Computational Statistics (2025) 40:1985–2003
https://doi.org/10.1007/s00180-024-01556-4

ORIGINAL PAPER

The role of the frailty in the evaluation of injury risk factors for National Basketball Association players

Ambra Macis[1]

Received: 7 December 2023 / Accepted: 5 September 2024 / Published online: 27 September 2024

Abstract

Injuries often occur in sports and, due to medical and economic reasons, it is important to understand the factors that mainly affect the risk of experiencing them. This work aims to explore this field in the context of the National Basketball Association (NBA) league. Thus, the main purpose is to identify the main individual players' characteristics that are associated to a higher risk of suffering an injury in a shorter time, taking into account ten seasons, from the beginning of 2010–2011 season until the end of 2019–2020 season. All the needed information has been retrieved from different big datasets regarding NBA players. The work stands in the survival data analysis framework and, for the purpose, a Cox regression model with frailty has been used. Results suggest that the player's position and the Body Mass Index have a significant effect on the injury's risk. From a methodological point of view, this manuscript provides an insight into the role of the frailty in the model, studying its relationship with the residuals of a mispecified Cox model.

Keywords Survival data modelling · Time-to-event · Frailty Cox model · Injury risk · Basketball

1 Introduction

Sports analytics has widespread in the last years to address many questions as, for example, the expected value of a game state, the probability of winning a match or a tournament, the evaluation of team strength and the use of sports betting market data (Baumer et al. 2023). It is straightforward that the role of statistics is very important in this field, and, even more importantly, so is the collaboration between sport scientists and statisticians (Sainani et al. 2020; Cleather et al. 2023). Moreover, besides practical applications, also the scientific research plays a fundamental role.

✉ Ambra Macis
ambra.macis@unibs.it

[1] Department of Economics and Management, University of Brescia, Contrada S. Chiara 50, 25122 Brescia, Italy

Therefore, it is important the need of an open science, that allows the reproducibility and replicability of the results (Borg et al. 2020; Bullock et al. 2023) and the role of meta-analyses, that allow to provide a quantitative method for analyzing and summarizing many research findings (Thomas and French 1986; Hagger 2006).

Nowadays many works have been published concerning the importance of statistics in different sports (Albert et al. 2005; Albert and Koning 2007; Winston 2012; Alamar 2013; Severini 2014; Miller 2015; Passos et al. 2016; Zuccolotto et al. 2017; Albert et al. 2017; Groll et al. 2018, 2019; Baumer et al. 2023; Dominicy and Ley 2023) and, among these, basketball stands out as one of the most examined. In this field, different aims have been considered, as, for example, performance analysis, the identification of factors that distinguish winning and losing teams, tracking player's analysis and analysis of the psychology of athletes (Zuccolotto and Manisera 2020).

Besides all these objectives, one of the most interesting topics is injuries prevention; indeed, injuries often occur in sports and have a negative impact on both teams and athletes. To this extent, it is worth highlighting the importance of specialized figures, such as the Sport Epidemiologist and Biostatistician (Casals and Finch 2017; Kerr 2023), able to apply appropriate statistical methods to injury data. Many contributions have considered this issue in different sports, trying to understand the factors that affect the risk of suffering an injury. Some focused on basketball (McKay et al. 2001; Drakos et al. 2010; McCarthy et al. 2013; Torres-Ronda et al. 2022; Lian et al. 2022), while others examined soccer (Venturelli et al. 2011; Zumeta-Olaskoaga et al. 2021), running (Buist et al. 2010), baseball (Jack et al. 2019) and hockey (Sochacki et al. 2019). Finally, many other works have examined injuries across various sports simultaneously (Beynnon et al. 2005; Malisoux et al. 2013; Nelson et al. 2016; Mai et al. 2017; Dekker et al. 2017; Lawrence et al. 2018; Kontos et al. 2019; Howell et al. 2019; Ekeland et al. 2020; Lu et al. 2022).

This work aims to determine, from a descriptive point of view, the main risk factors of getting injured in the National Basketball Association (NBA) League during 10 seasons (from the beginning of the 2010-2011 season until the end of the 2019-2020 season), using information retrieved from different datasets (injuries data, matches' data and players' data). From a methodological point of view, time-to-event models are the most appropriate for handling this kind of data (Ullah et al. 2014; Shrier et al. 2016; Nielsen et al. 2019a, b, 2020). These models stand in the survival analysis framework, a collection of methods firstly employed in the medical field, but now widely used in many other fields to study the occurrence of a given event of interest taking into account the time-to-event. In particular, a Cox regression model with frailty (Hanagal 2011) has been used, because it is a classical model that performs well in these contexts.

In sport analytics survival analysis has not only been used for studying injuries. For example, time-to-event models have also been employed to evaluate the career length of players (Fynn and Sonnenschein 2012), to investigate the role of team performance in the dismissal of coaches (Wangrow et al. 2018; Tozetto et al. 2019), to identify the factors that impact the time before the first substitution during a football match (Del Corral et al. 2008), to study the connection between specific attributes of young athletes and their attrition in different sports (Pion et al. 2015; Moulds et al. 2020; Smith and Weir 2022; Back et al. 2023), the duration of Olympic success (Gutiérrez et al. 2011;

Csurilla and Fertő 2022), the performance indicators able to predict the time before the first goal or the times between goals in football or ice hockey (Thomas 2007; Nevo and Ritov 2013; Pratas et al. 2016) or the skills determining a higher score in basketball (Macis et al. 2023).

Besides the practical application, this paper aims to provide a methodological contribution, drawing some insights about the role of the frailty component, studying its relationship with the residuals of a classical Cox model.

This work is organized as follows. Section 2 reports the methodological framework; then, Sect. 3 presents the case study, with two subsections devoted respectively to data and results (Sect. 3.1) and to the role of frailty in the model (Sect. 3.2). Concluding, some final remarks are reported in Sect. 4.

