Analysis of Success Probabilities in Field Hockey with Machine Learning

Jethro R. Lee, Northeastern University

Abstract

The main goal of this project is to use modern machine-learning techniques to analyze field hockey player performance. Field hockey is a relatively heretofore unexplored application, and there would be a great benefit in building a model to predict players' ability to score under different conditions. We have scraped public data from Northeastern's 2023 field hockey season, and implemented a prototype mixed effects logistic regression model with both glmer [2] and RStan [6]. The model investigates effects that could impact players' goal-scoring probability, including match location, score difference between the competing teams, and situational effects such as scoring off a penalty corner or not. The model also estimates random effects capturing individual players' likelihood to score a goal. By interacting player effects with whether a shot came after a penalty corner, the model supports an expectation from the Northeastern coaching staff that some players perform worse when shooting after penalty corners due to game strategy.

Introduction

Field hockey is a sport that has thus far been understudied. Yet, all sports contain layers that we should unfold. The appeal of sports is in their relative entropy—otherwise, no one would stay on their toes yearning to see the outcome of a match or feel the adrenaline of making nail-biting sports bets with their friends. This project seeks to implement machine-learning models that use play-by-play data from field hockey matches to better inform decisions on the field. We intend to use the model to evaluate factors that lead certain field hockey teams or players to perform better or worse throughout a season and predict player performance based on these factors. We will assess how players contribute to field hockey matches by investigating the significance of on-field events such as penalty corners, goals scored, and others, and we will assess the on-field value of the outcomes of these events, especially on a player basis.

Success probabilities have been studied extensively in other sports. In baseball, for example, a bivariate binomial distribution model for player performance was developed to model extra-base hit percentage as a probability nested within the first-level success probability of batting average [3]. While our ultimate goal is to implement this model within a field hockey framework, we first focus on developing lower-level models, including a mixed-effects logistic regression model, to assess goal-scoring ability in field hockey. Success probabilities have been investigated in the sport most related to field hockey, ice hockey, where researchers showed that it may be useful to consider success in a nested fashion; for example, winning a face-off is valuable for a team, but more valuable if it is a "clean" face-off win [1]. Other relevant studies include those that employed decision tree models to observe the factors that influenced the turnout of certain games in ice hockey matches [4], such as several studies evaluating the value of face-offs and/or players' ability in them in ice hockey [1][5][7]. As stated previously, field hockey is a novel application in the realm of sports analytics. Tackling that randomness in field hockey via statistical modeling is an extensive project that will be rewarding for those wanting to

see the value of how precise movements within a match impact the valuation and performance of an athlete and their impact on the team. Furthermore, while coaches cannot control all the precise movements made by their players, more research into field hockey could provide them general insight into what situations they should strive for their players to be in, whether it be setting them up for success in penalty corners, adapting field positioning, strategizing ball movement, etc. This model aims to enable coaches to evaluate the performance of specific players and what types of current circumstances boost their chances of scoring a goal, helping coaches predict the conditions they should aim to get for those players in future matches.

Methodology

Data Collection

Play-by-play data from Northeastern's field hockey website was scraped using the Python Pandas library to create a spreadsheet mapping time into a game to important events in a match (e.g., goal, penalty corner, assist). Other useful information was also manually recorded, such as whether a match occurred at home or away, how many points ahead or behind Northeastern was, and if a penalty corner led to an immediate shot. Only one season and one team (Northeastern University) were investigated due to time constraints and to eliminate the need to account for a player's aptitude for scoring varying between seasons. Keeping the data as homogenous as possible ensures the random effects best measure the average performance of each player within a season. With the season investigated, the model was able to use data from 18 field hockey matches and 19 hockey players, which included 270 shots, 117 penalty corners, and 50 goals.

Mixed effects logistic regression

$$\text{logit}(p_{\ell t}) = \beta_{p0}^{(0)} + \beta_{p1}^{(0)} z_{\ell t1} + \ldots + \beta_{pk}^{(0)} z_{\ell tK} = \mathbf{z}_{\ell t} \boldsymbol{\beta}_p'^{(0)}$$

Model 0: Logistic regression without random effects

Model 0 describes the baseline fixed effects logistic regression model. The left term, $\text{logit}(p_{\ell t})$, represents the probability a player makes a goal. Each z term represents the value of a different parameter affecting one's goal-scoring ability (e.g., score difference, home vs. away, location, assists, blocks, etc.). The β terms portray the extent to which the corresponding parameter affects goal-scoring ability. If the probability a player scores a goal increases as the value for a parameter decreases, the parameter's corresponding β will be negative, whereas parameters that positively impact the probability a player scores a goal as it increases get multiplied by positive β values. The β not multiplied by a z term is the intercept term, which functions as a bias term that enables the model to optimally agree with as much of the data as possible.

$$\text{logit}(p_{\ell t}) = \beta_{p0}^{(1)} + \beta_{p1}^{(1)} z_{\ell t1} + \ldots + \beta_{pk}^{(1)} z_{\ell tK} + a_{\ell}^{(1)} = \mathbf{z}_{\ell t} \boldsymbol{\beta}_p'^{(1)} + a_{\ell}^{(1)}$$

Model 1: Logistic regression with random effects

The drawback of Model 0 is that it assumes all players in the data set are equally proficient at goal scoring except for random noise; we can adjust for the likely differences in player ability by including random effects for players. In Model 1, the a terms are random effects, which alter the extent to which the parameter values (z) affect each individual's goal-scoring probability, logit(p_{it}). Players who miss shots that the model classifies as more difficult to make don't receive as harsh negative impacts to their value for logit(p_{it}) than players who miss "easier" shots. To illustrate, if players tend to score better if they are ahead of the opposing team by 3 points, then the value for logit(p_{it}) gets more negatively impacted for players who consistently miss their shots when their team is ahead by 3 points. Conversely, if the data portrays that players tend to miss their shots if they are made outside the circle, players who keep missing shots outside the circle receive mitigated negative effects to their value of logit(p_{it}).

While not illustrated in Model 1, in our implementation, we include an interaction between the player random effect term and the indicator covariate of whether a shot was taken after a penalty corner or not. This addition stems from a discussion with the Northeastern coaching staff where they suggested that certain players take shots directly after penalty corners not intending to score, but rather to force the ball to rebound to a player in a better position. Thus we expect that certain players will perform better from the field of play than immediately following penalty corners and wish to capture this effect. The inclusion of the interaction term adds another set of random effects so that there are now two per player; one which estimates the player's goal-scoring performance on penalty corner set plays, and one which estimates their goal-scoring performance on all other shots.

Code implementation

Modeling strategies

To implement the logistic regression models, two software packages were used. Their outputs were also compared. The first package is lme4, to employ the glmer function, which uses a numerical approximation technique (i.e., Laplace approximation) to obtain β values and random effects (See Appendix A) [2]. The other package was RStan, which approximates β values by fitting a Bayesian hierarchical model (See Appendix B) [6]. Both packages generally emulate the mixed effects logistic regression model in Model 1. Our intuition is that using Laplacian approximation for our intended final bivariate binomial model would be more beneficially complex to implement than using Bayesian modeling in RStan.

