the away team when all covariates are set to be 0. Remarkably, our model was able to distinguish the extremely short Type 2-S times (< 1 minute) out of Type 2 times under the assumption that Type 2-L times have same distribution as Type 1 times. Our estimates reveal that both score difference and betting odds have significant effects on

both home and away teams. Specifically, our analysis suggests that when a team is in the lead, the estimated rate of that team obtaining a corner kick be lower, leading to a longer waiting time for the next corner kicks. This observation is intuitive as leading teams tend to play cautiously and do not press forward as often (**?**). In addition, the effect of betting odds indicates that teams with larger odds (typically weaker teams) are expected to have longer waiting times for their next corner kicks. This is also intuitive as we expect weaker teams to have less possession, have less threatening positions, and consequently fewer corner kicks. Lastly, the estimated effect of the indicator of home team using pooled data suggests that the home team have shorter Type 2-L and Type 1 waiting times for corner kicks.

4.2 Applications of the Developed Model

4.2.1 Estimating probability of a corner kick

Predictions for corner kicks can be made in different scenarios with our fitted model. Figure 3 shows the estimated survivor functions by the developed model with 6 different combinations of the covariate values. The probability for the waiting time of the next corner kick to exceed a certain value can be obtained from the plot under different scenarios. For example, in sub-figure 3-1, when the home team is leading by 3 goals in the second half, no red card difference, and the betting odds of the home team are 10 (orange curve), the probability that the next corner kick performed by the home team will wait longer than 20 minutes is around 55% (95% CI: (51%, 59%)). We also note that the covariate effects on the corner kick times with the home and away teams are along the same directions, while the covariates effects are stronger associated with the home team.

4.2.2 Simulating times to corner kicks in a game

Another application of our fitted model is to simulate corner kicks. In sports betting, people may be interested in corner kick totals during a certain time period under a particular scenario. A simple approach to perform this kind of prediction is based on a Poisson regression model by treating the corner kick counts of the target period as the response variable; see for example **?**, and **?**. Alternatively, our fitted model can be used to simulate the corner kick process under certain scenarios. The extra match time, and covariates X_3 and X_4 in our model represent "in-game" information that needs to be simulated before simulating corner kick times. A practical approach involves employing grouped Kaplan-

Meier estimators to simulate X_3 and X_4. For a given set of covariates, the simulation of Type 1 and Type 2-L times to corner kicks can be achieved following the methodology outlined in **?**. Section 3 in the supplementary material presents a detailed and step-by-step explanation of our simulation process.

Although simulating one single game may not provide a realistic prediction, a large number of simulations may show some insights. To demonstrate our simulation procedure, we used the betting odds $X_5 = 5$ together with the covariates as the first half ($X_1 = 1$), home team ($X_2 = 1$) and simulated 1000 games. The computing time was only 1.11 minutes, using a laptop equipped with the 13th Gen Intel(R) Core(TM) i7-13700H 2.40 GHz processor. The computing time suggests that our simulation procedure can be potentially done in-game. Figure 4 shows 20 simulated games based on our developed model using CSL data. The simulation successfully reveals the mixture feature in the timing of corner kicks, showcasing where multiple corner kicks occur in a short time frame.

With the simulated games, different predictions can be made. For example, one may be interested in the total number of corner kicks by the home team during the first 10 minutes of the match. The prediction based on 1000 simulated games is 0.448 (2.5% quantile = 0, 97.5% quantile = 2), while the average number in our observed data is 0.468. The estimated number of corner kicks of the home team during the first 10 minutes of the second half is 0.512 (2.5% quantile = 0, 97.5% quantile = 2), while the average number in our observed data is 0.511. And the estimated number of corner kicks of the home team after 40 minutes in the second half is 0.553 (2.5% quantile = 0, 97.5% quantile = 2), while the average number in our observed data is 0.506. Other predictions can also be made depending on the subject of interest. Notably, although predicted numbers of corner kicks from our simulations closely align with the observed figures, formal evaluation of prediction performance is challenging. Our model assumes that the gap times between corner kicks are a mixture of short and long times. One would never have a clear-cut for a corner kick time being short or long in reality.

The simulation shown in this paper considers the fitted model using the CSL data. The proposed model in Section 3 can be fitted using different data to make predictions. In the scenario that the in-game information is known up to a certain time point, one can also plug in the avaliable information to modify the simulation procedure. For example, when the first half of the game is done, the score differential and differential in red cards should be known, and can be used to obtain improved prediction for the time until the first corner kick in the second half. However, our study does not separate training and testing sets

due to limited match data and the complexity of our full model. Evaluating predictions is challenging due to the mixed nature of observed times. Addressing these limitations, particularly assessing prediction accuracy, will be a focus of future investigation.