2 Methodological framework

Survival analysis aims to analyze the time necessary for the occurrence of a specific event during a given observation time period (follow-up). One of the main features of survival data is censoring, that occurs when the event of interest has not been observed for a particular individual during the follow-up, implying that the only information available for him/her is the last time the event did not occur (Collett 2015).

In this context, a subject i is denoted by three elements:

- a time point τ_i, either the event time t_i or the censoring time c_i;
- an event indicator δ_i assuming value 0 if the subject is censored, i.e. if $\tau_i = c_i$, and 1 otherwise;
- a vector of observed covariates $\boldsymbol{x}_i$.

Therefore, for the i^{th} subject the time-to-event can be defined as

$$\tau_i = \min(t_i, c_i) = \begin{cases} t_i & \text{if } \delta_i = 1 \\ c_i & \text{if } \delta_i = 0 \end{cases}.$$

Survival times are assumed to be the observations of a non-negative random variable T defined in $[0;\infty)$, with density function $f(t)$ and cumulative distribution function $F(t)$. The most used quantities in this setting are (*i*) the survival function and (*ii*) the hazard function. The survival function is a non-increasing and right-continuous function that measures the probability that the event of interest does not occur before time t. It can be defined as $S(t) = 1 - F(t)$, with $S(0) = 1$ and $\lim_{t\to+\infty} S(t) = 0$. Differently, the hazard function $h(t)$ measures the probability that an event occurs at a specific time t given that it has not occurred until that given timepoint. It is defined as

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}.$$

Survival data can be analyzed with statistical models or machine learning algorithms. For both approaches, a big number of possible solutions is available (Macis 2023).

Among statistical models, one of the most used methods is the Cox proportional hazards (Cox PH) regression model (Cox 1972), that is based on the proportionality hazards assumption, i.e. the ratio of the hazards of two individuals $\frac{h_1(t)}{h_2(t)}$, the so called *hazard ratio* (HR) or *relative hazard*, is constant at each time t. It aims to estimate the hazard of a subject i on the basis of a set of K covariates $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{Ki})'$:

$$h_i(t, \boldsymbol{x}_i) = h_0(t) e^{\sum_{k=1}^{K} \beta_k x_{ki}}, \tag{1}$$

where $h_i(t, \boldsymbol{x}_i)$ is the hazard of the i^{th} subject, $h_0(t)$ is the baseline hazard, x_{ki} is the observed value of the k^{th} explanatory variable for subject i (i.e. the k^{th} element of $\boldsymbol{x}_i$) and β_k is the corresponding coefficient, that is the k^{th} element of $\boldsymbol{\beta}$.

Since the expression in (1) is composed of a non-parametric component (the baseline hazard) and a parametric term, the Cox PH model is said to be semi-parametric.

Each coefficient in the vector $\boldsymbol{\beta}$ is estimated by maximizing the partial log-likelihood function:

$$\begin{aligned} \ln L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}' \boldsymbol{x}_i - \ln \left(\sum_{l \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_l} \right) \right] \\ &= \sum_{i=1}^{n} \delta_i \boldsymbol{\beta}' \boldsymbol{x}_i - \ln \left(\sum_{l \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_l} \right) \sum_{i=1}^{n} \delta_i \end{aligned} \tag{2}$$

where n is the sample size, $\boldsymbol{x}_i$ is the observed covariate vector for the i^{th} subject who experienced the event at the j^{th} ordered event time $t_{(j)}$ and R_j is the risk set, i.e. the set of subjects at risk at time $t_{(j)}$. According to this expression, censored observations ($\delta_i = 0$) contribute to the likelihood only indirectly because all the subjects are included in the risk set.

The β coefficients represent the estimated change in the logarithm of the hazard ratio in correspondence of a unit change of the corresponding covariate, independently of the other explanatory variables. Usually, for easiness of interpretation, their exponential is considered, representing the hazard ratio: for a quantitative covariate X_k, a value of e^{β_k} greater (lower) than 1 means that a one-unit increase in X_k determines an increase (decrease) in the hazard by e^{β_k}; for a categorical covariate X_k, a value of hazard ratio greater (lower) than 1 means that a subject in the corresponding group defined by X_k has a higher (lower) hazard - by an amount equal to e^{β_k} - than a subject in the reference group.

2.1 Frailty models for survival data

The Cox PH model is valid when all the events, and consequently survival times, are independent to each other; however, sometimes, survival times are not independent. A typical example is that of recurrent events, i.e. more events experienced by

the same subject; in this case, the event times of each subject are not independent (Collett 2015). Therefore, the individual heterogeneity should be taken into account to deal with this kind of data. A possible solution is to introduce a random effect, the so-called *frailty*, within the Cox PH model (Vaupel et al. 1979). In particular, the frailty Cox model is an extension of the Cox PH model that is based on the assumption that the hazard of a subject depends not only on the baseline hazard and on the set of observed covariates, but also on the frailty, a non-negative latent random variable Z, which acts multiplicatively on the baseline hazard function. Similarly to classical time-to-event models, the nature of the baseline hazard determines whether the frailty model may be classified as semi-parametric (no distributional assumption on $h_0(t)$) or parametric. The model is then defined as:

$$h_i(t, x_i) = z_i h_0(t) e^{\sum_{k=1}^{K} \beta_k x_{ki}}. \tag{3}$$

The hazard in (3) represents, therefore, the conditional hazard for an individual with a given value of the frailty z_i.

To ensure model's identifiability, Z is assumed to be scaled so that its expectation is equal to one (Balan and Putter 2020). Depending on the nature of data, the frailty Z can be defined individually or for groups of subjects (e.g. for clustered data a shared frailty model is more appropriate). In any case, it is assumed that the frailty increases or decreases the risk of experiencing the event(s) in a shorter time, suggesting an individual susceptibility to risk. In particular, two subjects with the same observed covariates can have a different risk of experiencing an event, depending on their frailty.