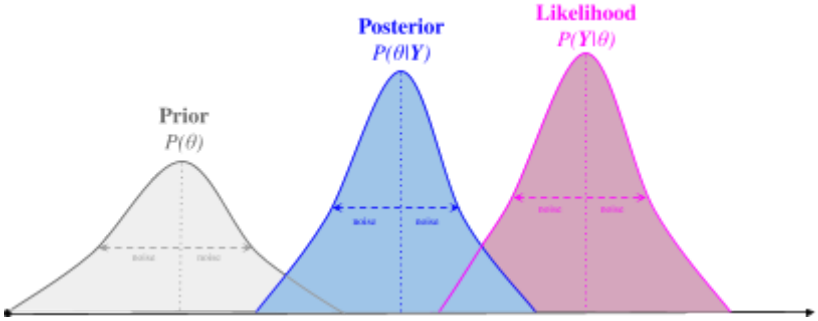

Figure 1: Bayesian approximation model

Here is a basic overview of how RStan calculates players' goal-scoring probabilities based on circumstances. In RStan, the prior distribution on each player's scoring probability is normal with a mean of 0 and variance σ^2 which is given a hyperprior distribution that is uniform from 0 to 10. Essentially, for each player, the model initially assigns them average goal-scoring probabilities, indicating each player's propensity to score a goal is the same. In Figure 1, the gray arch symbolizes a player's prior distribution, the distribution of the probability of a player scoring a goal, $P(\theta)$. However, since the model portrays factors such as time, scoring off a penalty corner or not, home-field advantage, and specific player random effects affecting players' goal-scoring abilities, the model adjusts each player's goal-scoring probability according to the data, producing the posterior distribution of a player's goal-scoring ability:

$$P(Y|\theta) = \frac{P(\theta|Y)P(Y)}{P(\theta)}$$

Equation 1: Bayes' theorem

The posterior distribution, $P(\theta|\mathbf{Y})$, defines the probability an event occurs (e.g., shooting off a penalty corner, match happening at home, score difference of +3) given a player. With the prior probability of a player scoring a goal, $P(\theta)$, and the posterior probability of an event occurring given a player, $P(\theta|\mathbf{Y})$, the model can use Bayes' theorem to find the probability a player scores a goal given the value of a parameter (e.g., home vs. away, score difference, shot off penalty corner or not), $P(\mathbf{Y}|\theta)$ (Equation 1). This value is the likelihood probability. Its distribution is shown in Figure 1 in pink.

Model Design

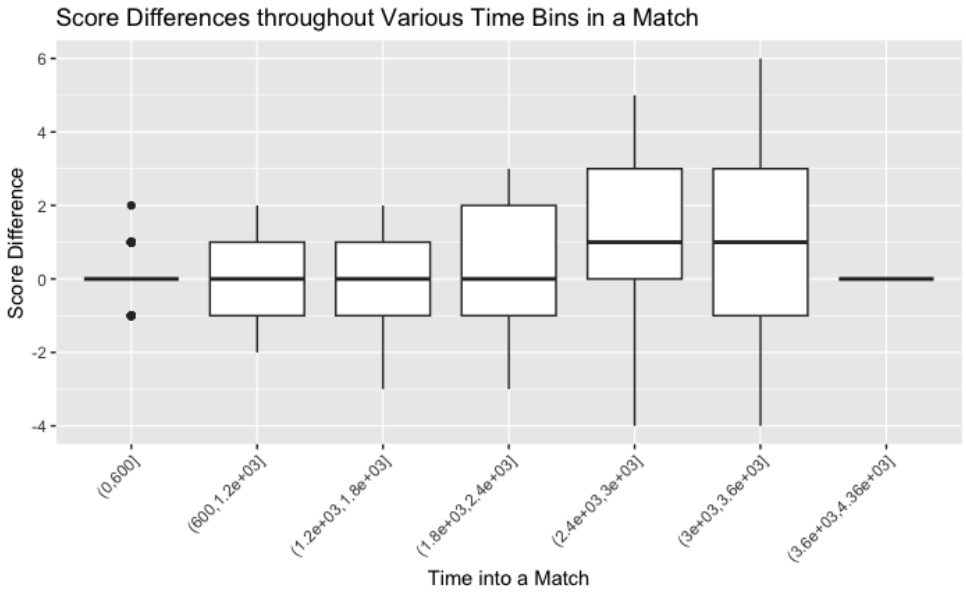

Figure 2: The boxplot shows the range of score differences varies considerably throughout a match, with opponents tending to have the same score at the start of a match or during overtime. The median score difference steadily increases with time along with the range of the boxes, suggesting the Northeastern team becomes increasingly ahead or behind (mostly ahead) with time.

GOAL is the output of interest, where logit($p_{\ell t}$) represents the log odds of scoring a goal (GOAL = 1). Home (whether a game was at home, Home = 1, or away, Home = 0), Shot.after.PC (whether a shot occurred after a penalty corner), and Blocked.or.goal.after.block (whether a shot was blocked or a goal happened after a blocked shot) are all indicator variables. The parameters match_seconds_bin (time into a match, in seconds) and Score.Difference (how many points ahead or behind the Northeastern team is compared to its opponent) are numeric variables, though we bin the time into a match to better capture the non-linear effects of time on goal scoring. We expect an interaction effect between the two variables. The coefficients involved with each interaction, their p values, and closer investigation of the data set support this assertion (See Appendices C and D, Figure 2).

Both models also include two random effects for players. The random effect for a player is defined as having an interaction with the binary outcome of whether a shot was made after a penalty corner. With glmer, this fact was accounted for via (1 | Player: Shot.after.PC), while in the RStan code, this fact was accounted for by defining each player to have two random effects organized in a matrix, with one column of values containing the random effects for when Shot.after.PC == 0 (FALSE), and one for when Shot.after.PC == 1 (TRUE) (Appendices E and F). Random effects that are near 0 represent players who are average in goal-scoring ability and/or have minimal data.

We decided to implement the specific interaction between the player and Shot.after.PC because a discussion with a Northeastern field hockey coach confirmed that some players strategically miss shots on purpose in hopes of a play leading to a deflection. As the results will show, this factor proved essential in predicting players' goal-scoring ability, corroborating the additional interaction.

Results

Fixed Effects Parameter Estimation

Model 1 was implemented in both glmer and with RStan. After running the model in glmer, it resulted in values in the "Estimate" column listing the β values for the parameter in the corresponding row (See Appendix C). In RStan, the values in the "mean" column of Appendix D contain the β values for different parameters. For 5 chains with 3,000 iterations each, RStan ran the model, which calculated coefficient values for each parameter each time. The "mean" column presents the average coefficient for each parameter from all the chains (See Appendix D).

