Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models

Lore Zumeta-Olaskoaga[1,2], Maximilian Weigert[3], Jon Larruskain[4], Eder Bikandi[4], Igor Setuain[5], Josean Lekue[4], Helmut Küchenhoff[3] and Dae-Jin Lee[1]

[1]BCAM - Basque Center for Applied Mathematics, Spain

[2]Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Spain

[3]Statistical Consulting Unit StaBLab, Ludwig-Maximilians Universität München, Germany

[4]Medical Services, Athletic Club, Spain

[5]Department of Health Sciences, Universidad Pública de Navarra, Spain

Abstract

Data-based methods and statistical models are given special attention to the study of sports injuries to gain in-depth understanding of its risk factors and mechanisms. The objective of this work is to evaluate the use of shared frailty Cox models for the prediction of occurring sports injuries, and to compare their performance with different sets of variables selected by several regularized variable selection approaches. The study is motivated by specific characteristics commonly found for sports injury data, that usually include reduced sample size and even fewer number of injuries, coupled with a large number of potentially influential variables. Hence, we conduct a simulation study to address these statistical challenges and to explore regularized Cox model strategies together with shared frailty models in different controlled situations. We show that predictive performance greatly improves as more player observations are available. Methods that result in sparse models and favour interpretability, e.g. best subset selection and boosting, are preferred when the sample size is small. We include a real case study of injuries of female football players of a Spanish football club.

Keywords— shared frailty models, regularized Cox methods, sports injury prevention, survival analysis

1 Introduction

Sports injuries are "complex" phenomena, as recent views in sports medicine and injury prevention suggest (Bolling et al., 2018). They result from the dynamic interaction of multiple risk factors and have serious consequences on the athletes' health. In this sense, vast efforts are directed to study the underpinning mechanisms of injuries, since some may be addressed with injury prevention programmes. That being so, the use of adequate statistical models could help medical staff in monitoring the athletes' health status by providing predictions on the risk of injury.

In recent years, an increasing tendency to use data-based analysis methods in the field of sports injury prevention can be noticed. An overview of existing strategies to monitor and predict occurrence and duration of sports injuries comprising classical statistical and machine learning models

*Corresponding author: dlee@bcamath.org

Accepted in AStA Advances in Statistical Analysis, 5 November 2021

is given by Ruddy et al. (2019) –who also mention several limitations due to the specific characteristics of sports injury data. A proper statistical model for such a purpose should encompass the complex time-varying and recurrent nature of injuries: a player's injury susceptibility may change over time, and moreover, a player can sustain more than one injury, as subsequent injuries are often influenced by previous ones (Hägglund et al., 2006; De Visser et al., 2012). On this basis, models for recurrent events are appealing for sports injuries prevention. Such models have been broadly used in many biomedical studies, but their application has insufficiently been exploited in sports injury research (Nielsen et al., 2016). Shared frailty Cox models are useful for analysing the association of exposure variables to injury risk over time, while also handling the possible dependency of recurrent injuries (Nielsen et al., 2019; Ullah et al., 2014). Applications of frailty models in the field of sports injury include studies that identify risk factors for contact injuries, accounting for recurrence, in professional rugby league players (Gabbett et al., 2012); analyse the training load and shoulder injuries in a large youth handball cohort (Møller et al., 2017); study the genetic association with hamstring injuries in soccer players (Larruskain et al., 2018). Also recently, several machine learning approaches have been applied, mainly based on classification techniques where a binary outcome (injured/non-injured) is predicted (e.g. Rossi et al., 2018). While machine learning methods are very appealing and a powerful tool in many applications, they usually require large sample size for training and hyperparameter tuning. Besides, most classical machine learning methods do not explicitly account for recurrent events or easily deal with imbalance classes (i.e. very few injured players).

Another important aspect to account for in sports injury data analysis is the screening tests made for monitoring the players' health status. A high number of functional tests is widely used in professional sports teams for injury prevention, rehabilitation or fitness conditioning. And the number of these functional screening tests requires special attention in the modelling process, mainly when the interest –from a practical perspective– lies on sparse and interpretable models. In addition, when time-to-event analysis is considered for a large number of predictors, the estimation of frailty terms is likely to undergo convergence problems when the number of parameters to be estimated considerably increases (McGilchrist and Aisbett, 1991; Therneau et al., 2003). This may especially be problematic for small sample data, as it often is the case for sports injury data. Thus, it is desirable to reduce the number of parameters to be estimated, and efficiently select a subset of relevant variables the risk of injury is associated with. Further, data are often limited to individual teams or a small sample of individual players, and from a statistical point of view, the total number of injuries is usually small.

In this study, we focus on lower-limb injuries that frequently occur in women football –one of the fastest growing sports worldwide. Lower-limb injuries are of great concern due to the severity of some of them, and given their high incidence in women football players (Crossley et al., 2020). During a regular season, the medical staff (which includes medical doctors, physiotherapists, strength and conditioning coaches etc.) conduct regular screening tests intended to identify players predisposed to injury, and consequently, to optimize player's (and team's) performance and increase their safety. They consist of a series of medical evaluations such as functional movement tests that assess biomechanical factors and muscle imbalance. In these assessments, in football players it is frequent to quantify inter-limb asymmetries. The asymmetry measurements may help to identify players who are at increased risk of lower-limb injuries. Some studies have shown that bilateral strength asymmetry may be a risk factor for musculoskeletal injury (Croisier et al., 2002, 2003; Knapik et al., 1991; Hewett et al., 2005). However, the overall scientific level of evidence for the screening tests remains scarce (see McCall et al., 2015; Bahr, 2016, for a review).

The aim of this paper is to study the adequacy and performance of a family of statistical methods for time-to-event data analysis based on regularization techniques and Cox regression in the context of sports injury data. We aim to compare the performance of frailty models that include different sets of previously selected variables, with respect to prediction accuracy. Due to limitations caused by the outlined characteristics of sports injury data, we explore several hypothetical controlled situations in a simulation study. Our work consists of two major components, (i) a real data example, containing information about a single team with 22 players, where the performance of the different approaches is compared, and (ii) a simulation framework where all considered variable selection methods are systematically evaluated for three different scenarios and varying size of the data.

In the following, in Section 2, we present the data that motivated the study and methods considered, including different regularized Cox models to perform variable selection and shared

frailty models to fit the data with a reduced number of variables. We also describe the simulation study carried out to evaluate the models' performance and accuracy. In Section 3, we present the results obtained from the case study data and from the simulation scenarios for various sample sizes and different dependence structure for the covariates effects. Finally, in Section 4, we conclude with a general discussion.

2 Data and Methods

2.1 Screening tests and lower-limb injury data

A female football team with 22 players, prospectively followed-up during the 2017-2018 season, is analysed. The data contain players' exposure, i.e. time spent training and competing in minutes, as well as time-loss non-contact lower-limb injuries that were recorded by the club's medical staff. Lower-limb non-contact injuries were captured when a player was unable to participate in a future training session or match due to a physical complaint resulting from football training or match play, and was considered injured until the medical staff cleared the player for full participation in training and match play (Fuller et al., 2006). During the season, players completed three times, biomechanical and functional conditioning screening tests. A total of 28 variables were pre-selected by the medical experts' criteria, out of 200 variables measured in each of these screening tests, see Supplementary Material Table 1. These variables gathered anthropometrics data and further values of biomechanical functional tests, assessed as bilateral strength asymmetries of the lower limbs based on Impellizzeri et al. (2007).

Figure 1 shows a complete picture of the injuries sustained by each player, days lost until the complete recovery and return to competition, together with the screening tests completed by each player. The information recorded on the third series of screening tests was not used, since the players' follow-up ended at that moment. A total number of 12 players suffered a lower-limb injury from a total of 19 injuries that occurred in the team: 7 players were injured once, 4 players twice and one player four times, while 45% of players never got injured. The median exposure time of a player was 13,302 minutes and the total exposure time of the team roughly 250,000 minutes. The team injury incidence and injury burden (i.e. the total number of days lost to injury per 1,000 hours of exposure) were 4.56 injuries and 178.67 days lost due to injury, respectively.