The model in (3) can also be written in terms of $w_i = \ln(z_i)$ as

$$h_i(t, x_i) = h_0(t) e^{w_i + \sum_{k=1}^{K} \beta_k x_{ki}}.$$

Usually, when some covariates are included in the model, the PH assumption holds conditional on the frailty (Balan and Putter 2020).

Besides the conditional hazard, the marginal hazard, can be defined as:

$$\bar{h}(t) = E[Z|T \geq t]h(t).$$

This function can be interpreted as a weighted average of individual hazards of subjects alive at time t, where the weighting depends on the distribution of Z among the subjects at risk at that timepoint. The conditional and marginal hazards coincide only if all the frailties are equal to 1 (Balan and Putter 2020).

Different distributions can be assumed for the frailty Z. Frailty models can then be expressed in terms of Laplace transform. Once the Laplace transform of frailty distribution is obtained, the parameters of frailty models can be easily estimated (Hanagal 2011).

Among the possible distributions for the frailty, there are the log-normal distribution and the family of the Power Variance Functions (PVF) proposed by Hougaard (2000). In particular, this family includes, as particular cases, the Gamma, the inverse Gaussian (IG), the Hougaard, the compound Poisson and the positive stable (PS) distributions. The choice of the frailty distribution is important because each

choice implies a different marginal model for the hazard. For example, if a Gamma distribution is chosen, more emphasis is given to late dependence of the observations. On the contrary, if, for example, the PS frailty is assumed, emphasis is set on early dependence; the IG is instead a middle ground between the two (Balan and Putter 2020). The log-normal frailty is a popular assumption, but this distribution is not infinitely divisible and its Laplace transform and expressions for the distribution of survivors are not easily obtained in closed form (Balan and Putter 2020). Further details may be found in Hanagal (2011).

On top of the distribution of the frailty, it is useful to take into account the variance of Z, because it provides information about the degree of heterogeneity in the data and, consequently, the appropriateness of the introduction of this random effect in the model.

According to the chosen frailty distribution, different estimation strategies can be set up. Among these, there are *(i)* the penalized likelihood method (for a Gamma or a log-normal frailty), *(ii)* the Laplace approximation (for a log-normal frailty), and *(iii)* the Monte Carlo expectation-maximization (EM) algorithm (for a PVF or log-normal frailty).

Further details on these estimation procedures can be found in Balan and Putter (2020).

From a computational point of view, many R packages are available for fitting frailty models. Among the existing R packages it is worth to mention:

- `survival`, which allows to estimate semi-parametric Gamma and log-normal frailty models through the penalized likelihood method;
- `frailtyEM`, which allows to estimate semi-parametric frailty models (PVF family) through the profile EM algorithm;
- `phmm`, which uses the Monte Carlo EM algorithm for log-normal frailty models;
- `frailtyHL`, which uses the h-likelihood for log-normal and Gamma frailty models;
- `frailtySurv`, which supports some of the infinitely divisible distributions from the PVF family and uses the pseudo-likelihood approach;
- `coxme`, which fits a Cox model with a Gaussian frailty, using a penalized likelihood approach.
- `frailtypack`, which allows to estimate parametric frailty models for the Gamma and log-normal distributions.
- `parfm`, which supports the PVF family distributions in a parametric setting.

3 Injury risk for NBA players

Injuries are very common in basketball, and the topic has gained more and more interest, particularly in discussions about the role of load management in injury risk in the NBA (Lewis 2018). In this case study, the injury risk for NBA basketball players in the 10 seasons from 2010–2011 to 2019–2020 has been investigated, taking into account the nature of this data, i.e. the presence of recurrent events.

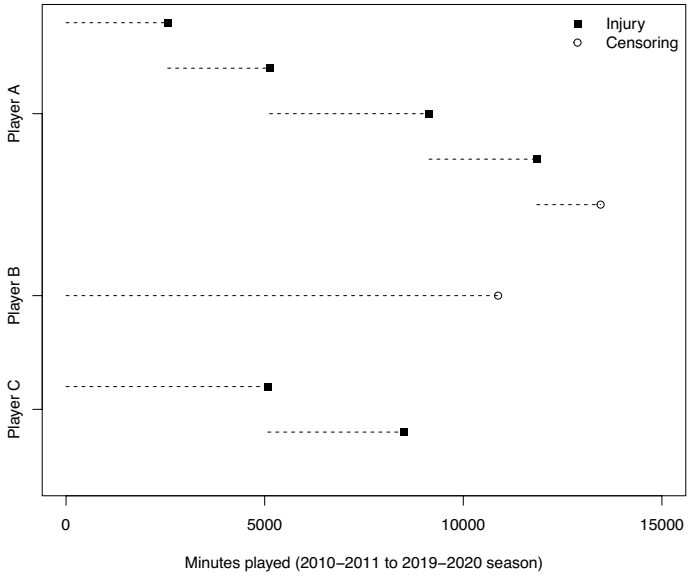

Fig. 1 Example of the data structure for three players (black square = injury, white dot = censoring)

The outcome variable is then composed of *(i)* the time-to-event variable, expressed as the minutes played by each player between two consecutive event times and *(ii)* the event indicator (taking value 1 if the player got injured in that time-interval). In particular, right-censored data have been considered: the only censored data are those of players not injured at the end of the last observed season.

A novel contribution of this work is the definition of the time-to-event variable. Indeed, instead of using matches dates, the playing time has been used as time-to-event variable. This particular choice has been made because two injuries occurred to a player at a short distance time frame (in terms of matches dates) may have a very different meaning according to the total amount of minutes played between them. Figure 1 shows an example of the data structure for three players who had respectively four, zero and two injuries in the observation period, with two censored players and one who exited the study injured.