Beta1 values [1] to [7] are coefficients for increasing time bins (0 to 600 seconds into a match, 600 to 1,200 seconds into a match, 1,200 to 1,800 seconds into a match, 1,800 to 2,400 seconds into a match, 2,400 to 3,000 seconds into a match, 3,000 to 3,600 seconds into a match, and 3,600 to 4,360 seconds into a match, respectively), beta1[8] represents Score.Difference, beta1[9] represents Home, beta1[10] represents Blocked.or.goal.after.block, beta1[11] to beta1[17] represents each time bin interacting with Score.Difference, and beta2[1] represents Shot.after.PC.

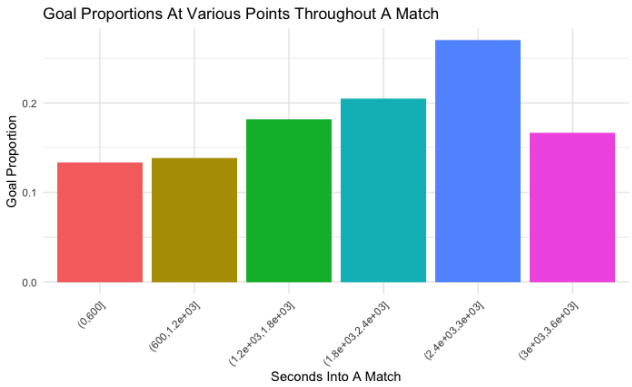

Figure 3: Northeastern generally makes more goals as time progresses

With glmer, there are considerable values for each match_seconds_bin coefficient. Disregarding the value of -14.2380 for the parameter identified with 3,600 to 4,360 seconds into a game (overtime), the match seconds bin parameter coefficients range from 0.3425 to 1.0188 (See Appendix C). The coefficients generally increase as match_seconds_bin holds larger values, meaning the Northeastern team tends to score more goals as time passes (except for overtime). The RStan model, again disregarding the mean coefficient for overtime (-77.38), assigns beta values ranging from 6.61 to 7.14 to each time bin, with the means again steadily increasing with time bins representing seconds further into a match (See Appendix D). Indeed, there appears to be a steady increase in goal proportion among the Northeastern hockey players as time within a match increases (Figure 3). Since overtime's coefficient has a very large standard error (-14.2380) and a p-value very close to 1 (0.99196) in glmer as well as an extreme standard deviation in RStan (65.97), we can disregard the overtime data as outlier data (See Appendices C and D).

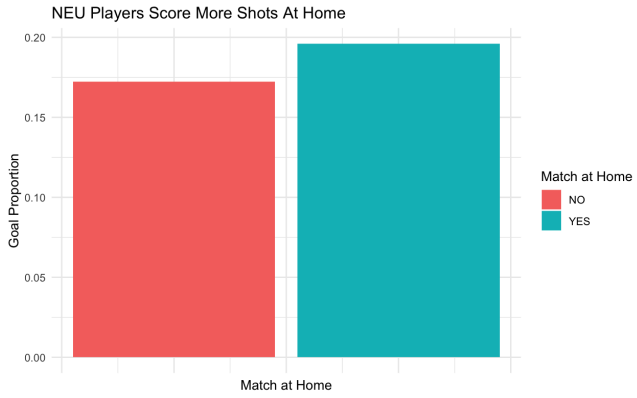

Figure 4: Northeastern tends to score about 2.5% more of its shots at home rather than away

Conversely, the positive values of the coefficient for home vs. away (0.5611 in the glmer model and 0.72 in the RStan model) suggest Northeastern's goal proficiency is boosted by

home-field advantage (See Appendices C and D). The team does indeed score about 2.5% more of its shots into the net when at home rather than away (Figure 4).

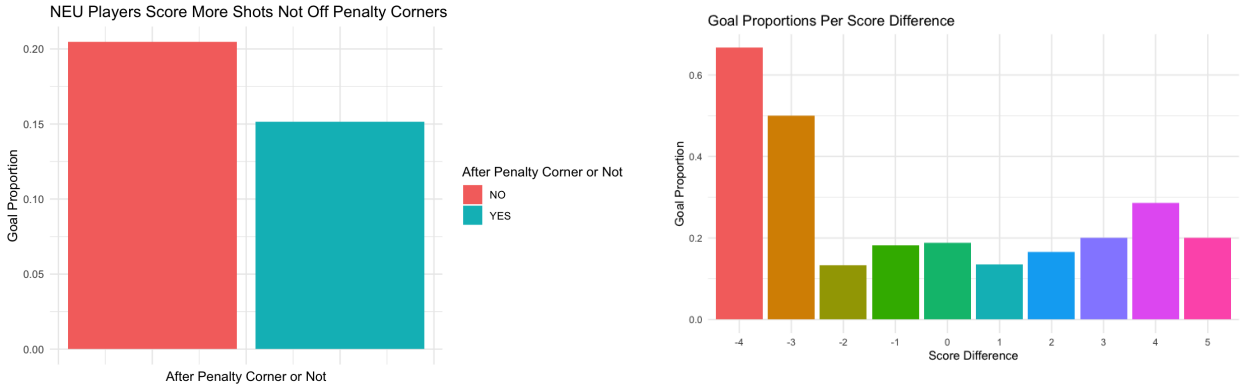

Figure 5: (a) Northeastern scores more shots not after penalty corners (left). (b) It also tends to score more when ahead except when it is significantly trailing (right).

In the glmer model, the Score.Difference coefficient is positive (1.1885) while the Shot.after.PC coefficient is negative (-0.2258), indicating players tend to score more when ahead of the opposing team and are shooting not after penalty corners (See Appendix C). The greater absolute value for the coefficient of Score.Difference suggests that the parameter has a greater impact on a player's goal-scoring probability than whether they are shooting off a penalty corner. In the RStan model, the Score.Difference parameter is negative (-1.45) while the Shot.after.PC coefficient is also negative (-0.19), indicating a discrepancy between the two models that need to be addressed (See Appendix D). Both models align with how the data supports players having a better chance of scoring a shot not off a penalty corner (Figure 5a). Yet, the glmer Score.Difference coefficient better aligns with the data, except for when the Northeastern team is very behind (by 3 or 4 points) in which its goal proportion actually peaked, which may explain the inconsistency between the Score.Difference coefficients in glmer and RStan (Figure 5b). Still, RStan also indicates that Score.Difference more greatly impacts scoring ability than Shot.after.PC (albeit more negatively).

Specifically, the Northeastern team scored about 20.468% of its shots not after penalty corners. In comparison, it only scored about 15.152% of its shots after penalty corners. (Figure 5a). Therefore, for every 6.667 shots off penalty corners, the Northeastern team is able to score a goal (1/proportion of shots made after penalty corners = 1/0.152 = 6.667). The Northeastern hockey coach reported that world-class teams score 1 goal for every 3 penalty corners. We hope that some of the insights from the player random effects can help the team improve upon their penalty corner performance, inching them closer to world-class standards.