5 DISCUSSION

This paper is concerned with formulation of the corner kick process in soccer games. In particular, we address the issues of right-censored corner kick times and a mixture distribution of the times of corner kicks. It is worth noting that the sports analytics literature has yet to explore right censoring and mixture distributions in the context of corner kick event time analysis. Leveraging the power of machine learning, we employed advanced statistical techniques to handle these challenges.

We observe notably that Type 2-L times bear similarities to the time elapsed until the occurrence of the first corner kick after a game restart. In estimation of model parameters, we made a novel and reasonable assumption that Model-1 (equation (1)) shares the same parameters as Model 2-L (2). The idea allows to borrow information from Type 1 times to distinguish Type 2-S and Type 2-L times, then use both Type 1 and Type 2-L times to facilitate the estimation of parameters in Model 2-L (2) and Model 2-S (3). We emphasize that our model distinguishes Type 2 short and Type 2 long times without relying on any predetermined assumptions regarding their length.

With the proposed modeling, we analyzed the CSL data. The modeling outcomes enables us to reveal important covariates that impact the timing of corner kicks and provide sporting insights. We also showcased the simulation of the corner kick times with the estimated model. Although our main focus remains on simulating corner kick times, our preliminary simulations of other events (goal, red cards, extra match time) lay a solid groundwork for event simulations in soccer. Future investigations can delve deeper into exploring covariates effects, and potential dependencies between different events. It is important to underscore that the analytical framework presented in this paper transcends soccer and can be seamlessly extended to explore events in other sports and other scientific domains, and with machine learning playing a pivotal role in enhancing predictive accuracy.

Few limitations of the current paper suggest future research opportunities. Firstly, using our soccer intuition, we have identified covariates that impact the time until corner kicks (e.g., goal deficit, team strength, etc.). We believe that we have identified the most important ones, and therefore, conditional on these covariates, the times to corner kicks should

be roughly independent. The potential correlation of the times to corner kicks might be investigated by introducing random effects or using copula models in future investigations. Secondly, the corner kick processes of the home team and the away team in the same game can be correlated to each other. Instead of considering two separate processes for the home team and away team, one may consider one process where an occurrence of a corner kick is either by the home team or the away team. One may handle it following the idea of competing-risk.

In our analysis, corner kicks are recorded at the time when the corner kicks are awarded in the available data. However, to assist modelling, there might be a gap between the corner kick being awarded and taken. Moving forward, one may consider the taken-to-awarded as the gap time between corner kicks. This would make Type 2-S times even shorter, and easier to be distinguished. Although such information may not be available in the event data, it can be likely derived from tracking data.

Supplementary information

The supplementary material contains the additional analyses for Section 2.1, Section 4.1.1, Section 4.2.2 in the main text.

Acknowledgments

This research is partially supported by the grants of Natural Sciences and Engineering Research Council of Canada (NSERC) to Hu and Swartz, and the CRT (Collaborative Research Team) in Sports Analytics of the Canadian Statistical Sciences Institute (CANSSI) led by Swartz. The authors thank Daniel Stenz, former Technical Director of Shandong Luneng Taishan FC for providing the data discussed in this paper.

Declarations

The authors declare no conflict of interest.