The data analyzed in this case study were obtained through a meticulous data preparation process. In particular, three different databases have been merged together for obtaining all the needed information. The three used data sources are:

- *Inj*, the injury dataset, containing all the data about players who entered in the NBA injury list from the beginning of the 2010–2011 season until the end of the 2019-2020 season.[1] In this dataset, each row (total number of rows 27,105) included the name of the player, the date of the event (start or the end of the sick leave), the affiliated team and some notes describing the event.

[1] https://www.kaggle.com/datasets/ghopkins/nba-injuries-2010-2018.

- *PbP*, the play-by-play data of the 10 analyzed NBA seasons, consisting in almost 6,000,000 game events recorded, for more than 12,000 matches. This data have been made available by BigDataBall.[2]
- Inf_{Pl}, a SQLite database containing information about more than 64,000 matches, 4800 players, and 30 Teams.[3]

The preparation of the final data began with some preliminary cleaning and alignment operations among the three datasets. More specifically:

- all the sick leaves not due to injuries (e.g. flu, COVID-19) were excluded;
- the injuries officially recorded in the days after the match were assumed to have occurred in the match played immediately before (recovered from *PbP*);
- the extensions of the sick leave (without return of the player to the lineup) have been considered as a single injury (merged to the previous injury);
- the players' names have been harmonized across the three datasets because, sometimes, the same player was recorded with different names (*e.g* 'Mo Bamba' and 'Mohamed Bamba').

Therefore, injuries have been recorded based on the injury list. No difference has been made between time-loss and medical attention injuries (Bahr et al. 2020; Rodas et al. 2019). However, usually, players who enter in the injury list sustained time-loss injuries.

Finally, we merged the three datasets and we obtained the final data in long format (life-table form); therefore, there were different rows for each player, depending on the number of suffered injuries. The dataset included information about the total amount of minutes played by each player in the examined seasons (follow-up duration), the starting and ending point of each considered time-interval expressed as minutes played up to the occurrence of an injury or the player's exit from the study, the indicator of injury's occurrence and some players' features. The example shown in Fig. 1 has been also reported in tabular form (Table 1) for showing the structure of the dataset.

3.1 Data description and model estimation

The overall sample consisted of 1299 NBA players observed over 10 seasons. The players were firstly classified in seven different roles depending on the main assigned play position (center, center-forward, forward-center, guard, guard-forward, forward-guard and forward). Then, we considered three macro-categories: center (center, center-forward and forward-center), guard (guard, guard-forward and forward-guard) and forward, in order to gather players who have basically a similar play-style. Using these three categories, it was observed that 635 players

[2] www.bigdataball.com.

[3] https://www.kaggle.com/datasets/wyattowalsh/basketball.

Table 1 Example of the dataset structure (life-table form)

Player	Follow-up time	Tstart	Tstop	Injury	BMI	Position
A	13456	0	2560	1	28	Guard
A	13456	2560	5123	1	28	Guard
A	13456	5123	9138	1	28	Guard
A	13456	9138	11850	1	28	Guard
A	13456	11850	13456	0	28	Guard
B	11850	0	11850	0	24	Forward
C	8500	0	5080	1	22	Center
C	8500	5080	8500	1	22	Center

The data reported in the table are those represented in Fig. 1. Tstart and Tstop are respectively the starting and ending timepoints of each observed time interval. Injury is the event indicator

BMI: Body Mass Index

(48.9%) were guards, 373 (28.7%) were forwards and 291 players (22.4%) were centers. The mean Body Mass Index (BMI) of the players included in the sample was equal to 24.91 kg/m^2 (with standard deviation SD equal to 1.7).

The observed players, on average, played 4733.9 minutes (SD=6189.4) during the follow-up.

Over the 10 seasons, 9463 injuries were observed. In the overall sample the mean number of observed injuries was equal to 7.3 (SD=8.4), while among injured players it was 8.6 (SD=8.5).

It emerged that 84.4% of the players ($n = 1096$) suffered from at least one injury and, among these, 288 players ended the study injured. Therefore, the percentage of censored players (not injured at the end of the follow-up) in the sample was equal to 37.8%. More in details, we can see that just 9.2% of the players sustained only one injury, while most of the players suffered from at least two injuries. Moreover, it emerged that 40.9% of centers had more than 7 injuries; this percentage is lower for forwards and guards (32.9% and 32.0% respectively).

These preliminary evidences are interesting; however, the number of injuries may be also influenced by the duration of the observation period for each player; therefore, we also evaluated the injury incidence rate for each player (Lee and Zumeta-Olaskoaga 2022). To make the interpretation easier, this rate has been evaluated per 100 player-match exposure:

$$I_{r,i} = \frac{n\_inj_i}{\Delta T_i} \times 48 \times 100,$$

where $I_{r,i}$ is the injury incidence rate of the i^{th} player, n_inj_i is the number of injuries he suffered, and ΔT_i is the total number of minutes he played over the ten seasons.

The mean injury incidence rate in the sample was equal to 47.4, indicating that on average almost 48 injuries per 100 players occurred in 48 minutes played.

Table 2 Descriptive statistics of the sample according to the injury incidence rate

	< 3.19	[3.19; 8)	[8; 20.9)	≥ 20.9	*Overall*
N (%)	325 (25%)	324 (25%)	325 (25%)	325 (25%)	1299 (100%)
Position					
Center, n (%)	52 (16.0%)	67 (20.7%)	87 (26.8%)	84 (25.9%)	291 (22.4%)
Forward, n (%)	99 (30.5%)	93 (28.7%)	85 (26.2%)	96 (29.6%)	373 (28.7%)
Guard, n (%)	174 (53.5%)	164 (50.6%)	153 (47.1%)	144 (44.4%)	635 (48.9%)
BMI, Mean (SD)	24.9 (1.8)	25.1 (1.6)	24.9 (1.6)	24.7 (1.9)	24.9 (1.7)

For position frequencies and percentages (conditional on the injury incidence rate) have been reported. For BMI mean values and standard deviation have been provided