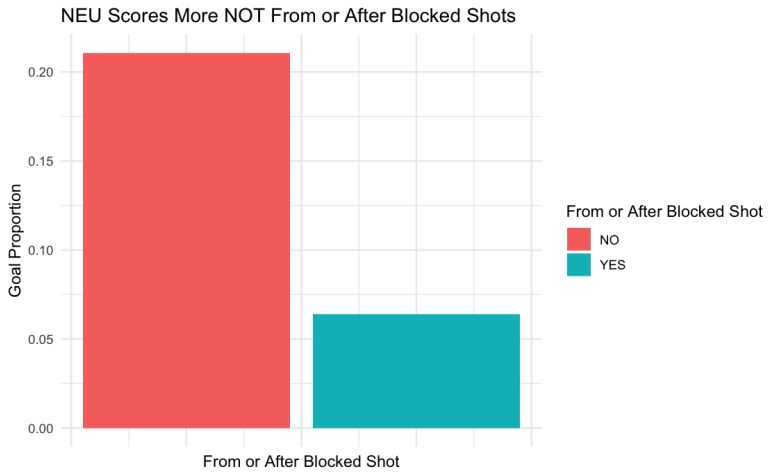

Figure 6: Many more goals are made not from or after a blocked shot

Both models also have a coefficient for the parameter Blocked.or.goal.after.block, which represents whether a player made a shot that was blocked or a goal after a blocked shot. A goal made after a blocked shot indicates a successful deflection (where a defender tries to block a shot that ends up in the net). The coefficient for this parameter is negative in the glmer model (-1.3918), which aligns with how most players are better at making goals that aren't blocked or following blocked shots (See Appendix C, Figure 6). The RStan model also assigns an appropriately negative coefficient for the parameter (-1.84) (See Appendix D).

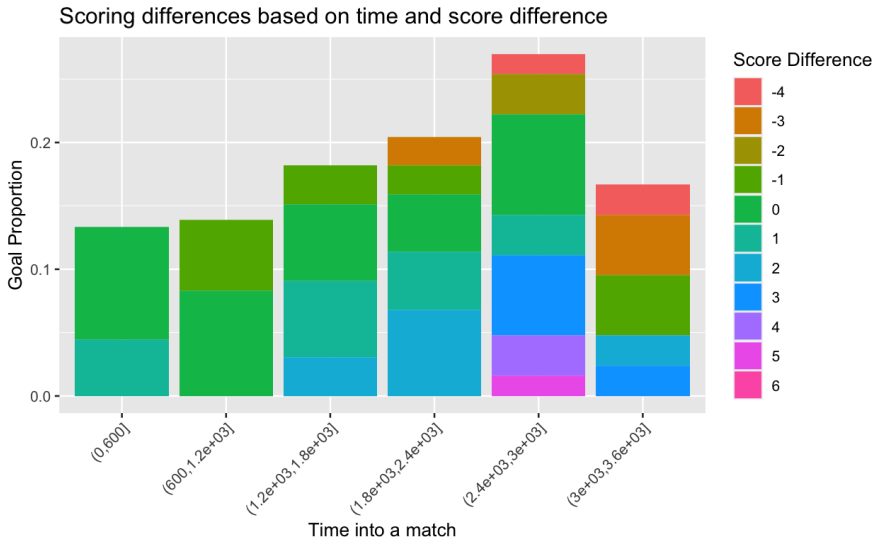

Figure 7: Over time, the Northeastern team usually scores more goals. Between 2,400 and 3,000 seconds into a match, the team makes goals in the most various score difference situations, though it notably tends to have the smallest goal proportion (goals/shots) in that timeframe when tied with the other team

We also observed an interaction between time and score difference. Appendix C presents the glmer model's coefficients for the interaction between time and score difference. The values in the first column represent Score.Difference interacting with each unique value for

match_seconds_bin. The p values in the last column are not close to 1. Appendix D contains the RStan model's coefficients for the interaction between time and score difference. The coefficients in the "mean" column are also not glaringly extraneous except for beta1[7]. Hence, the use of the interaction is justified.

Regarding the interaction, the negative coefficients for each interaction parameter in the glmer model (e.g., match_seconds_bin(600, 1.2e+03]:Score Difference, match_seconds_bin(1.2e+03, 1.8e+03]:Score.Difference) suggest that as the values in a time bin increase with Score.Difference, the likelihood of a player making a goal decreases (See Appendix C). In other words, within each sixth of a match, as the Northeastern team exceeds the other team by more points, it scores fewer goals, which the data tangentially agrees with (Figure 7). Throughout a match, as the legend indicates, Northeastern is mostly behind or tied with the other team when it scores. Beta1[11] to beta1[17], representing the interaction between different time bins and score difference in the RStan model, are positive, indicating another discrepancy that may be due to the inconsistent trend among time, score difference, and goals (See Appendix D).

Player random effects and penalty corners

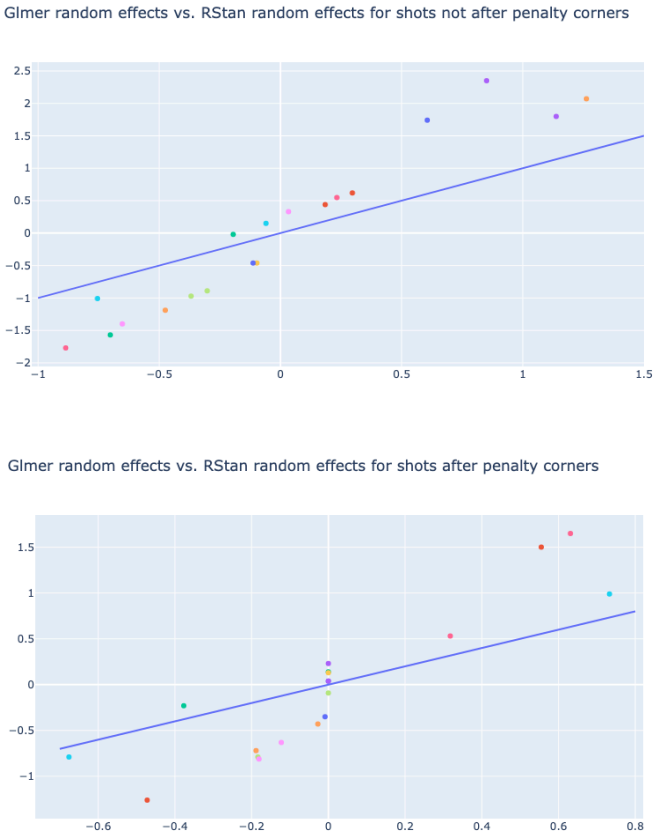

Figure 8: Comparison of random effects between glmer and RStan

In Appendix E, the random effects from the glmer model, interacted with Shot.after.PC, are presented. The "mean" column in Appendix F lists players' random effects from the RStan model, which also interact with Shot.after.PC. Even though the corresponding random effects are

different between glmer and RStan, there is a general agreement between random effects for the same player in the same penalty corner situation between the models (Figure 8).