A Figures

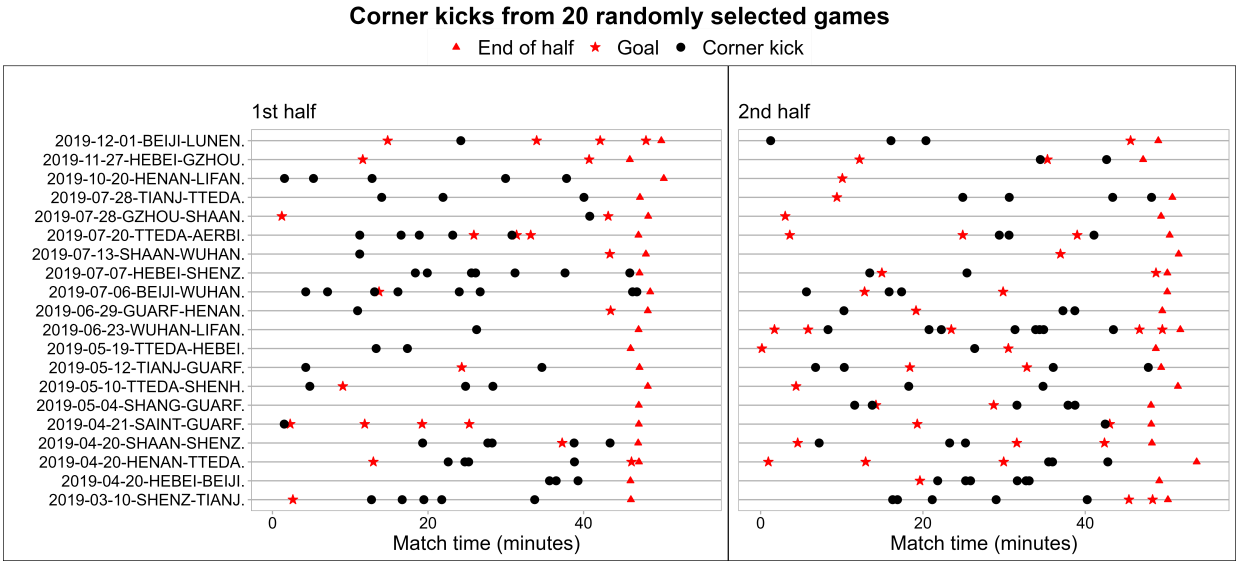

Figure 1: This plot presents corner kicks from 20 randomly selected games from the CSL. Dots are corner kicks, stars are goals, and triangles denote the end of the half. Only the corner kicks of the home teams are shown.

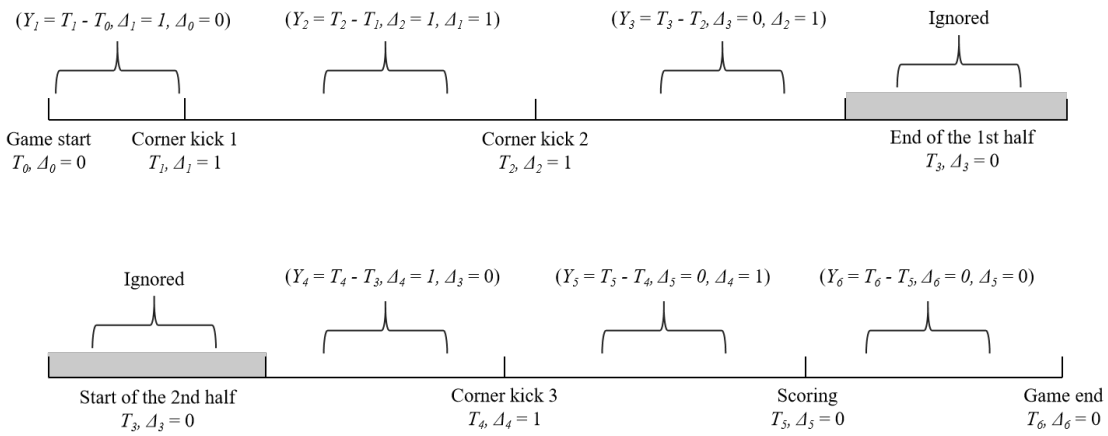

Figure 2: This figure illustrates the notation introduced in Section 3.1, where T_k is the time to an event of interest, associated with a team in a game. The time Y_1 is the observed gap between the game start and the first corner kick, an observed Type 1 time with $\Delta_1 = 1, \Delta_0 = 0$; Y_2, the time between corner kick 1 and corner kick 2, an observed Type 2 time with $\Delta_2 = 1, \Delta_1 = 1$; Y_3, the time between corner kick 2 and the end of the 1st half, a right censored Type 2 time with $\Delta_3 = 0, \Delta_2 = 1$; Y_6, the time between scoring to the game end, a right censored Type 1 time with $\Delta_5 = 0, \Delta_6 = 0$. The shadowed part represents the breaking time between first and second halves. It is not considered in our analysis since the game is paused during that period.