BMI: Body Mass Index; SD: Standard Deviation

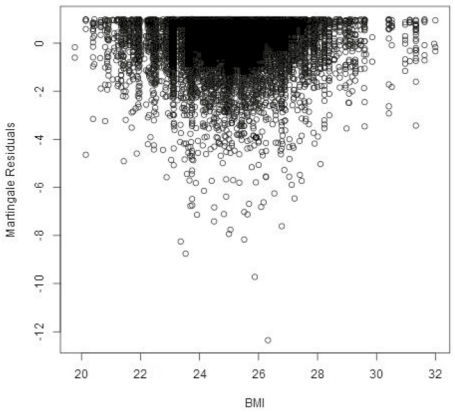

Fig. 2 Plot of Martingale residuals against the observed values of the Body Mass Index

Then, we used the injury incidence rate for classifying the players into four groups of players using quartiles. Some summary statistics for these players groups are shown in Table 2.

We can see that, with respect to the overall statistics, a higher percentage of guards had a lower injury incidence rate. Conversely, a higher percentage of centers had higher injury incidence rates. So, these preliminary statistics highlight that centers suffered a higher number of injuries than guards and forwards.

The final aim of this work is to analyze the risk of getting injured, trying to identify the individual players' characteristics that affect this risk. To this extent, a Cox model with a Gamma frailty has been fitted, and estimates have been obtained by using the profile EM algorithm implemented in the `frailtyEM` R package. We chose to use the Gamma distribution because of its simplicity and flexibility (Wienke et al. 2003). Furthermore, many trials have been carried out considering different frailty distributions (e.g. Inverse Gaussian and positive stable distribution), but there were some computational issues.

The survival outcome consisted of the time-to event (composed of the starting and ending point of each observed interval) and the event indicator, assuming

value 1 if the player suffered an injury in that time interval. The explanatory variables included in the model were the BMI, the player's position and their interaction. In particular, since it was found that the BMI had a non-linear effect on the risk of getting injured (Fig. 2), the following transformation of the BMI has been considered:

$$\log\left[(BMI - Mean(BMI))^2 + 1\right].$$

The covariates included in the final model are the following:

- the transformation of the BMI (BMI_{transf}),
- the players' position (with three categories: guard, forward and center),
- the interaction effect between them.

In particular, we obtained a significant parameter for the interaction effect considering in the interaction term the position with only two categories ($Position_{aggr}$ - forward and center, guard):

$$h_i(t) = h_0(t)e^{\beta_1 Position + \beta_2 BMI_{transf} + \beta_3 BMI_{transf} \times Position_{aggr}}$$

For the following analyses the significance level α has been fixed equal to 0.05.

The results of the final model are shown in Table 3. Interestingly, we can see that centers and forwards had a 61% and 32% higher probability of experiencing an injury earlier compared to guards (reference category). Moreover, the interaction ($p = 0.045$) suggests that the effect of BMI on the injury risk is different depending on the players' position. In particular, the following effect holds for the guards: as the BMI deviates from the mean, the risk of getting injured earlier increases. Therefore, for the guards, when the BMI is very low or very high compared to the mean BMI, the risk of injury is higher.

The estimate of the frailty variance is 0.75 and its related 95% confidence interval is [0.68; 0.84]. The random effect is highly significant ($p < 0.001$), as indicated by the Commenges-Andersen test. This result confirms that including the frailty component in the model is appropriate because the data are heterogeneous. Moreover,

Table 3 Results of the Cox PH frailty model

	β estimate [95% CI]	Hazard ratio [95% CI]	p value
Position			
Guard	ref	ref	ref
Forward	0.274 [0.078;0.470]	1.315 [1.081;1.600]	0.006
Center	0.476 [0.364;0.588]	1.609 [1.439;1.800]	< 0.001
$\mathbf{BMI_{transf}}$	0.122 [0.010;0.234]	1.130 [1.010;1.264]	0.033
$\mathbf{BMI_{transf} \times Position_{aggr}}$	−0.150 [−0.296; − 0.004]	0.860 [0.743;0.996]	0.045

CI: Confidence interval; BMI: Body Mass Index

$\mathbf{BMI_{transf}} = \log\left[(BMI - Mean(BMI))^2 + 1\right]$

we carried out the proportionality hazards test and we observed that the Gamma frailty model satisfies the conditional proportional hazards assumption and, therefore, allows to explain the marginal non-proportionality (Balan and Putter 2019).

Figure 3 aims to show the effect of BMI and position on the injury risk, for two different values of frailty. In particular, the hazard has been estimated at a given timepoint (100 minutes) according to the position (guard, forward and center) and to the value of the frailty (1.3 and 2.3). Then, the hazard has been evaluated for different values of the BMI (ranging from 19 to 33). First of all, the figure highlights the non-linear effect of the BMI on the risk for the guards (red lines): the injury risk assumes higher values when the BMI is lower or higher than the mean value (24.9). Conversely, this effect is not observed for centers and forwards, who have a similar risk behaviour. Moreover, we can see that on average, centers (green lines) are those at highest risk of getting injured. Finally, the figure shows that, as expected, lower values of frailty (dashed lines) imply a lower injury risk.

Machine learning methods may be an interesting alternative for analysing this data; however, to the best of my knowledge, there are relatively few algorithms capable of effectively accounting for recurrent events and incorporating the frailty term into their models. Some preliminary analyses have been carried out fitting multivariate survival trees (Su and Fan 2004; Calhoun et al. 2018) and random survival forests (Ishwaran et al. 2008, 2022), and in both cases there were some issues, due to convergence and lack of packages documentation.