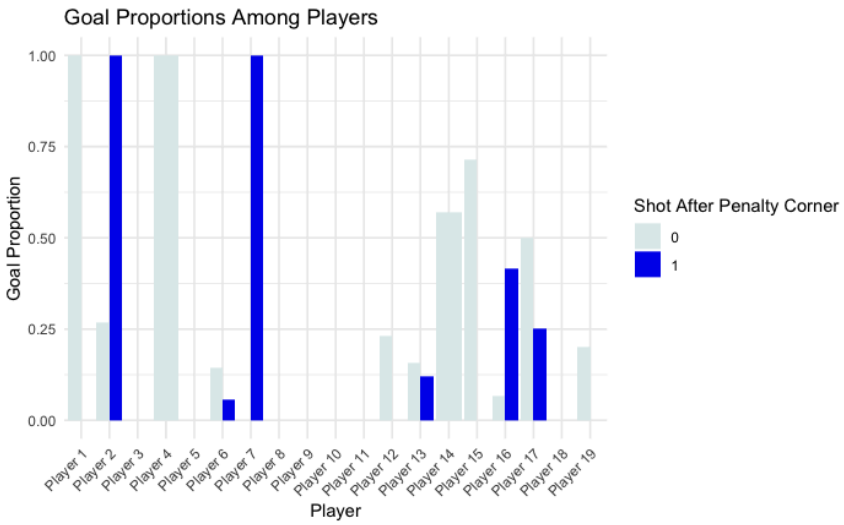

Figure 9: Differences in goal proportions among players when they score after penalty corners vs. when they don't score after penalty corners.

Higher random effects values are associated with players who are better at scoring goals considering circumstantial difficulties defined by time, whether a match is home or away, score difference, whether a shot happened after a penalty corner, and whether a shot is blocked or after a previously blocked shot (deflection). Both models propose the players represented by Player 14/psi[14] and Player 15/psi[15] are the best scorers in terms of shots not off penalty corners, as they each yield the highest random effect values under the condition that Shot.after.PC is FALSE across both models (See Appendices E and F). These players also have among the highest scoring percentages in the studied season. The data, showing that those players have the high scoring percentages regarding goals made not after penalty corners, justifies this finding (Figure 9). Players 1 and 4 both only made 1 shot not after a penalty corner that happened to be a goal. Therefore, both the glmer and RStan models were able to correctly discern that Player 1 and Player 4's goal proportions of 1.000 for shots not off penalty corners are not comparable due to a lack of data.

Conversely, Player 2/psi[2], Player 7/psi[7], and Player 16/psi[16] demonstrate the best ability to score shots after penalty corners, as they each yield the highest random effect values under the condition that Shot.after.PC is TRUE across both models (See Appendices E and F). The data also presents those players as having among the highest scoring percentages in terms of shots off penalty corners (Figure 9). Players such as Player 11/psi[11] with very low random effects tended to have scoring percentages of 0 regardless of the penalty corner situation (See Appendices E and F, Figure 9).

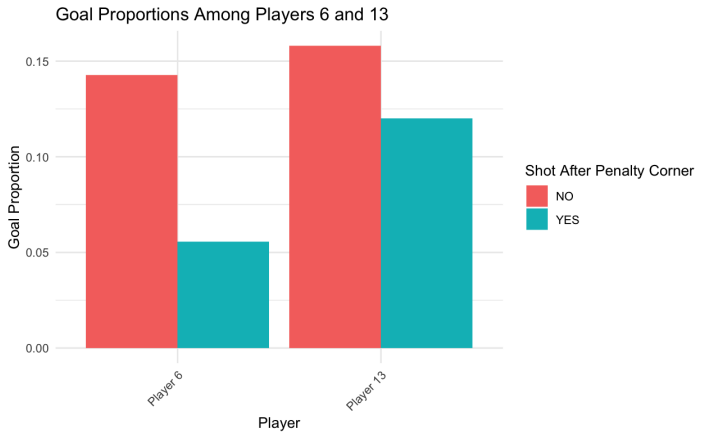

Figure 10: Players 6 and 13 sometimes miss shots after penalty corners on purpose, which is why their goal proportions for shots made after penalty corners are lower than theirs for shots made not after penalty corners

In both models, the random effect for player interacts with the parameter of whether a shot was made after a penalty corner since, as mentioned previously, discussion with the coaching staff taught me some players such as Player 6/psi[6] and Player 13/psi[13] occasionally take "bad" shots after penalty corners for strategic purposes. Hence, some players score many more goals not after penalty corners than after. The data and random effect values corroborate this statement (See Appendices E and F, Figure 10). The random effect values for player 6's ability to score goals after penalty corners are -0.676 in the glmer model and -0.79 in the RStan model while their random effect values for their ability to score not after penalty corners are higher: -0.059 in the glmer model and 0.15 in the Rstan model. Similarly, player 13's random effect values for their ability to score goals after penalty corners are -0.377 in the glmer model and -0.23 in the RStan model, while those indicating their ability to make goals not after penalty corners are -0.195 in the glmer model and -0.02 in the RStan model. These values show that while due to strategic reasons they are considered poor scorers after penalty corners, on all other shots, they are much closer to average in ability.

Interpreting the coefficients

$$log(\frac{p}{1-p}) = a + \beta X$$

Equation 2: Solving for $\frac{p}{1-p}$ retrieves the probability of scoring a goal over the probability of not scoring a goal under one condition

To illustrate what a parameter's coefficient actually means, consider the effect of Home on goal-scoring probability based on the RStan model. Equation 2 calculates the probability of scoring a goal over the probability of not scoring a goal under certain conditions. For a, input the model's intercept term (-9.35 in the RStan model (See Appendix D)). For β, input the coefficient for the parameter of interest (RStan's coefficient for Home was 0.72 (See Appendix D)). For X, input a value for the parameter of interest. To investigate Northeastern's probability of scoring a goal over its probability of not scoring a goal when the team plays at home, X would be set to 1. To investigate the same effect when the team plays away, X would be set to 0.

$$\frac{p}{1-p} = e^{0.72(1)} = 2.05443$$

Equation 3: The probability of scoring a goal over the probability of not scoring a goal at home

To isolate the effect of home-field advantage on goal-scoring probability, the intercept can actually be disregarded. Hence, Equation 3 shows that when solving for $\frac{p}{1-p}$, there is about a 105.443% chance ((2.05443 - 1) x 100) of Northeastern scoring a goal over not scoring a goal when playing at home rather than away. Hence, Northeastern was about twice as likely to score a goal at home turf vs. away throughout the season.

$$log(\frac{p}{1-p}) = a + \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n$$

Equation 4: Solving for $\frac{p}{1-p}$ retrieves the probability of scoring a goal over the probability of not scoring a goal under n conditions

The effect that multiple parameters have on the probability of scoring a goal over not scoring a goal can eventually be calculated using Equation 4. Similarly, a represents the intercept of the model, while each β is the coefficient for a different parameter being multiplied by X, the actual value of the same parameter.