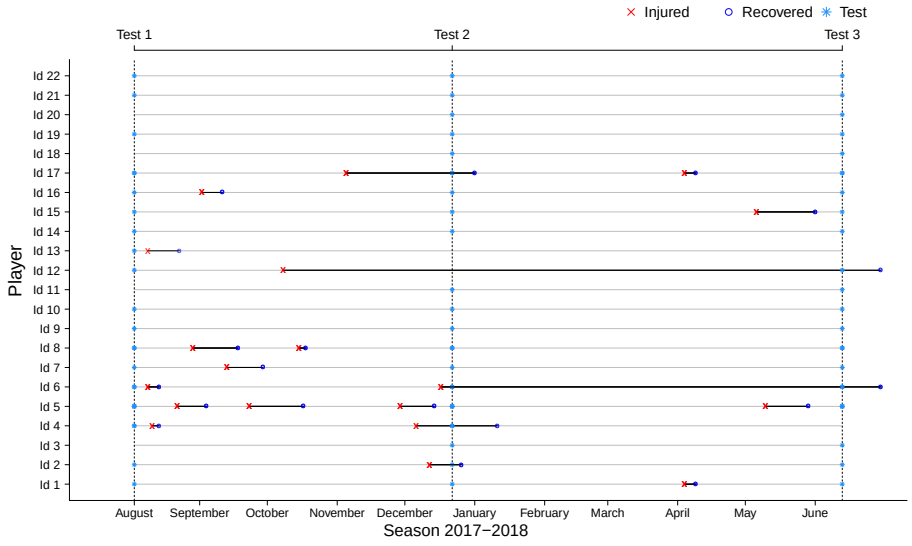

Figure 1: Overview of the screening tests and lower-limb injury data. The timeline of the football team players' is depicted horizontally, the red cross indicates the exact moment of the injury occurrence, the blue circle the moment when the player is totally recovered, the bold black line the time lost due to injury, and the three vertical lines correspond to the moment when the screening tests were performed

2.2 The models

Regularized Cox models and shared frailty models are applied to the screening tests and lower-limb injury data. A two-step strategy is followed, motivated by the large number of potential covariates present in the data. In a first step, different variable selection techniques based on regularized Cox models are applied which do not explicitly account for repeated measures. Secondly, the most relevant variables, i.e. sets containing a reduced number of variables selected by each of the methods in the previous step, are used to fit shared frailty models. Following, the notation for the Cox model and regularized Cox methods is first introduced. Afterwards, in the "Frailty models" section, the notation is extended to the case of recurrent events.

The primary outcome variable is defined as a player's exposure time in minutes until the player gets injured, denoted as a non-negative random variable T. Players' follow-up started at the beginning of the season, the time the first screening test was conducted, and continued until mid-season. Having collected new covariates at this moment, the time origin of the primary outcome is set again to 0. Hence, when a player does not get injured in the first or second half of the season, the exposure time is set the time of the second screening –mid-season– or/and the end of the season, respectively, and considered censored. In the following, the censorship is defined as a random variable C.

The observed data are then composed by the set $\{(Y_i, \delta_i, X_i),\ i = 1, \ldots, N\}$, where $Y_i = \min\{T_i, C_i\}$ and $\delta_i = \mathbb{I}\{T_i \leq C_i\}$ is the censorship indicator and N the total number of observations. We assume that censoring is non-informative and that given $\boldsymbol{x}_i$, y_i and δ_i are independent.

The Cox model

All models used throughout this paper are based on the Cox proportional hazards model (Cox, 1972, 1975). Focused on sports injury data, the model assumes that hazard of a player i being injured is related to covariates through

$$\lambda_i(t|\boldsymbol{x_i}) = \lambda_0(t)\exp(\boldsymbol{x}_i\boldsymbol{\beta}), \qquad i = 1, \ldots, N,$$

where $\lambda_0(t)$ is a non-parametric baseline hazard function and $\boldsymbol{\beta}$ is a vector of regression parameters that relates the vector of covariates $\boldsymbol{x}_i$. The estimation of $\boldsymbol{\beta}$ regression coefficients is performed by maximising Cox partial likelihood (Cox, 1975), $pL(\boldsymbol{\beta})$:

$$pL(\boldsymbol{\beta}) = \prod_{i=1}^{N}\left(\frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{l\in R(Y_i)}\exp(\boldsymbol{x}_l\boldsymbol{\beta})}\right)^{\delta_i},$$

where $R(Y_i)$ is the risk set, set of players at risk at the moment y_i, i.e. $R(Y_i) = \{l : Y_l \geq y_i\}$.

Regularized Cox methods

Six different regularized Cox models are studied, namely, Best Subset Selection (BeSS) (Wen et al., 2020), Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, 1997), Elastic Net (Zou and Hastie, 2005), Ridge regression (Hoerl and Kennard, 1976), Group Lasso (Yuan and Lin, 2006) and Boosting in Cox regression (Bühlmann et al., 2007). Except for the latter, estimation of these models, in the context of survival analysis, is performed maximising the penalized Cox partial log-likelihood (Cox, 1972, 1975). That is,

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_0 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}\mathbb{I}\{\beta_j \neq 0\}) \qquad \text{(Best Subset Selection)}$$

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}|\beta_j| \qquad \text{(Lasso regression)}$$

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_2^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\left(\sum_{j=1}^{p}\beta_j^2\right) \qquad \text{(Ridge regression)}$$

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\left((1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right) \qquad \text{(Elastic Net)}$$

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\sum_{g=1}^{G}\|\boldsymbol{\beta}_l\|_2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, pl(\boldsymbol{\beta}) - \lambda\sum_{g=1}^{G}\sqrt{(\beta_1^2 + \ldots + \beta_{n_g}^2)} \qquad \text{(Group Lasso)}$$

where $\lambda \geq 0$ and $\alpha \in (0, 1)$ are the regularization tuning parameters and $pl(\boldsymbol{\beta})$ is the Cox partial log-likelihood to be maximized subject to a constraint, a penalty function to be multiple of a L_1 or L_2-norm, or a L_0-seminorm. For Group Lasso, the vector of coefficients is partitioned in G groups of size n_g, i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_G')'$. The type of test is used as grouping factor, and all these G groups are equally penalized, see Supplementary Material Table 1. For the sake of simplicity, Elastic Net with $\alpha = 0.5$ is considered only. The best regularization parameter λ is estimated by 10-fold cross-validation, for which the same cross-validation splits are used across all models to enable a fair comparison of their performance. It is worth noticing that, although the Ridge regression technique itself is not a variable selection method, it is included as a regularized method for the comparisons. Hence, the estimated coefficients' 95% confidence intervals are generated via bootstrap, and it is checked whether the interval includes zero or not. The variables are assumed to be selected when zero is not included in their corresponding coefficients' 95% confidence interval (Chatterjee and Lahiri, 2010; Sartori, 2011).

The sixth regularization method considered, a Boosting approach in Cox regression, relies on a rather different idea. Instead of directly optimizing the penalized likelihood, coefficients are obtained via an iterative process. For the scope of this study, we focus on likelihood-based boosting (Tutz and Binder, 2006). The negative partial log likelihood is used as a loss function $f(\cdot)$ in the negative gradient algorithm –or L_2-Boosting. The algorithm results in refitting residuals multiple times, so that the solution of the partial log likelihood is updated by a small factor in each boosting iteration. Regularization is implicitly achieved through early stopping the algorithm, and through updating a single coefficient in each iteration variable selection is performed. The number of boosting iterations m_{stop}, the tuning parameter, is selected via a 10-fold cross validation.

Frailty models

The occurrence of non-contact lower-limb injuries is fitted by shared frailty Cox models (Hougaard, 1995; McGilchrist and Aisbett, 1991). Such a model considers the dependence, that observations within the same player possibly share, by including a subject-specific random effect that acts on the baseline hazard in a multiplicative way. The frailty term aims to account for unobserved heterogeneity, since observations within each player might be correlated and individual characteristics (variables that make individual players different from one another) might often be unobserved –and might sometimes be unmeasurable.

Now, let specify the total number of players by K, where the k-th player has n_k observations indexed by j, so that the repeated measures are explicitly accounted for the data observed, i.e. $\{(Y_{jk}, \delta_{jk}, X_{jk}),\ k = 1, \ldots, K \text{ and } j = 1, \ldots, n_k\}$, being N the total number of observations, the sum of each player's number of observations $N = \sum_{k=1}^{K} n_k$. For Y_{jk} recurrent events the so-called gap time approach (Kelly and Lim, 2000; Ullah et al., 2014) is considered. This determines the risk interval of each player, in such a way that a new risk interval is set every time the player has totally recovered from a sustained injury and starts to train. Thus, each recurrent event is represented by a separate interval and each recurrent episode is "at-risk" from the starting point of the previous event recovery, which is reset to time 0, see Figure 2.