3.2 The role of frailty

From a methodological point of view we need to consider further the role of the frailty within the model. One of the main aims of the paper is therefore to explore it. Indeed, the name of this random effect and its definition may suggest that a higher frailty indicates that a player is frailer (in absolute terms). However, from an exploratory analysis, it resulted that the frailty is not associated, for example, with the number of suffered injuries. This is due to the fact that, instead,

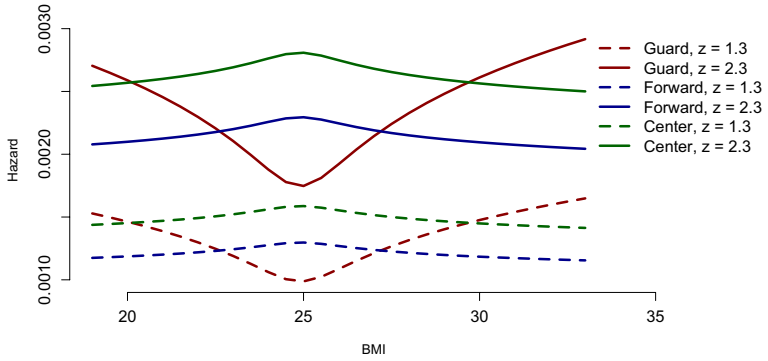

Fig. 3 Estimated hazard at a given timepoint ($t = 100$ minutes) as a function of the BMI and player's position for players with two different values of frailty ($z = 1.3, z = 2.3$)

the frailty represents a sort of intrinsic frailty of the subject with respect to the estimated risk. To confirm this assertion, the relationship between the residuals of a mispecified Cox model (*i.e.* not taking into account the presence of recurrent events and, therefore, without frailty) and the frailty values has been explored. The obtained results are represented in Fig. 4. It emerged that the negative residuals of a mispecified model, indicating an overestimate of the risk (i.e. the player got injured later than expected), are associated with frailties lower than 1. Therefore, a value of the frailty lesser than 1 lowers the estimated risk of the player. On the contrary, positive residuals, that show an underestimate of the risk (i.e. the player suffered an injury sooner than expected), are mainly associated with a frailty greater than 1. Thus, in this case, the frailty increases the risk for the player. So, the frailty indicates whether the player is expected to suffer a higher/lower number of injuries than another player with the same covariates and observed for the same time period. This is clearly shown in Fig. 3: given a set of covariates, i.e. same BMI and same position, players with a lower value of frailty had a lower injury risk than players with a higher frailty.

In conclusion, the frailty is a random effect needed for catching the heterogeneity in the data improving the estimate of risk that, otherwise, would be overestimated or underestimated. Therefore, it is important to include this term when data are heterogeneous (e.g. recurrent events, clustered data) to obtain more accurate estimates.

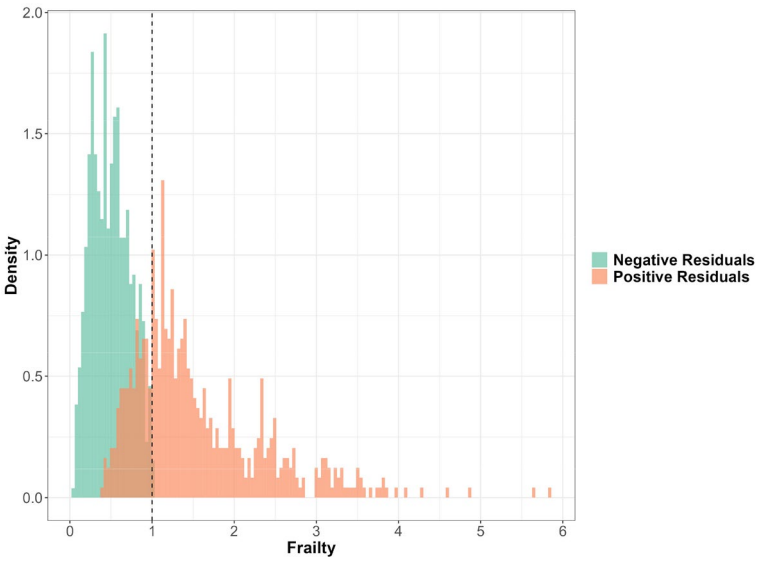

Fig. 4 Histogram of the values of the frailty of a Cox PH regression model for recurrent events (bars coloured according to the sign of the residuals of a Cox PH regression model without frailty)

4 Concluding remarks

In this paper, the risk of the NBA players of getting injured has been evaluated using a frailty model for survival analysis (namely, the Cox PH with frailty). This model, indeed, allows to address the issue of dependent survival times occurring when dealing with recurrent events. The data used for the analyses have been obtained by the merge of three different datasets, and data preparation required a huge effort for harmonizing them. In particular, a wide dataset containing all the injuries occurred to NBA players starting from season 2010–2011 to season 2019–2020 has been used and merged with the play-by-play data of all the matches played in the same period and a dataset containing players' individual characteristics.

From a practical point of view, the novelty of this work stands in the definition of the time-to-event as playing time and, therefore, measured as minutes played until the injury.

The descriptive statistics highlighted that centers had a higher injury incidence rate compared to forwards and guards. This evidence is also confirmed by the results of the Cox PH model with frailty. Indeed, some interesting relationships between the individual injury risk and individual characteristics of the players emerged. In particular, we saw that guards were expected to get injured later than forwards and centers and that centers were the players at highest risk of injury. Moreover, an interesting result was obtained for the BMI: this variable, indeed, had a different effect on injury risk depending on the player's position. In detail, it was observed that for the guards values far away from the mean BMI were associated to a highest risk of getting injured in a shorter time. We chose to consider the BMI and position as covariates for many reasons. In particular, the BMI allows to evaluate the player's physical build, while player's position enables us to also consider his playing style. We did not include height and/or weight for avoiding issues of multicollinearity.