Discussion and Future Work

While the random effects logistic regression model provides some useful insight into the player's goal-scoring ability and some strategy, there are still two major steps we wish to take in furthering this analysis. Ultimately, we hope to move to a bivariate binomial model for accounting for the nested structure of some of the success probabilities: goals scored from penalty corners are directly nested in the success of the penalty corner itself. If the team's overall probability of scoring a goal after a successful penalty corner is truly lower than that after the absence of a successful penalty corner, coaches should not only encourage players to practice scoring off penalty corners more often but also devise more alternative methods of scoring well-suited for the team and its strengths. This nested probability model should also account for the probability of a player missing penalty shots for specific reasons. Certain athletes missing a shot off a penalty corner leads to a greater chance for a deflection that some coaches believe could better enable their team to score compared to if a player attempted to score directly off a penalty corner shot. Hence, the nested probability model should not classify all missed shots off penalty corners as detrimental to a player's goal-scoring probability.

We currently have a bivariate binomial model that first uses logistic regression to calculate the likelihood of an individual shooting off the penalty corner being observed. The implementation of this model in RStan mirrors somewhat that of the logistic regression model in Appendix B. It calculates weights for each parameter affecting the probability of a player shooting off a penalty corner, as well as a random effect value accounting for the inherent differences among players that make them more or less prone to shooting off a penalty corner.

If a shot is made off an observed penalty corner, the bivariate binomial model then calculates the likelihood of that shot leading to a goal. Here, weights are found for each

parameter affecting the likelihood of a shot made off a penalty corner leading to a goal. Furthermore, random effects values are calculated to address the inherent differences among players that make them more or less prone to making a goal off a penalty corner. While we have some preliminary results from this bivariate binomial model, we are not prepared to present them here.

To ensure the bivariate binomial model ran as expected, we simulated data on fake values and observed whether pre-defined fixed parameter and random effect values aligned with those recovered by the model. Our current simulation studies have been successful, but more advanced simulations will be performed to verify the validity of the model.

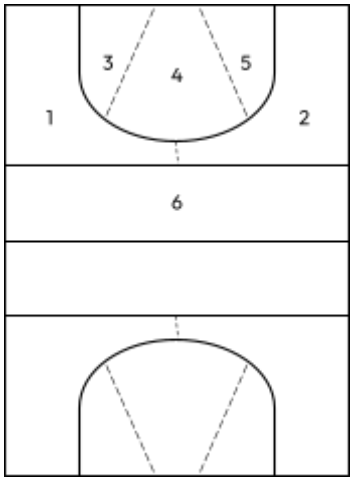

Figure 11: The system used for determining the location of match events

The other major step we wish to take is to add breadth to the model with field location data for each important event in a match and where they happened among the 6 sites in Figure 11. This diagram is only one framework for measuring location and may change depending on the focus of the analysis. The Northeastern coaching staff has also expressed interest in examining when and where a passed ball enters the circle before shooting versus when the player makes the shot and/or goal. To increase the number of locations, we may also only consider the half of the field where the team on offense is defending to assess defensive abilities.

With the location data, the model can include an additional interaction between a player and location, causing the model to output more random effects. It would be useful to have additional random effects accounting for the probability of a player scoring a goal based on their location depending on whether they are scoring off a penalty corner. If Figure 11 were to be used in the final model, then given a player is scoring off a penalty corner, it may be useful if the model could create different random effects accounting for the probability of whether the player is shooting from zones 1-6. Conversely, if the player is not scoring off a penalty corner, the model could determine random effects covering the probability of whether a player is scoring from the left, middle, or right portion of the circle. The main concern of this aspiration is overfitting the model with too many parameters, especially given a relative lack of data, restricting the model's ability to make generalized predictions for a player to score in certain match situations. Having too many parameters puts the model at risk of focusing too much on

irrelevant details of players' situations when predicting their ability to score a goal under different conditions.

References

[1] Czuzoj-Shulman, N., Yu, D., Boucher, C., Bornn, L., & Javan, M. (2019). Winning Is Not Everything: A contextual analysis of hockey face-offs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1902.02397

[2] *glmer: Fitting Generalized Linear Mixed-Effects Models*. (n.d.). RDocumentation. Retrieved December 11, 2023, from https://www.rdocumentation.org/packages/lme4/versions/1.1-35.1/topics/glmer

[3] Han, Y., Kim, J., Ng, H. K. T., & Kim, S. W. (2022). Logistic Regression Model for a Bivariate Binomial Distribution with Applications in Baseball Data Analysis. *Entropy*, *24*(8), 1138. https://doi.org/10.3390/e24081138

[4] Kim, M. (2023). A study on the winning and losing factors of para ice hockey using Data Mining-Based Decision Tree analysis. *Applied Sciences*, *13*(3), 1334. https://doi.org/10.3390/app13031334

[5] Liardi, V. L., & Carron, A. V. (2011). An analysis of National Hockey League face-offs: Implications for the home advantage. *International Journal of Sport and Exercise Psychology*, *9*(2), 102–109. https://doi.org/10.1080/1612197X.2011.567100

[6] *rstan package - RDocumentation*. (n.d.). RDocumentation. Retrieved December 11, 2023, from https://www.rdocumentation.org/packages/rstan/versions/2.32.3

[7] Schuckers, M., Pasquali, T., Curro, J., & Statistical Sports Consulting, LLC. (2012). *An Analysis of NHL Faceoffs* [Undergraduate Honors Theses]. St. Lawrence University.

Appendix A

Glmer Implementation

This appendix consists of the code used to implement the logistic regression model with the glmer package.

```
model <-
  glmer(GOAL~ (1|Player:Shot.after.PC) + match_seconds_bin * Score.Difference + Home + Shot.after.PC + Blocked.or.goal.after.block,
        family="binomial", control = glmerControl(optimizer="bobyqa",
                                                  optCtrl = list(maxfun=2e30)),
      data=nu_shots)
```