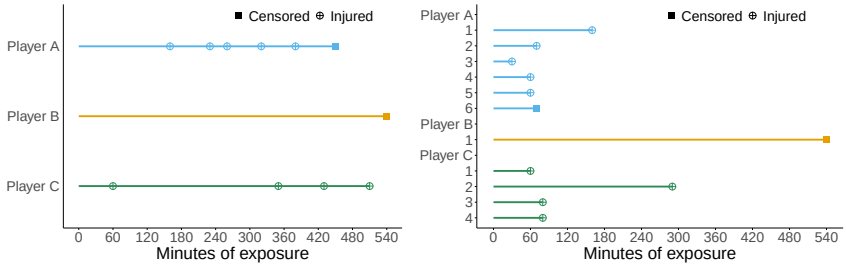

Figure 2: Illustration of a risk interval formulation. Left a general case study; right the gap time approach. Each time to an event or censoring is a separate risk interval

Technically, each observation of the data set corresponds to a single player: some players had not got injured at all, during the follow-up, and contributed to censored survival times; others, sustained at least an injury and are thus represented by one or multiple survival times.

The hazard rate of injury at time t, for the j-th observation of the k-th player, is given by:

$$\begin{aligned}\lambda_{jk}(t|\alpha_k, \boldsymbol{x_{jk}}) &= \alpha_k \lambda_0(t) \exp(\boldsymbol{x_{jk}}\boldsymbol{\beta}) = \\ &= \lambda_0(t) \exp(\boldsymbol{x_{jk}}\boldsymbol{\beta} + \boldsymbol{z_{jk}} b_k), \qquad j = 1, \ldots, n_k,\ k = 1, \ldots, K,\end{aligned} \tag{1}$$

where λ_0 is the unspecified baseline hazard, p the number of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ the vector of coefficients, $\boldsymbol{x_{jk}}$ the corresponding row of this observation in the design matrix $\boldsymbol{X}$and α_k or $b_k = \ln(\alpha_k)$, the player's frailty term, where matrix $\boldsymbol{Z}$ is a sparse matrix $N \times K$ such that $\boldsymbol{z}_{jk} = 1$, when j-th observation corresponds to player k and 0 otherwise. Penalized partial likelihood is used to estimate the regression coefficients and the frailty terms (Ripatti and Palmgren, 2000). As stated by Gasparini et al. (2019), the choice of a particular parametric frailty distribution has minimal impact on the estimation and testing of regression coefficients. Here, we assume the log-normal distribution for the frailty term α_k, i.e. the Gaussian distribution for $b_k = \ln(\alpha_k)$, following that the models fitted to real data gave best fits with this distribution according to the Akaike information criterion.

Equivalently, the marginal survival function, i.e. the probability of a player not sustaining an injury at time t, given the covariates, can be derived from Equation (1) and integrating out the frailty term from the conditional survival probability,

$$\begin{aligned}S(t|\boldsymbol{x}) &= \int_{-\infty}^{\infty} S(t|\boldsymbol{b}, \boldsymbol{x}) g(\boldsymbol{b})\, \mathrm{d}\boldsymbol{b} = \\ &= \int_{-\infty}^{\infty} S_0(t)^{\exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b})} g(\boldsymbol{b})\, \mathrm{d}\boldsymbol{b},\end{aligned} \tag{2}$$

where $\boldsymbol{b} = (b_1, ..., b_K)$ represents the vector of the frailties, which follows a Gaussian distribution, and $g(\cdot)$ its density function.

Based on the survival function, it is possible to predict players' injury probabilities for times $t > 0$. In particular, the marginal approach of the survival function, i.e. a population-averaged probability, Equation (2), includes predictions for new players which have not been part of the data used to fit the model. Contrary to the conditional survival probability approach, that do not allow estimating predictions of new players –in this case player-specific predictions– since their frailties are unknown. The chosen gap time approach for recurrent events allows predictions to be made for future survival times after injury, constrained to the maximum event time followed in the data used to fit the model.

Evaluation of frailty models

The predictive performance of frailty models is assessed through the Brier score (BS) and the Integrated Brier Score (IBS), i.e. the area under the BS curve (Gerds and Schumacher, 2006; Graf et al., 1999). The choice of adequate performance measures is an essential task for a valid comparison of different models. The BS is a time-dependent predictive measure to calculate a model's overall performance (Steyerberg et al., 2010). It is commonly used in survival analysis since it copes with the point that risk prediction in this field is made in terms of probabilities.

Formally, the BS at time point t is a weighted mean squared error between predicted survival probability and observed survival status. Besides, inverse probability of censoring weighting (IPCW) (Gerds and Schumacher, 2006) is used to account for observations under risk, regardless they are eventually censored or not, and thus to make use of all available. Let $G(t) = P(C < t)$ be the censoring distribution and N be the total number of observations. Then, the BS is formulated as

$$\mathrm{BS}(t|\hat{S}(t|x)) = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \frac{(0 - \hat{S}(t|x_i))^2}{\hat{G}(t_i)} & t_i \leq t,\ \delta_i = 1 \\ \frac{(1 - \hat{S}(t|x_i))^2}{\hat{G}(t)} & t_i > t \\ 0 & t_i \leq t,\ \delta_i = 0 \end{cases} \tag{3}$$

The BS has a range between 0 and 1, with smaller values corresponding to a better prediction. Then, the IBS can be calculated, following Equation (3), as an overall measure of the model performance at all available time points:

$$\mathrm{IBS}(\mathrm{BS(t)}, \tau) = \frac{1}{\tau} \int_0^{\tau} \mathrm{BS}(u, \hat{S}) du$$

where $\tau = t_{\max}$, or $0 < \tau < t_{\max}$.

Due to the lack of external validation data and to avoid overfitting, the so-called "bootstrap .632+ approach" (Efron and Tibshirani, 1997) is used which has been demonstrated to lead to more accurate estimations (Binder and Schumacher, 2008) and balances both the apparent and the bootstrap BS estimate. See section B in the Supplementary Material for more detail on this estimate.

2.3 Simulation framework

Our simulation study aims to evaluate the applicability and robustness of the outlined statistical approaches, by establishing three hypothetical controlled situations that may stem in the context of sports injury data. Apart from assessing model performance in these scenarios, the impact of varying sample sizes is of particular interest.

The simulation procedure can be summed up in the following three steps, following in part the structured strategy proposed by Morris et al. (2019) to the planning of simulation studies:

Step 1: Generation of the data

The underlying data generating process is designed to mimic the original sports injury data as closely as possible, i.e. time-to-event outcome variable, highly right-skewed, many censored observations and a high number of covariates. Three different scenarios are considered: (i) augmenting the original case study, of screening tests and lower-limb injury data, by bootstrap re-sampling and adding a random noise; and generating time-to-event observations that arise from covariates that share (ii) a weak correlation, and lastly, from covariates with (iii) a high correlation. The three scenarios considered are explained in detail in the following specific subsection *"Parameters defining simulation setting"*.

In this regard, a modified version of the random spline method proposed by Harden and Kropko (2019) is used to produce the true underlying data generating process. This method does not assume any distributional form for the baseline hazard function, thus it matches the Cox model's inherent flexibility. It requires to initially determine the number of points –or knots– to be drawn to fit a cubic spline for the baseline hazard function. To represent the recurrent nature of sports injury data, the method is modified in a way to include a multiplicative random effect, i.e. a frailty term.

Step 2: Fitting the models

The six regularized Cox methods described in the previous Section 2.2 are fitted to the data to preselect small sets of variables. Afterwards, for each of the six sets of selected variables shared frailty Cox models (1) are fitted to the data, accounting for the players' unobserved variability. It is assumed that the frailty term follows a Gaussian distribution, in accordance with the real case study.