From a methodological point of view, an interesting contribution of this work is related to the interpretation of the frailty. Indeed, given its name and its definition, one might infer that it represents an individual measure of vulnerability in absolute terms. However, it is merely a random effect that is needed to take into account the within-subjects correlation and to adjust the individual estimate of risk of an observation with a given set of covariates. Therefore, including this random effect into the model allows to improve the estimate of risk according to the individual vulnerability of subjects. In this way, a frailty less than 1 is required to lower the risk of a player whose risk would be, otherwise, overestimated. The vice versa holds if a frailty greater than 1 is observed. So, the frailty improves the parameters estimation, allowing to reduce the error that would occur if the data structure were ignored. Indeed, as plotted in Fig. 4, it effectively captures the residuals of the mispecified model.

In this context, therefore, the frailty indicates whether the injury risk of a player is higher/lower than that of another player observed for the same time period and with the same covariates. This effect is also clearly shown in Fig. 3: given a set of covariates and an observation time, the injury risk of a player increases/decreases according to his frailty.

Besides the importance of introducing this random term from a methodological point of view, frailty models are also useful from a practical standpoint. To this extent, the frailty models enable us to focus on players who are at highest risk of getting injured. Thus, knowing a player's frailty may enable coaches to offer better protection and tailor specific training sessions, while also paying close attention to load management, which can impact the risk of injury (Lewis 2018).

From a practical point of view, future research could focus on the definition of a new variable able to account for the injury's severity. Moreover, a deeper investigation of the relationship between the injury risk and other covariates could be done for further improving the identification of injury risk factors. Furthermore, developments will focus on considering only the players who got injured at least once, in order to explore whether the previous injuries affect the risk of injury recurrence.

In addition, from a methodological point of view, subject heterogeneity could be studied with random effects models also taking into account for confounding variables not included in this analysis as, for example, load variables, age, injury severity. Furthermore, the model could also be adjusted for left truncation, since players may have sustained other injuries at the beginning of the study.

Then, further research will pay attention to different approaches for modelling recurrent events, e.g. models developed in the framework of counting processes.

Finally, our interest will also focus on machine learning methods to recurrent times-to-event data from both a methodological and practical point of view.

Acknowledgements I would like to thank Prof. Paola Zuccolotto, Prof. Marica Manisera and Dr. Marco Sandri for their precious help in developing this work. I would also like to thank the two anonymous reviewers for their valuable comments, which greatly improved the paper, the Editor and the Guest Editors. The research was carried out in collaboration with the Big &Open Data Innovation Laboratory at University of Brescia (BODaI-Lab, bodai.unibs.it), project "Big Data Analytics in Sport" (BDsports, bdsports.unibs.it).

Funding Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE Agreement. Research Project PRIN 2022, granted by European Union - Next Generation EU, "Statistical Models and AlgoRiThms in sports (SMARTsports). Applications in professional and amateur contexts, with able-bodied and disabled athletes", project nr. 2022R74PLE, CUP: D53D23005950006.

Declarations

Conflict of interest The author has no conficts of interest to disclose.

References

Alamar BC (2013) Sports analytics: a guide for coaches, managers, and other decision makers. Columbia University Press, Columbia

Albert J, Bennett J, Cochran JJ (2005) Anthology of statistics in sports, vol 16. SIAM. https://doi.org/10.1137/1.9780898718386

Albert J, Koning RH (2007) Statistical thinking in sports. CRC Press, Boca Raton

Albert J, Glickman ME, Swartz TB et al (2017) Handbook of statistical methods and analyses in sports. CRC Press, Boca Raton

Back FA, Hino AAF, Bojarski WG et al (2023) Evening chronotype predicts dropout of physical exercise: a prospective analysis. Sport Sci Health 19(1):309–319. https://doi.org/10.1007/s11332-022-00963-8

Bahr R, Clarsen B, Derman W et al. (2020) International Olympic Committee consensus statement: methods for recording and reporting of epidemiological data on injury and illness in sports 2020 (including the STROBE extension for sports injury and illness surveillance (STROBE-SIIS)). Orthop J Sports Med 8(2). https://doi.org/10.1177/2325967120902908

Balan TA, Putter H (2019) frailtyEM: an R package for estimating semiparametric shared frailty models. J Stat Softw 90(7):1–2. https://doi.org/10.18637/jss.v090.i07

Balan TA, Putter H (2020) A tutorial on frailty models. Stat Methods Med Res 29(11):3424–345. https://doi.org/10.1177/09622802209218

Baumer BS, Matthews GJ, Nguyen Q (2023) Big ideas in sports analytics and statistical tools for their investigation. Wiley Interdiscip Rev Comput Stat 15(6):e161. https://doi.org/10.1002/wics.1612

Beynnon BD, Vacek PM, Murphy D et al (2005) First-time inversion ankle ligament trauma: the effects of sex, level of competition, and sport on the incidence of injury. Am J Sports Med 33(10):1485–149. https://doi.org/10.1177/0363546505275490

Borg DN, Bon J, Sainani K, et al. (2020) Sharing data and code: a comment on the call for the adoption of more transparent research practices in sport and exercise scienhttps://doi.org/10.31236/osf.io/ftdgj.

Buist I, Bredeweg SW, Bessem B et al (2010) Incidence and risk factors of running-related injuries during preparation for a 4-mile recreational running event. Br J Sports Med 44(8):598–60. https://doi.org/10.1136/bjsm.2007.044677

Bullock GS, Ward P, Impellizzeri FM et al (2023) Up front and open? Shrouded in secrecy? Or Somewhere in between? A meta-research systematic review of open science practices in sport medicine research. J Orthop Sports Phys Ther 53(12):735–74. https://doi.org/10.2519/jospt.2023.12016

Calhoun P, Su X, Nunn M et al. (2018) Constructing multivariate survival trees: the MST package for R. J Stat Softw 83(12):1–21. https://doi.org/10.18637/jss.v083.i12

Casals M, Finch CF (2017) Sports Biostatistician: a critical member of all sports science and medicine teams for injury prevention. Inj Prev 23(6):423–42. https://doi.org/10.1136/injuryprev-2016-042211