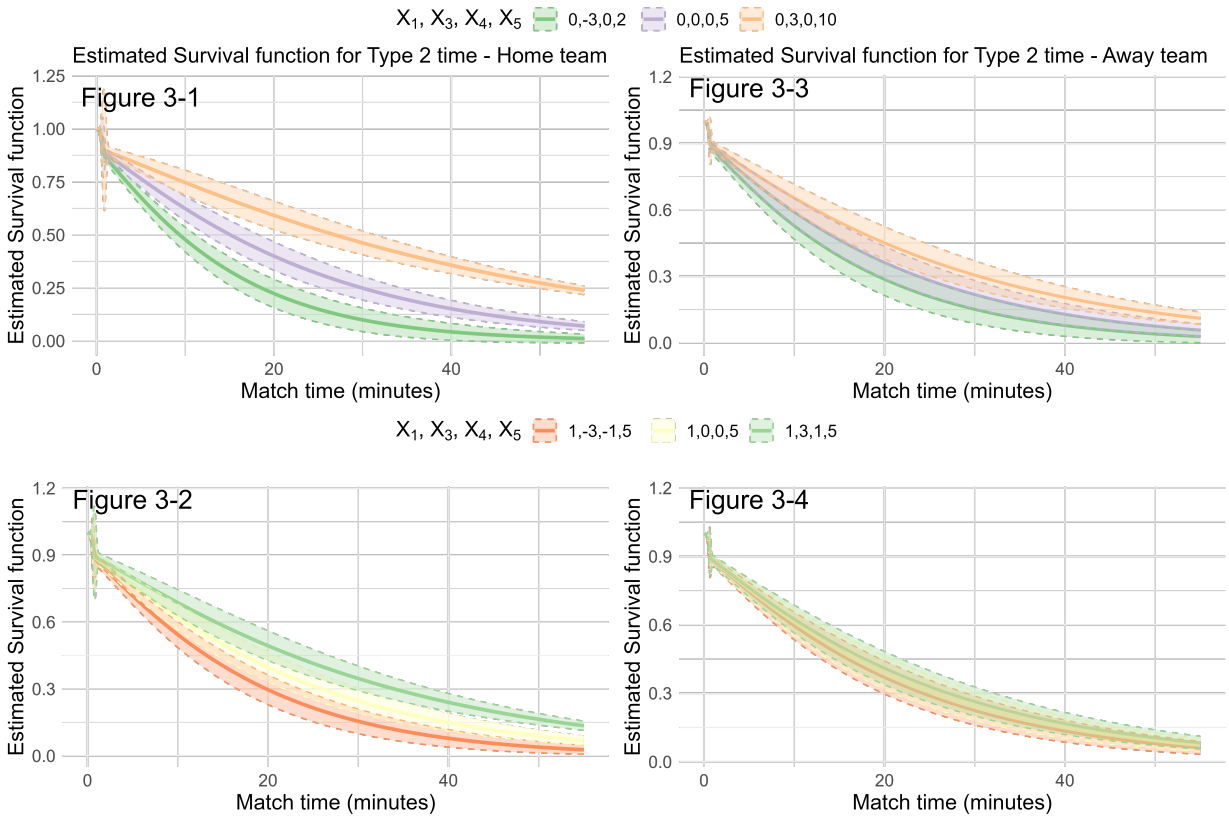

Figure 3: Comparison of the estimates of survival functions of Type 2 times and their 95% CIs based on the final (developed) model. The standard error estimates were obtained using the multivariate delta method. We consider 6 combinations of diverse covariate patterns where X_1 is the indicator of first half, X_3 is the score difference, X_4 is the differential in red cards, and X_5 is the betting odds of the team.