Step 3: Performance Measures

The previous two steps are repeated N_{sim} times, for a given prespecified configuration. Finally, the models are assessed by a number of different measures that evaluate both, (i) the model performance and (ii) the predictive accuracy of the final shared frailty models. On the one hand, we compare the model performance by assessing how well the estimated models represent the underlying true model. Thus, the selection of significant variables is evaluated through the measures presented in Table 1; and additionally, differences in estimated and true coefficients are measured by the mean squared error (MSE), defined as,

$$\text{MSE} = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \sum_{j=1}^{p} (\hat{\beta}_j^{(i)} - \beta_j)^2.$$

Table 1: Summary of the measures used to evaluate the performance of variable selection methods. The optimal value in the first column refers to the value one would obtain if the variable selection method always found the correct model. When the value is more than one –for average model size– it refers to each one of the settings.

Measure (Optimal value)	**Description (Abbreviation)**
Average model size (2,6,4,5,5,5,5,5,5)	The average number of variables included in model (AMS).
Average number of falsely selected variables (0)	The average number of variables incorrectly selected (ANFS).
Average number of falsely non-selected variables (0)	The average number of incorrectly excluded variables, i.e. variables that really have an effect and their corresponding coefficient is estimated as zero (ANFNS).

On the other hand, the predictive accuracy of the shared frailty models is assessed by means of the measures presented in Section 2.2, i.e. the BS and the IBS. This evaluation is repeated for each N_{sim} replica. To summarize the overall predictive accuracy, the medians of the IBS in the [0,1000] and [0,3500] minutes time intervals are reported.

Parameters defining simulation settings

In this section, the three scenarios considered are described in detail including the parameters that define the scenarios. See Table 2 and Table 3 for a summary. Table 2 shows parameters that are fixed throughout the settings, and Table 3 parameters that are specific for each setting, i.e. parameters that are allowed to vary across the settings such as the true vector of coefficients, true vector of frailties, number of players and number of observations per each.

The first scenario is designed according to the results obtained from the real data analysis. Three different settings are considered based on the estimated coefficients obtained from these data. In the first setting, the vector of coefficients $\boldsymbol{\beta}$, that generates the underlying true data, is fixed to be the vector of coefficients estimated by the frailty model based on BeSS selected variables. The second and third settings are constructed in the same way. The second setting uses the vector of coefficients $\boldsymbol{\beta}$ obtained by the frailty model based on Lasso, Elastic Net and Ridge regression; whereas the third setting uses the one obtained by the frailty model based on Cox Boosting selected variables. Further, it is assumed that the data consists of 66 players, i.e. three average-sized teams, $K = 66$, each with three repeated observations, $n_k = 3$ for all $k = 1, \ldots, K$ and $p = 28$ variables. The design matrix of the real data is augmented through resampled rows, by drawing bootstrap samples with replacement and repeating each sampled value three times with an added random noise. The grouping vector for performing Group Lasso is kept the same as in real data, which was determined concerning the type of screening test.

Conversely, the second scenario and third scenario's design matrices are generated from equally distributed normal variables, the vector of coefficients $\boldsymbol{\beta}$ is set to be the same in both scenarios (see Table 3) and the frailty term is assumed to follow a normal distribution centered at 0 and standard deviation of 0.3. The scenarios differ in the correlation structure among the variables. Scenario 2 assumes independent variables, whereas Scenario 3 considers a pairwise correlation between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, $\rho_{i,j}$ to be $0.65^{|i-j|}$. Within both scenarios, four different sample sizes are regarded, varying the number of players K, $K \in \{22, 66, 132, 220\}$ players –or 1, 3, 6 and 10 football teams with an average number of 22 players– each with a different random number of observations of $p = 50$ variables. The number of observations per each player is generated following a truncated Poisson distribution with mean 3, that gives rise to a total number of 60, 191, 391 and 670 number of observations. Note that the number of observations per each player and the vector of frailties is the same among the settings of the second and third scenarios. The grouping vector for performing the Group Lasso is set to be ten groups of five variables each, i.e. $G = 10$ and $n_g = 5$.

Table 2: Fixed parameters of the simulation study.

Parameter	**Value**
Number of simulated data (N_{sim})	100
Maximum observed time ($T_{\max}$)	4000
Censorship	75%
Frailty distribution	Gaussian
Knots	500

Table 3: Parameter settings for each scenario of the simulation study.

Scenario	**Vector of coefficients** $\boldsymbol{\beta}$	**Frailty term** $\boldsymbol{\alpha_k}$	**Sample size** $N_{\text{obs}} = \sum_{k=1}^{K} n_k$
Scenario 1			
True model: frailty (BeSS)	$\beta_6 = -0.754, \beta_{15} = 1.034,$ otherwise $\beta_l = 0$	Estimated frailties in frailty (BeSS)	198
True model: frailty (Lasso, Elastic Net, Ridge)	$\beta_5 = 0.175,\ \beta_6 = -0.731,$ $\beta_{10} = -0.825,\ \beta_{12} = 0.155,$ $\beta_{15} = 0.424,\ \beta_{24} = -0.580$ otherwise $\beta_l = 0$	Estimated frailties in frailty (Lasso)	198
True model: frailty (Boosting)	$\beta_6 = -1.048,\ \beta_{10} = -0.552,$ $\beta_{12} = 0.076,\ \beta_{15} = 0.990,$ otherwise $\beta_l = 0$	Estimated frailties in frailty (Boosting)	198
Scenario 2	$\beta_1 = 0.4, \beta_2 = 0.2, \beta_3 = 0.2,$ $\beta_4 = 0.2, \beta_5 = 0.2,$ otherwise $\beta_l = 0$	$\sim N(0, 0.3^2)$	60, 191, 391, 670
Scenario 3	$\beta_1 = 0.4, \beta_2 = 0.2, \beta_3 = 0.2,$ $\beta_4 = 0.2, \beta_5 = 0.2,$ otherwise $\beta_l = 0$	$\sim N(0, 0.3^2)$	60, 191, 391, 670

2.4 Software issues

All computations are performed with the open source statistical software R 3.6.2 (R Core Team, 2019), in a 64-bit linux platform with an Intel Core 2.2 GHz CPU of 2 cores and 243.4 GByte RAM. The six regularized methods are implemented through `BeSS` (Wen et al., 2020), `glmnet` (Friedman et al., 2010), `grpreg` (Breheny and Huang, 2015) and `CoxBoost` (Binder, 2013) packages. Shared frailty models are fitted using the `coxph` function of the `survival` (Therneau, 2020) package and `pec` (Mogensen et al., 2012) is used for the BS and IBS computations. The code for the simulations is available in the following GitHub repository: `https://github.com/lzumeta/TimeToEvent-InjurySim`.

3 Results

3.1 Results of the screening tests and lower-limb injury data

The results derived from the variable selection techniques emphasize the distinct characteristics of each method. Group Lasso tends to select more variables –all variables within the same group– whereas all other techniques, BeSS, Lasso, Elastic Net, Ridge regression and Cox Boosting, are more restrictive in selecting relevant variables. Figure 3 graphically displays a summary of the variables

selected and not selected by each of the methods. In this respect, BeSS leads to the sparsest model, it estimates 2 out of 28 to be non-zero. Lasso, Elastic Net and Ridge regression, though using different penalizations, select the same variables. *'ASLR lumbar strength LSI'* and *'Drop jump vertical propulsion LSI'* are the only variables selected by all considered methods, followed by *'Horizontal jumping impact forces LSI'* and *'Drop jump mechanical power LSI'* which are selected by five of them.