Appendix B

RStan Implementation

This appendix consists of the code used to implement the logistic regression model with the RStan package. The implementation with RStan is presented in the top image, while the Stan code used to develop the model itself is shown in the bottom image.

```
new_dat <- list(N=nrow(neu_shots_only), M = length(unique(neu_shots_only$Player)),
                p=17, GOAL=neu_shots_only$GOAL, x=cbind(neu_shots_only$'match_seconds_bin_(0,600]', neu_shots_only$'match_seconds_bin_(600,1.2e+03]',
neu_shots_only$'match_seconds_bin_(1.2e+03,1.8e+03]', neu_shots_only$'match_seconds_bin_(1.8e+03,2.4e+03]',
neu_shots_only$'match_seconds_bin_(2.4e+03,3e+03]', neu_shots_only$'match_seconds_bin_(3e+03,3.6e+03]',
neu_shots_only$'match_seconds_bin_(3.6e+03,4.36e+03]', neu_shots_only$Score.Difference, neu_shots_only$Home,
neu_shots_only$Blocked.or.goal.after.block,  neu_shots_only$'match_seconds_bin_(0,600]'*neu_shots_only$Score.Difference,
neu_shots_only$'match_seconds_bin_(600,1.2e+03]'*neu_shots_only$Score.Difference,
neu_shots_only$'match_seconds_bin_(1.2e+03,1.8e+03]'*neu_shots_only$Score.Difference,
neu_shots_only$'match_seconds_bin_(1.8e+03,2.4e+03]'*neu_shots_only$Score.Difference,
neu_shots_only$'match_seconds_bin_(2.4e+03,3e+03]'*neu_shots_only$Score.Difference,
neu_shots_only$'match_seconds_bin_(3e+03,3.6e+03]'*neu_shots_only$Score.Difference,
neu_shots_only$"match_seconds_bin_(3.6e+03,4.36e+03]"*neu_shots_only$Score.Difference),
                spc=cbind(neu_shots_only$Shot.after.PC),
                g=as.integer(as.factor(neu_shots_only$Player)))
```

```
data {
  // Define variables in data
  // Number of observations (an integer; the number of shots)
  int<lower=0> N;

  // Number of parameters (an integer; the number of predictors, i.e. Score_Difference, Home, etc.)
  // NOT including Shot.after.PC
  int<lower=0> p;

  // Number of groups (an integer; the number of unique players)
  int<lower=0> M;

  // Outcome
  int<lower=0, upper=1> GOAL[N]; // defining the binary goal outcome variable

  // Predictors
  row_vector[p] x[N]; // throwing all predictors into a vector (except spc)
  row_vector[1] spc[N]; // spc by itself to use in estimating random effect

  // Mapping observations (shots) to groups (players)
  int g[N];
}

parameters {
  // Define parameters to estimate
  real alpha;
  vector[p] beta1; // fixed effects, excluding intercept
  vector[1] beta2; // specifically for spc
  matrix[2,M] psi; // random effects, two for each player
  vector<lower=0, upper=10>[2] sigma; // error scale (captures the noise)
}

model {
  // Prior part of Bayesian inference (flat if unspecified)
  alpha ~ normal(0,100);
  psi[1] ~ normal(0,sigma[1]);
  psi[2] ~ normal(0,sigma[2]);
  beta1 ~ normal(0,100);
  beta2 ~ normal(0,100);

  // Likelihood part of Bayesian inference
  for (n in 1:N) {
    GOAL[n] ~ bernoulli_logit(alpha +
    psi[1,g[n]]*(1-spc[n]) + psi[2,g[n]]*spc[n] +
    x[n]*beta1 + spc[n]*beta2);
  }
}
```

Appendix C

Recovered Coefficients from Glmer

This appendix consists of the weights recovered for each coefficient of the logistic regression model using the glmer package. The model was run for 2 x 10^{30} iterations.

```
Fixed effects:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -2.1302     0.6801  -3.132  0.00173 **
match_seconds_bin(600,1.2e+03]                     0.3425     0.7545   0.454  0.64986
match_seconds_bin(1.2e+03,1.8e+03]                 0.8046     0.7576   1.062  0.28820
match_seconds_bin(1.8e+03,2.4e+03]                 0.6982     0.6871   1.016  0.30960
match_seconds_bin(2.4e+03,3e+03]                   1.0188     0.6519   1.563  0.11812
match_seconds_bin(3e+03,3.6e+03]                   0.5159     0.7447   0.693  0.48843
match_seconds_bin(3.6e+03,4.36e+03]              -14.2380  1412.9115  -0.010  0.99196
Score.Difference                                   1.1885     0.9753   1.219  0.22298
Home                                               0.5611     0.3933   1.427  0.15369
Shot.after.PC                                     -0.2258     0.6346  -0.356  0.72198
Blocked.or.goal.after.block                       -1.3918     0.6467  -2.152  0.03138 *
match_seconds_bin(600,1.2e+03]:Score.Difference   -1.7771     1.0785  -1.648  0.09939 .
match_seconds_bin(1.2e+03,1.8e+03]:Score.Difference -1.3352   1.0567  -1.263  0.20642
match_seconds_bin(1.8e+03,2.4e+03]:Score.Difference -1.2019   1.0047  -1.196  0.23156
match_seconds_bin(2.4e+03,3e+03]:Score.Difference  -1.1958     0.9904  -1.207  0.22727
match_seconds_bin(3e+03,3.6e+03]:Score.Difference  -1.7661     1.0018  -1.763  0.07793 .
```

Appendix D

Recovered Coefficients from RStan

This appendix consists of the weights recovered for each coefficient of the logistic regression model using the RStan package. The model was sampled 3,000 times over 5 chains.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	-9.35	2.22	36.03	-80.05	-34.11	-9.20	15.70	61.36	263	1.02
beta1[1]	6.61	2.23	36.05	-64.42	-18.52	6.46	31.21	77.45	262	1.02
beta1[2]	6.97	2.23	36.06	-63.63	-18.17	6.85	31.80	77.75	262	1.02
beta1[3]	7.58	2.23	36.07	-63.09	-17.62	7.46	32.37	78.50	262	1.02
beta1[4]	7.43	2.23	36.05	-63.09	-17.76	7.35	32.18	78.32	262	1.02
beta1[5]	7.84	2.23	36.05	-62.84	-17.36	7.81	32.62	78.64	262	1.02
beta1[6]	7.14	2.23	36.06	-63.51	-17.97	7.09	32.02	78.10	262	1.02
beta1[7]	-77.38	1.81	65.97	-226.05	-117.08	-69.33	-29.90	31.58	1331	1.00
beta1[8]	-1.45	2.07	37.77	-72.77	-27.42	-2.71	23.05	74.64	334	1.01
beta1[9]	0.72	0.01	0.45	-0.14	0.41	0.72	1.02	1.65	1871	1.00
beta1[10]	-1.84	0.02	0.77	-3.49	-2.32	-1.79	-1.30	-0.49	1634	1.01
beta1[11]	2.72	2.07	37.76	-73.04	-21.75	3.99	28.63	73.92	334	1.01
beta1[12]	0.76	2.07	37.77	-75.79	-23.84	2.07	26.74	72.03	333	1.01
beta1[13]	1.29	2.07	37.77	-74.56	-23.13	2.52	27.28	72.55	334	1.01
beta1[14]	1.45	2.07	37.76	-74.69	-23.03	2.71	27.48	72.70	334	1.01
beta1[15]	1.42	2.07	37.77	-74.64	-23.21	2.71	27.36	72.61	333	1.01
beta1[16]	0.75	2.07	37.77	-75.33	-23.84	2.01	26.70	71.82	334	1.01
beta1[17]	2.51	1.88	99.61	-195.03	-63.83	2.08	70.03	195.18	2811	1.00
beta2[1]	-0.19	0.10	1.27	-2.96	-0.82	-0.19	0.44	2.37	147	1.03