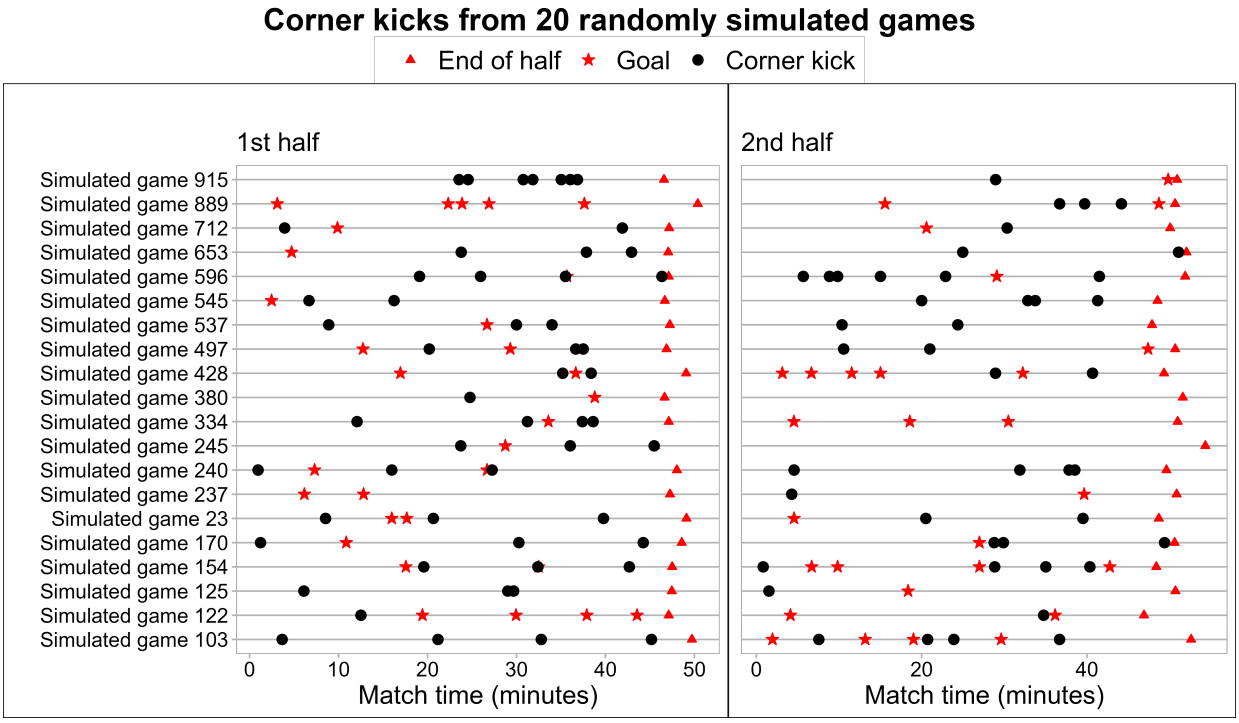

Figure 4: This plot presents corner kicks from 20 simulated games based on the developed (final) model. Dots are corner kicks, stars are goals, and triangles denote the end of the half.

B Tables

Covariates	**Description**
Indicator of first half (X_1)	1 if first half, 0 if second half
Indicator of home team (X_2)	1 if home team, 0 if away team
Score difference (X_3)	Positive X_3 means the team is leading, Negative X_3 means the team is losing, Zero X_3 means a draw
Differential in red cards (X_4)	Positive X_4 means the team has more players, negative X_4 means the team has fewer players. In soccer, a red card prevents a player from playing for the remainder of the match and as a result reduces the number of players that are available to a team. Therefore, the differential in red cards reflects the differential in number of players.
Betting odds of the team (X_5)	Decimal odds (European odds) are a common way to represent betting odds in soccer and other sports. They are presented as decimal numbers such as 2.00, 1.50, or 3.75 and represent the potential payout from a successful wager. Higher odds offer greater potential profits but are associated with lower probabilities of winning

Table 1: Covariate list and descriptions

Parameter	**All data Estimates** (SE)	**Home team Estimates** (SE)	**Away team Estimates** (SE)
Rate of baseline hazard (λ_2^S)	1.618 (.014)	1.463 (.057)	1.611 (.045)
[a]Shape of baseline hazard (γ_2^S)	**4.834 (.251)**	**4.293 (.297)**	**5.012 (.359)**
[b]Home/Away team (β_{22}^S)	**-0.276 (.059)**	.	.
Rate of baseline hazard (λ_2^L)	0.050 (.001)	0.054 (.001)	0.051 (.003)
[a]Shape of baseline hazard (γ_2^L)	**1.104 (.028)**	**1.127 (.036)**	**1.087 (.031)**
[b]First/Second half (β_{21}^L)	-0.025 (.048)	0.007 (.080)	-0.082 (.070)
[b]Home/Away team (β_{22}^L)	**0.097 (.044)**	.	.
[b]Score difference (β_{23}^L)	**-0.088 (.017)**	**-0.122 (.028)**	**-0.055 (.025)**
[b]Differential in red cards (β_{24}^L)	-0.015 (.069)	0.062 (.097)	0.104 (.098)
[b]Betting odds (β_{25}^L)	**-0.027 (.013)**	**-0.057 (.026)**	**-0.021 (.010)**
[c]P(Type 2 times to be long) (ρ)	**0.910 (.007)**	**0.908 (.011)**	**0.910 (.011)**

[a] Bold estimates of shape parameter indicate significance deviating from 1.
[b] Bold estimates of regression coefficients (β's) indicate significant covariate effects.
[c] Bold estimates of ρ indicate significance deviating from 1, which indicate the importance of the mixture.

Table 2: Estimates of the baseline parameters and regression coefficients in the final (developed) model, assuming Model (1) = Model (2). The column "All data" shows the results using pooled data from both home and away team; the columns "Home team", "Away team" show the results of the stratified analysis where the home team and the away team are considered separately.