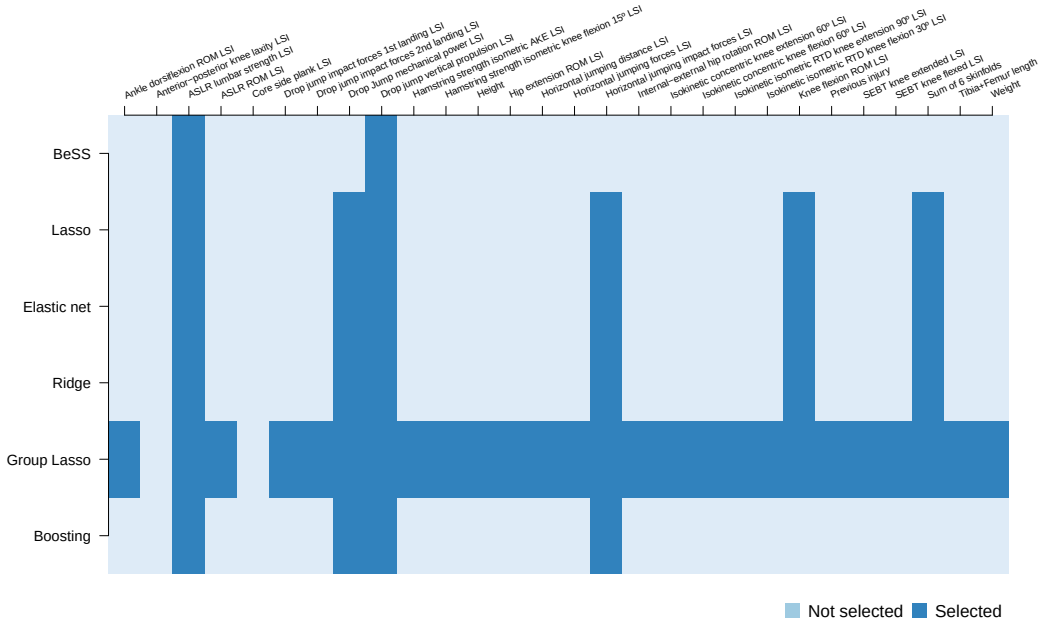

Figure 3: Summary of the selected variables (in dark blue) by each of the regularized Cox method

In Figure 4, we show the effect of the selected variables on the injury risk along with their confidence intervals, except the ones from Group Lasso. The frailty term is found to be significant for all models, emphasizing the need for such a multiplicative random effect. Further, the predictive performance of each shared frailty model is shown in Figure 5, calculated as the prediction error curve, i.e. Brier Score using the Bootstrap 0.632+ strategy. In general, except for the model based on Group Lasso selected variables, only small differences in the prediction error are visible between the different regularization approaches. The model fitted with variables that BeSS selected shows the best predictive behaviour, followed by the model based on Cox Boosting selected variables. Models based on Lasso, Elastic Net and Ridge regression selected variables show slightly better performance than the model without any covariate information and no frailty term, i.e. a Kaplan-Meier curve for all observations. Among all models, the one fitted with variables that Group Lasso selected gives the poorest prediction performance, which might be explained by the fact that the variables selected by the other regularization methods come from several groups of screening tests. In general terms, the prediction error at early times is low and similar for all models. Later time points lead to higher errors, since less information is available.

3.2 Simulation results

Table 4 summarizes the simulation results of the three different settings considered within Scenario 1. In the first setting, data were generated based on the results obtained from the frailty model fitted to the two variables BeSS selected on the real data. The best performing model is the frailty model based on BeSS selection. The second best method is Ridge regression, with an average number of wrongly selected variables of 4.17, a MSE of 3.78 and medians of the IBS between $[0, 1000]$ and $[0, 3500]$ exposure time intervals, of 0.045 and 0.086 respectively. Also, the frailty model based on

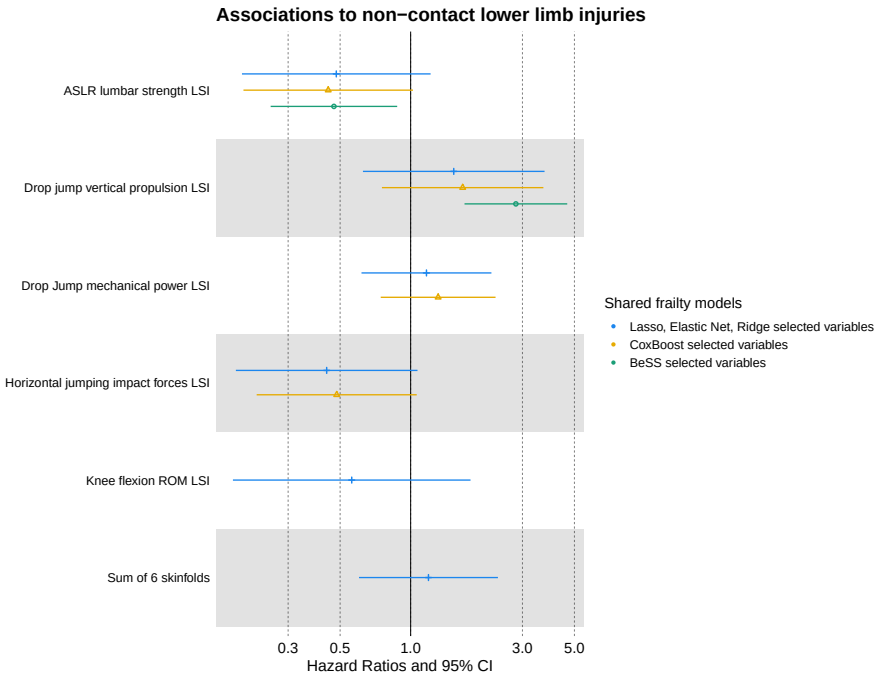

Figure 4: Hazard ratios and confidence intervals of shared frailty models fitted with the set of variables that BeSS, Lasso, Elastic Net, Ridge regression and Cox Boosting selected. Log-scale is used for the x-axis

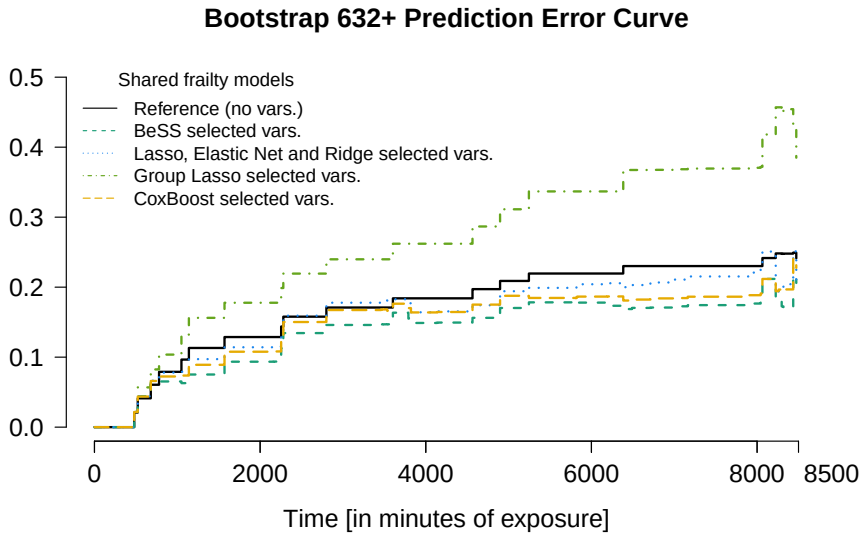

Figure 5: Bootstrap .632+ estimates of the Brier Score curves, i.e. prediction error curve, based on 30 bootstrap samples. Four Gaussian frailty models, each fitted with the set of variables that selected BeSS, Lasso –and Elastic Net and Ridge regression–, Group Lasso, Cox Boosting and a Kaplan-Meier estimator –no variables and no frailty term considered– referred to as the reference model, are compared

Cox Boosting selection performs well. Regarding the second setting, there are no clear differences in

the methods and model performances from which the setting has been generated from –that is the frailty model containing six variables of the real data, the ones that selected Lasso, Elastic Net and Ridge regression– and other selection methods. In fact, among the aforementioned three methods, Ridge is the best in terms of MSE. Subsequently, BeSS is the second best performing model in terms of the average number of wrongly selected variables (1.35), MSE (5.05) and Brier Score. Cox Boosting also behaves well with respect to the medians of the IBS. Last, with regards to the latter setting of Scenario 1, results indicate that not only Cox Boosting performs well –the method in which the setting is based on–, but also BeSS and Ridge regression methods are adequate. Indeed, when it comes to some metrics, BeSS method is notably better, particularly with respect to the average number of wrongly selected variables (0.62 versus 6.89 of Cox Boosting) and MSE (5.24 versus 6.17 of Cox Boosting). Both have similar IBS median values (0.044 BeSS and 0.045 Cox Boosting between $[0, 1000]$ and 0.080 BeSS and 0.082 Cox Boosting between $[0, 3500]$).