Appendix E

Random Effect Coefficients with Glmer

In this section, the random effects from glmer, interacted with Shot.after.PC, are presented. 0 represents the random effect for a player based on their ability to score goals not after a penalty corner, while 1 represents their ability to score shots off a penalty corner.

```
$`Player:Shot.after.PC`
                     (Intercept)
Player 1:0          0.606041250
Player 1:1         -0.008553571
Player 2:0          0.296860216
Player 2:1          0.555044449
Player 3:0         -0.701968352
Player 4:0          0.850710398
Player 5:0         -0.475082345
Player 5:1         -0.188451199
Player 6:0         -0.059494820
Player 6:1         -0.676104742
Player 7:0         -0.885961605
Player 7:1          0.631003393
Player 8:0         -0.301999995
Player 9:0         -0.652326202
Player 9:1         -0.122669872
Player 10:0        -0.097304192
Player 11:0        -0.112828236
Player 12:0         0.184766995
Player 12:1        -0.472281410
Player 13:0        -0.194821008
Player 13:1        -0.376869431
Player 14:0         1.137924599
Player 15:0         1.262636480
Player 15:1        -0.027334758
Player 16:0        -0.754930891
Player 16:1         0.732862746
Player 17:0         0.232972396
Player 17:1         0.317820482
Player 18:0        -0.368638665
Player 18:1        -0.183646308
Player 19:0         0.033443825
Player 19:1        -0.180819844
```

Appendix F

Random Effect Coefficients with RStan

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
psi[1,1]	1.74	0.12	2.12	-1.19	0.26	1.29	2.77	7.01	332	1.02
psi[1,2]	0.62	0.05	0.81	-0.79	0.06	0.53	1.07	2.42	307	1.01
psi[1,3]	-1.57	0.10	1.91	-6.20	-2.30	-1.17	-0.33	0.78	340	1.02
psi[1,4]	2.35	0.14	2.34	-0.82	0.58	1.91	3.62	7.93	292	1.02
psi[1,5]	-1.19	0.06	1.71	-5.51	-1.97	-0.86	-0.06	1.30	800	1.01
psi[1,6]	0.15	0.04	0.96	-1.73	-0.44	0.10	0.71	2.20	488	1.00
psi[1,7]	-1.77	0.08	1.83	-6.31	-2.56	-1.41	-0.54	0.56	474	1.02
psi[1,8]	-0.89	0.07	1.84	-5.55	-1.72	-0.57	0.18	2.14	676	1.01
psi[1,9]	-1.40	0.08	1.72	-5.86	-2.19	-1.06	-0.22	1.01	409	1.02
psi[1,10]	-0.46	0.06	1.91	-4.94	-1.34	-0.22	0.60	2.93	1177	1.00
psi[1,11]	-0.46	0.06	1.92	-4.89	-1.29	-0.22	0.57	2.88	992	1.00
psi[1,12]	0.44	0.05	0.92	-1.27	-0.13	0.34	0.96	2.44	403	1.01
psi[1,13]	-0.02	0.04	0.86	-1.68	-0.54	-0.04	0.46	1.87	443	1.01
psi[1,14]	1.80	0.08	1.20	-0.10	0.92	1.71	2.55	4.45	253	1.02
psi[1,15]	2.07	0.08	1.30	-0.02	1.12	1.96	2.89	4.93	250	1.03
psi[1,16]	-1.01	0.04	1.09	-3.50	-1.65	-0.90	-0.26	0.84	771	1.01
psi[1,17]	0.55	0.05	1.29	-1.81	-0.23	0.42	1.26	3.40	799	1.01
psi[1,18]	-0.97	0.07	1.84	-5.42	-1.73	-0.63	0.12	1.70	682	1.02
psi[1,19]	0.33	0.05	0.81	-1.08	-0.18	0.24	0.77	2.19	298	1.01
psi[2,1]	-0.35	0.17	2.25	-5.88	-0.98	-0.07	0.62	3.48	175	1.04
psi[2,2]	1.50	0.25	2.49	-1.22	0.04	0.78	2.16	8.61	97	1.08
psi[2,3]	0.14	0.17	2.58	-5.06	-0.75	0.01	0.82	6.81	237	1.01
psi[2,4]	0.23	0.22	2.55	-4.29	-0.72	0.03	0.91	6.74	136	1.04
psi[2,5]	-0.72	0.16	2.00	-6.10	-1.36	-0.30	0.28	2.26	161	1.04
psi[2,6]	-0.79	0.10	1.24	-3.64	-1.43	-0.59	-0.05	1.39	168	1.03
psi[2,7]	1.65	0.28	2.50	-1.05	0.10	0.90	2.38	8.53	78	1.08
psi[2,8]	-0.09	0.16	2.54	-5.67	-0.87	-0.02	0.79	4.98	257	1.00
psi[2,9]	-0.63	0.17	2.31	-6.91	-1.13	-0.19	0.41	2.49	192	1.03
psi[2,10]	0.13	0.11	2.32	-4.66	-0.75	0.02	0.91	5.44	468	1.00
psi[2,11]	0.04	0.14	2.48	-5.23	-0.80	0.02	0.84	5.53	304	1.01
psi[2,12]	-1.26	0.27	2.35	-8.13	-1.81	-0.61	0.02	1.40	73	1.08
psi[2,13]	-0.23	0.08	1.05	-2.26	-0.74	-0.22	0.20	2.31	166	1.04
psi[2,14]	0.04	0.12	2.37	-5.55	-0.74	0.03	0.86	4.96	425	1.01
psi[2,15]	-0.43	0.16	2.31	-6.72	-0.99	-0.10	0.58	3.20	221	1.03
psi[2,16]	0.99	0.09	1.19	-0.90	0.22	0.83	1.55	4.04	169	1.04
psi[2,17]	0.53	0.09	1.35	-2.02	-0.20	0.33	1.18	3.77	228	1.03
psi[2,18]	-0.79	0.23	2.27	-7.28	-1.34	-0.30	0.29	2.29	101	1.05
psi[2,19]	-0.81	0.25	2.36	-7.89	-1.29	-0.24	0.33	2.12	93	1.06

In this appendix, the random effects from RStan, interacted with Shot.after.PC, are presented. The "mean" column lists players' random effects, which also interact with Shot.after.PC. Listed in the first column are the random effect coefficients. The first bracketed value after "psi" represents whether the random effect corresponds to a player's ability to score after a penalty corner (1 for shots not made after penalty corners and 2 for shots made after penalty corners) while the second bracketed value indicates a player, each of whom is represented with the same number in the glmer model (e.g., psi [1, 8] represents Player 8's ability to score not after penalty corners)