Table 4: Simulation results for the three different settings within Scenario 1, for 66 players with 3 observations each, that results in a sample size of 198. Average model size (AMS), the average number of falsely selected variables (ANFS), the average number of coefficients incorrectly estimated as zero (ANFNS), mean squared error (MSE) and the median of the integrated Brier scores between [0, 1000] and [0, 3500] time intervals, for all models, are reported.

Model	**AMS** (2,6,4)	**ANFS** (0)	**ANFNS** (0)	**MSE** (0)	**IBS** (0)	
					[0,1000]	**[0, 3500]**
True model: frailty (BeSS)						
BeSS	2.62	0.72	0.10	3.85	0.045	0.084
Lasso	7.91	5.93	0.02	4.29	0.046	0.087
Elastic Net	11.36	9.38	0.02	4.84	0.047	0.089
Ridge	6.11	4.17	0.06	3.78	0.045	0.086
Group Lasso	16.61	14.66	0.05	9.24	0.048	0.095
Boosting	7.38	5.39	0.01	4.29	0.046	0.086
True model: frailty (Lasso)						
BeSS	4.94	1.35	2.41	5.05	0.050	0.083
Lasso	11.47	6.41	0.94	5.67	0.050	0.085
Elastic Net	14.15	8.76	0.61	6.26	0.051	0.087
Ridge	8	3.33	1.33	4.66	0.050	0.084
Group Lasso	22.51	16.8	0.29	14.57	0.056	0.097
Boosting	11.47	6.36	0.89	5.65	0.050	0.085
True model: frailty (Boosting)						
BeSS	3.42	0.62	1.20	5.24	0.044	0.080
Lasso	10.19	6.77	0.58	6.06	0.045	0.083
Elastic Net	13.46	9.84	0.38	7.15	0.045	0.085
Ridge	7.32	4.16	0.84	5.50	0.045	0.082
Group Lasso	19.79	15.86	0.07	12.38	0.048	0.090
Boosting	10.10	6.89	0.59	6.17	0.045	0.082

Table 5 shows the simulation results within Scenario 2 and Scenario 3, for sample sizes $N_{\text{obs}} = \{60, 191, 391\}$. The results for the setting with 670 observations can be found in the Supplementary Material. On the whole, differences between frailty models diminish when the sample size is increased, either concerning the MSE or the IBS. As the boxplots of the IBS, in the follow-up time interval, of the fitted models in Figure 6 show, prediction errors get smaller with more observations available. For example, in Scenario 2, for the frailty models based on BeSS selected variables and for the frailty models based on Group Lasso selected variables, the range of prediction errors (i.e. $\max$ prediction error $-$ $\min$ prediction error) decreases by 67.9% (from 0.131 to 0.042) and 72.1% (0.176 to 0.049), respectively, when the number of teams –and thus the sample size– is increased from 1 to 10 teams. In Scenario 3 the decrease in the range of prediction errors, for previously mentioned models, is of 67.8% (from 0.112 to 0.036) and 78.1% (from 0.183 to 0.04), respectively. An important point is that the prediction error depends on the time interval considered. In a [0, 1000] time interval the IBS of different models are similar, as the predictions of the frailty models

Table 5: Simulation results for Scenarios 2 and 3, that consider different correlation structures of covariates, $\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$, for different number of players $K \in \{22, 66, 132\}$ that give rise to $N_{\text{obs}} \in \{60, 191, 391\}$ number of observations. Average model size (AMS), average number of falsely selected variables (ANFS), average number of coefficients incorrectly estimated as zero (ANFNS), mean squared error (MSE) and the median of the integrated Brier scores between $[0, 1000]$ and $[0, 3500]$ time intervals, for all models, are reported.

Sample size (N_{obs})	**Correlation structure** ($i \neq j$)	**Frailty model including vars. that selected**	**AMS** (5)	**ANFS** (0)	**ANFNS** (0)	**MSE** (0)	**IBS** (0)	
							[0,1000]	**[0, 3500]**
$N_{\text{obs}} = 60$	$\rho_{ij} = 0$	BeSS	1.65	1.45	4.80	3.69	0.030	0.108
		Lasso	1.74	1.48	4.74	56.37	0.030	0.115
		Elastic Net	2.91	2.53	4.62	271.34	0.031	0.116
		Ridge	4.83	4.28	4.45	57.02	0.033	0.121
		Group Lasso	4.05	3.50	4.45	$> 10^5$	0.034	0.125
		Cox Boosting	1.86	1.55	4.69	42.77	0.030	0.113
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	1.67	1.04	4.37	4.90	0.040	0.109
		Lasso	2.92	2	4.08	2.72	0.040	0.118
		Elastic Net	5.31	3.83	3.52	2758.8	0.044	0.130
		Ridge	7.49	5.36	2.87	776.0	0.047	0.138
		Group Lasso	8.65	6.70	3.05	$> 10^5$	0.052	0.144
		Cox Boosting	2.87	1.92	4.05	2.92	0.040	0.118
$N_{\text{obs}} = 191$	$\rho_{ij} = 0$	BeSS	1.98	1.08	4.10	0.96	0.037	0.114
		Lasso	4.75	3.28	3.53	1.10	0.037	0.114
		Elastic Net	6.23	4.47	3.24	1.21	0.037	0.114
		Ridge	6.23	4.23	3.00	1.21	0.037	0.114
		Group Lasso	15.05	11.2	1.15	3.39	0.030	0.123
		Cox Boosting	4.73	3.15	3.42	1.18	0.037	0.112
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	2.23	0.85	3.62	1.28	0.039	0.108
		Lasso	7.33	4.78	2.45	1.54	0.039	0.106
		Elastic Net	9.72	6.69	1.97	1.78	0.040	0.108
		Ridge	8.61	4.84	1.23	1.50	0.040	0.109
		Group Lasso	19.4	14.7	0.30	$> 10^5$	0.044	0.122
		Cox Boosting	6.44	3.97	2.53	1.48	0.039	0.107
$N_{\text{obs}} = 391$	$\rho_{ij} = 0$	BeSS	2.16	0.49	3.33	0.57	0.034	0.109
		Lasso	7.67	4.78	2.11	0.82	0.035	0.106
		Elastic Net	9.87	6.67	1.80	0.89	0.035	0.107
		Ridge	6.74	3.72	1.98	0.74	0.034	0.107
		Group Lasso	17.7	12.7	0	1.09	0.035	0.112
		Cox Boosting	6.22	3.62	2.40	0.79	0.034	0.106
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	2.82	1.11	3.29	1.02	0.039	0.107
		Lasso	12.14	8.53	1.39	1.18	0.039	0.104
		Elastic Net	14.66	10.64	0.98	1.27	0.039	0.105
		Ridge	8.90	4.82	0.92	0.99	0.039	0.105
		Group Lasso	25.95	20.95	0	2.40	0.040	0.114
		Cox Boosting	8.74	5.57	1.83	1.12	0.039	0.103

at early times are similar; while at farther times, differences in IBS are more apparent between the models studied. At [0, 1000] the medians of the IBS of the settings of Scenario 2 are slightly lower than those of Scenario 3, but conversely, at [0, 3500] the medians of the IBS of the settings of Scenario 3 are slightly lower with respect to the ones of Scenario 2. When there is almost no dependence between covariates of the data sets, sparse models, i.e. models with fewer variables, are obtained. Following this fact, the average model sizes of the models of Scenario 2, when $\rho_{ij} = 0$ for all $i \neq j$, are closer to the average size of the true model, whose size is five. In both scenarios, the lower the sample size, the bigger the differences in the prediction error curves, as well as in the different variable selection methods identifying the true effects. For the smallest sample size, prediction errors for Group Lasso and to some extent Ridge regression and the Elastic Net show a remarkably high range and median compared to the other methods. This tendency is even enhanced with correlated covariates as in Scenario 3.

BeSS is the method that selects the fewest number of variables, followed by Ridge regression. In contrast, Group Lasso is the method selecting the largest number of variables, and leading to the most complex models. In this sense, the average number of falsely selected variables and average number of coefficients incorrectly estimated as 0, are in line with the number of variables that each method tends to select. Since BeSS selects fewer variables, the number of falsely selected variables is generally lower. That is, the variables selected by this method are few and in addition, correspond to the true effect. On the other hand, with respect to Group Lasso, the average number of falsely selected variables is very high due to its tendency to select more variables. Inversely, Group Lasso yields a low average number of coefficients incorrectly estimated as 0, i.e. almost every true effect

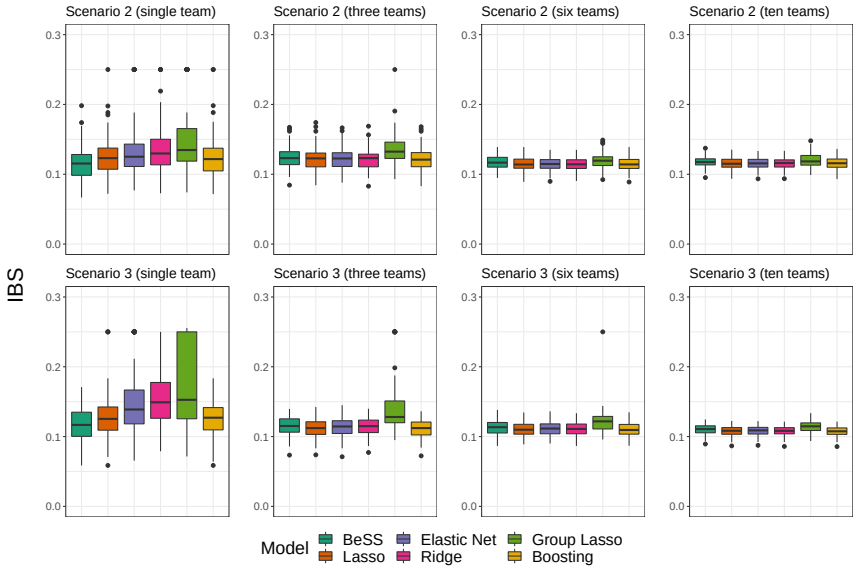

Figure 6: Comparison of the six frailty models' predictive performance by means of IBS for Scenarios 2 and 3, that consider different correlation structure of covariates, $\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$, for different number of teams consisting of $K \in \{22, 66, 132, 220\}$ players.

is correctly captured, since lower number of coefficients are estimated as 0, whereas BeSS performs worse regarding this metric. Overall, the obtained simulation results from all considered scenarios show that methods that result in sparse models, i.e. best subset selection and boosting, perform best, particularly when the sample size is small, and as sample size increases, differences between models disappear except for Group Lasso.

4 Discussion

Recurrent event models have been broadly used in many biomedical studies, but to our knowledge, only a few studies have applied them in the field of sports injury prevention. The present study aimed at providing an appropriate statistical modelling strategy for sports injury data in football, which holds some challenges, rather than providing evidence about lower-limb injuries' risk factors. Research on sports injuries is undergoing a remarkable shift, with an increasing emphasis on more powerful analytical methods. We believe that overcoming the limitations of the illustrated real case study, recurrent time-to-event methods hold great potential for sports injury research. Further investigation in larger cohorts, including several subsequent seasons and/or various sports teams, appears to be required to apply the proposed methodological approach and increase knowledge about the risk factors for sports injuries. In this respect, it should be noted that other levels of nesting would possibly be introduced, another random effect, for instance, accounting for team-specific issues due to different training styles (see Rondeau et al., 2012).

The analyses and simulation studies performed in this work suggest that the methodology presented is useful to identify screening tests associated with the risk of injury (variable selection), addressing the recurrent time-varying nature of sports injury data (frailty), for the sports medicine practice in a professional football team. However, as statisticians it is important to convey to medical service in a professional sports team that despite obtaining a large number of screening tests, the small sample size of individuals and lower-limb injury events limit the usefulness for predicting the risk of injury.

Moreover, the results obtained from our simulation study underline the described assumptions on reliability and robustness of estimated effects based on such small data. Across all scenarios, the predictive performance of the estimated models for a sample size of 22 players corresponding to the real data example, heavily depends on the chosen variable selection technique. Our results imply that Best Subset Selection and likelihood-based Boosting cope best with the specific situation of

small sample sports injury data present in many practical applications. Obtained prediction errors for Group Lasso and to some extent Ridge regression and the Elastic Net, however, indicate that these type of models are not the right choice in this case. Except for Group Lasso, differences between methods are negligible when the number of players increases. Consequently, the variable selection technique for larger cohorts is a less crucial choice.

Our study puts attention on the Cox model as one of the most classical approaches for modelling time-to-event data, and its extension to recurrent events data, the shared frailty Cox model. In a first step regularized Cox models were applied, motivated by the large number of covariates present in the data; and secondly, shared frailty Cox models were fitted to a set of reduced number of variables selected. We are aware though that the techniques used in the first step do not consider the correlation between groups of observations, and note that this could consequently lead to a flawed selection. A preferable choice would be to jointly perform both steps, i.e. selecting the important variables and fitting the model. We now discuss the choice of the methods employed and also, the ongoing research on the (simultaneous) regularization of frailty models.

Though we focused on statistical regularization techniques, modelling approaches enabling variable selection are not restricted to them. More methods exist and are useful (Witten and Tibshirani, 2010), mostly coming from the field of machine learning, such as tree-based survival techniques –recursive partitioning, random forests–, survival principal component analysis, support vector machines etc (LeBlanc and Crowley, 1992; Bair et al., 2006; Li and Luan, 2002). However, the underlying theory of most machine learning algorithms assumes that the training data is independent and identically distributed (i.i.d), they also require large training sets and are said to not perform so well with imbalanced cases (i.e. in our context very low number of injuries). Thus, without modifications, most machine learning algorithms are not directly applicable to non i.i.d data. Future research intends to integrate such machine learning survival techniques into the comparison of available methods for sports injury data. Benchmark studies taking into account machine learning survival approaches already exist for some survival tasks from other research areas, e.g. for modelling disease outcomes with genome data (Herrmann et al., 2020) or multivariate and random survival trees (Su and Fan, 2004; Ishwaran et al., 2008), but most of those methods relies on large sample sizes to train and tune the parameters.

On the other hand, other alternative survival methods for recurrent events data might be considered, such as the parametric survival models, variance-corrected Cox models and spline-based survival models. But, for our statistical analyses based on real data, shared frailty model was the preferred method over all these alternatives. Parametric survival models (e.g. accelerated failure time models (Pan, 2001)), require distributional assumptions to be made for the time-to-event outcome. Variance-corrected Cox models (e.g. Andersen-Gill 1982, Prentice-Williams-Peterson 1981, and Wei-Lin-Weissfeld 1989), account for correlation using robust standard errors, modelling the marginal distribution of each event time with corrected variance; as opposed to the shared frailty Cox model that corrects dependence among recurrent event times considering a random effect, i.e. assuming that some players are intrinsically more or less prone to experience an injury. Further, spline-based survival approaches offer a very compelling framework, such as generalized survival models (Liu et al., 2017) or piece-wise exponential additive mixed models (Bender et al., 2018), which can estimate the baseline hazard with smooth functions, include random effects and also allow high flexibility for a variety of covariate effects; nonetheless, they require more parameters to be estimated and are more data hungry than the frailty Cox approach.

Lastly, the literature provides some strategies to simultaneously perform variable selection and frailty model estimation. A first approach was proposed by Fan and Li (2002), who used a penalized likelihood estimator with smoothly clipped absolute deviation penalty (SCAD), to variable selection for gamma frailty models. Androulakis et al. (2012) extend this penalized gamma frailty model methodology to other frailty distributions. However, no open source software implementation is available yet. A recent penalization approach was developed by Groll et al. (2017) to obtain variable selection in frailty models with time-varying coefficients such that single varying effects are either included, included in the form of a constant effects or totally excluded. The method is implemented in the `PenCoxFrail` R package (Groll, 2016). These method was beyond the scope of our study, since we do not consider time-varying covariate nor time-varying effect settings. Newly, Hohberg and Groll (2020) proposed a more general Lasso Cox frailty approach allowing to perform variable selection, also for non-time-varying covariates.

In conclusion, the results of our study reveal existing potential of considered shared frailty Cox models in the context of sports injury prediction. Though conclusions from models based on

a very small sample size (e.g. a single football/sports team of about 20 players) should always be drawn with caution due to high variability, taking into account three or six teams already lead to strong improvements. Thus, we want to emphasize that –regardless of the chosen modelling strategy– a major key to increase predictive performance of the models and to obtain more accurate predictions of players' injuries lies in the size of available data. To really gain valuable new insights into prediction and monitoring of sports injuries based on data analytics, a practical advice to sports clubs is to put effort into the collection of as many data as possible, e.g. by performing regular screening tests for several of their teams.

Acknowledgements

This research was supported by the Basque Government through the BERC Programme 2018-2021 by the Spanish Ministry of Science, Innovation and Universities MICINN and FEDER: BCAM Severo Ochoa excellence accreditation SEV-2017-0718, Grant PRE2018-084007 funded by MCIN /AEI/10.13039/501100011033 and by FSE *"invest in your future"*, project PID2020-115882RB-I 00 funded by AEI/FEDER, UE and acronym "S3M1P4R" and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibility for its content. Furthermore, the authors are thankful to the two anonymous reviewers for their valuable and constructive comments which led to an improved manuscript.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120.

Androulakis, E., Koukouvinos, C., and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine*, 31(20):2223–2239.

Bahr, R. (2016). Why screening tests to predict injury do not work—and probably never will…: a critical review. *British journal of sports medicine*, 50(13):776–780.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Bender, A., Groll, A., and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321.

Binder, H. (2013). *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*. R package version 1.4.

Binder, H. and Schumacher, M. (2008). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Bolling, C., Van Mechelen, W., Pasman, H. R., and Verhagen, E. (2018). Context matters: revisiting the first step of the 'sequence of prevention'of sports injuries. *Sports medicine*, 48(10):2227–2234.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25:173–187.

Bühlmann, P., Hothorn, T., et al. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505.

Chatterjee, A. and Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Croisier, J.-L., Forthomme, B., Namurois, M.-H., Vanderthommen, M., and Crielaard, J.-M. (2002). Hamstring muscle strain recurrence and strength performance disorders. *The American journal of sports medicine*, 30(2):199–203.

Croisier, J.-L., Réveillon, V., Ferret, J., Cotte, T., Genty, M., Popovic, N., Mohty, F., Faryniuk, J., Ganteaume, S., and Crielaard, J.-M. (2003). Isokinetic assessment of knee flexors and extensors in professional soccer players. *Isokinetics and Exercise Science*, 11(1):61–62.

Crossley, K. M., Patterson, B. E., Culvenor, A. G., Bruder, A. M., Mosler, A. B., and Mentiplay, B. F. (2020). Making football safer for women: a systematic review and meta-analysis of injury prevention programmes in 11 773 female football (soccer) players. *British journal of sports medicine*.

De Visser, H., Reijman, M., Heijboer, M., and Bos, P. (2012). Risk factors of recurrent hamstring injuries: a systematic review. *British Journal of sports medicine*, 46(2):124–130.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, pages 74–99.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., Hägglund, M., McCrory, P., and Meeuwisse, W. H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Scandinavian journal of medicine & science in sports*, 16(2):83–92.

Gabbett, T. J., Ullah, S., and Finch, C. F. (2012). Identifying risk factors for contact injury in professional rugby league players–application of a frailty model for recurrent injury. *Journal of Science and Medicine in Sport*, 15(6):496–504.

Gasparini, A., Clements, M. S., Abrams, K. R., and Crowther, M. J. (2019). Impact of model misspecification in shared frailty survival models. *Statistics in medicine*, 38(23):4477–4502.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545.

Groll, A. (2016). *PenCoxFrail: Regularization in Cox Frailty Models*. R package version 1.0.1.

Groll, A., Hastie, T., and Tutz, G. (2017). Selection of effects in cox frailty models by regularization methods. *Biometrics*, 73(3):846–856.

Hägglund, M., Waldén, M., and Ekstrand, J. (2006). Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *British journal of sports medicine*, 40(9):767–772.

Harden, J. J. and Kropko, J. (2019). Simulating duration data for the cox model. *Political Science Research and Methods*, 7(4):921–928.

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. (2020). Large-scale benchmark study of survival prediction methods using multi-omics data. *arXiv preprint arXiv:2003.03621*.

Hewett, T. E., Myer, G. D., Ford, K. R., Heidt Jr, R. S., Colosimo, A. J., McLean, S. G., Van den Bogert, A. J., Paterno, M. V., and Succop, P. (2005). Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. *The American journal of sports medicine*, 33(4):492–501.

Hoerl, A. E. and Kennard, R. W. (1976). Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, 5(1):77–88.

Hohberg, M. and Groll, A. (2020). A flexible adaptive lasso cox frailty model based on the full likelihood. *arXiv preprint arXiv:2003.14118*.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis*, 1(3):255–273.

Impellizzeri, F. M., Rampinini, E., Maffiuletti, N., and Marcora, S. M. (2007). A vertical jump force test for assessing bilateral strength asymmetry in athletes. *Medicine & Science in Sports & Exercise*, 39(11):2044–2050.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.

Kelly, P. J. and Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in medicine*, 19(1):13–33.

Knapik, J. J., Bauman, C. L., Jones, B. H., Harris, J. M., and Vaughan, L. (1991). Preseason strength and flexibility imbalances associated with athletic injuries in female collegiate athletes. *The American journal of sports medicine*, 19(1):76–81.

Larruskain, J., Celorrio, D., Barrio, I., Odriozola, A., Gil, S. M., Fernandez-Lopez, J. R., Nozal, R., Ortuzar, I., Lekue, J. A., and Aznar, J. M. (2018). Genetic variants and hamstring injury in soccer: an association and validation study. *Medicine and science in sports and exercise*, 50(2):361–368.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, pages 411–425.

Li, H. and Luan, Y. (2002). Kernel cox regression models for linking gene expression profiles to censored survival data. In *Biocomputing 2003*, pages 65–76. World Scientific.

Liu, X.-R., Pawitan, Y., and Clements, M. S. (2017). Generalized survival models for correlated time-to-event data. *Statistics in medicine*, 36(29):4743–4762.

McCall, A., Carling, C., Davison, M., Nedelec, M., Le Gall, F., Berthoin, S., and Dupont, G. (2015). Injury risk factors, screening tests and preventative strategies: a systematic review of the evidence that underpins the perceptions and practices of 44 football (soccer) teams from various premier leagues. *British journal of sports medicine*, 49(9):583–589.

McGilchrist, C. and Aisbett, C. (1991). Regression with frailty in survival analysis. *Biometrics*, pages 461–466.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50(11):1.

Møller, M., Nielsen, R., Attermann, J., Wedderkopp, N., Lind, M., Sørensen, H., and Myklebust, G. (2017). Handball load and shoulder injury rate: a 31-week cohort study of 679 elite youth handball players. *British journal of sports medicine*, 51(4):231–237.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Nielsen, R. O., Bertelsen, M. L., Ramskov, D., Møller, M., Hulme, A., Theisen, D., Finch, C. F., Fortington, L. V., Mansournia, M. A., and Parner, E. T. (2019). Time-to-event analysis for sports injury research part 2: time-varying outcomes. *British journal of sports medicine*, 53(1):70–78.

Nielsen, R. Ø., Malisoux, L., Møller, M., Theisen, D., and Parner, E. T. (2016). Shedding light on the etiology of sports injuries: a look behind the scenes of time-to-event analyses. *journal of orthopaedic & sports physical therapy*, 46(4):300–311.

Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, 7(1):55–64.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022.

Rondeau, V., Mazroui, Y., and Gonzalez, J. R. (2012). Frailtypack: An r package for the analysis of correlated data with frailty models using the penalized likelihood estimation. *Journal Of Statistical Software*, 47(4).

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., and Medina, D. (2018). Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, 13(7):e0201264.

Ruddy, J. D., Cormack, S. J., Whiteley, R., Williams, M. D., Timmins, R. G., and Opar, D. A. (2019). Modeling the risk of team sport injuries: a narrative review of different statistical approaches. *Frontiers in physiology*, 10.

Sartori, S. (2011). Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.

Su, X. and Fan, J. (2004). Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics*, 60(1):93–99.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.2-7.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1):156–175.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971.

Ullah, S., Gabbett, T. J., and Finch, C. F. (2014). Statistical modelling for recurrent events: an application to sports injuries. *Br J Sports Med*, 48(17):1287–1293.

Wei, L.-J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408):1065–1073.

Wen, C., Zhang, A., Quan, S., and Wang, X. (2020). Bess: An r package for best subset selection in linear, logistic and cox proportional hazards models. *Journal of Statistical Software*, 94(4):1–24.

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.