Cleather DJ, Hopkins W, Drinkwater EJ, et al (2023) Improving collaboration between statisticians and sports scientists. Br J Sports Med

Collett D (2015) Modelling survival data in medical research. CRC Press, Boca Raton

Cox DR (1972) Regression models and life-tables. J R Stat Soc Ser B Methodol 34(2):187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Csurilla G, Fertő I (2022) How long does a medal win last? Survival analysis of the duration of olympic success. Appl Econ 54(43):5006–502. https://doi.org/10.1080/00036846.2022.2039370

Dekker TJ, Godin JA, Dale KM et al (2017) Return to sport after pediatric anterior cruciate ligament reconstruction and its effect on subsequent anterior cruciate ligament injury. J Bone Jt Surg 99(11):897–90. https://doi.org/10.2106/JBJS.16.00758

Del Corral J, Barros CP, Prieto-Rodriguez J (2008) The determinants of soccer player substitutions: a survival analysis of the Spanish soccer league. J Sports Econ 9(2):160–167. https://doi.org/10.1177/1527002507308309

Dominicy Y, Ley C (2023) Statistics meets sports: what we can learn from sports data. Cambridge Scholars Publishing, Cambridge

Drakos MC, Domb B, Starkey C et al (2010) Injury in the national basketball association: a 17-year overview. Sports Health 2(4):284–290. https://doi.org/10.1177/1941738109357303

Ekeland A, Engebretsen L, Fenstad AM et al (2020) Similar risk of ACL graft revision for alpine skiers, football and handball players: the graft revision rate is influenced by age and graft choice. Br J Sports Med 54(1):33–3. https://doi.org/10.1136/bjsports-2018-100020
Fynn KD, Sonnenschein M (2012) An analysis of the career length of professional basketball players. Macalester Rev 2(2):3
Groll A, Manisera M, Schauberger G et al (2018) Guest editorial ' statistical modelling for sports analytics'. Stat Model 18(5–6):385–387. https://doi.org/10.1177/1471082x18810264
Groll A, Manisera M, Schauberger G et al (2019) Guest editorial ' statistical modelling for sports analytics'. Stat Model 19(1):3. https://doi.org/10.1177/1471082x18810965
Gutiérrez E, Lozano S, González JR (2011) A recurrent-events survival analysis of the duration of Olympic records. IMA J Manag Math 22(2):115–12. https://doi.org/10.1093/imaman/dpq005
Hagger M (2006) Meta-analysis in sport and exercise research: review, recent developments, and recommendations. Eur J Sports Sci 6(2):103–11. https://doi.org/10.1080/17461390500528527
Hanagal DD (2011) Modeling survival data using frailty models. Springer, Berlin
Hougaard H (2000) Analysis of multivariate survival data. Springer, Berlin
Howell DR, Potter MN, Kirkwood MW et al (2019) Clinical predictors of symptom resolution for children and adolescents with sport-related concussion. J Neurosurg Pediatr 24(1):54–6. https://doi.org/10.3171/2018.11.PEDS18626
Ishwaran H, Kogalur UB, Blackstone EH et al (2008) Random survival forests. Ann Appl St. https://doi.org/10.1214/08-AOAS169
Ishwaran H, Kogalur UB, Kogalur MUB (2022) Package 'randomforestsrc'. breast 6:1
Jack RA, Sochacki KR, Hirase T, et al (2019) Performance and return to sport after hip arthroscopic surgery in major league baseball players. Orthop J Sports Med 7(2). https://doi.org/10.1177/2325967119825835
Kerr ZY (2023) No, my first name ain't 'Biostatistician'. It's 'Epidemiologist'(Dr. Kerr, if you're nasty). Br J Sports Med
Kontos AP, Elbin R, Sufrinko A et al (2019) Recovery following sport-related concussion: integrating pre-and postinjury factors into multidisciplinary care. J Head Trauma Rehabil 34(6):394–40. https://doi.org/10.1097/HTR.0000000000000536
Lawrence DW, Richards D, Comper P et al (2018) Earlier time to aerobic exercise is associated with faster recovery following acute sport concussion. PLoS One 13(4):e0196062. https://doi.org/10.1371/journal.pone.0196062
Lee DJ, Zumeta-Olaskoaga L (2022) Can we really predict injuries in team sports? Bol Estad Invest Oper 38(3):149
Lewis M (2018) It's a hard-knock life: game load, fatigue, and injury risk in the National basketball association. J Athl Train 53(5):503–50. https://doi.org/10.4085/1062-6050-243-17
Lian J, Sewani F, Dayan I et al (2022) Early ACLR and risk and timing of secondary meniscal injury compared with delayed ACLR or nonoperative treatment: a time-to-event analysis using machine learning. Am J Sports Med 50(5):1416–1429. https://doi.org/10.1177/0363546521101450
Lu Y, Jurgensmeier K, Till SE et al (2022) Early ACLR and risk and timing of secondary meniscal injury compared with delayed ACLR or nonoperative treatment: a time-to-event analysis using machine learning. Am J Sports Med 50(13):3544–3556. https://doi.org/10.1177/03635465221124258
Macis A (2023) Statistical models and machine learning for survival data analysis. PhD thesis, University of Brescia
Macis A, Manisera M, Zuccolotto P, et al. (2023) A survival analysis to discover which skills determine a higher scoring in basketball. Stat Applicata-Italian J Appl Stat 35(2). https://doi.org/10.26398/IJAS.0035-009
Mai HT, Chun DS, Schneider AD et al (2017) Performance-based outcomes after anterior cruciate ligament reconstruction in professional athletes differ between sports. Am J Sports Med 45(10):2226–223. https://doi.org/10.1177/0363546517704834
Malisoux L, Frisch A, Urhausen A et al (2013) Monitoring of sport participation and injury risk in young athletes. J Sci Med Sport 16(6):504–50. https://doi.org/10.1016/j.jsams.2013.01.008
McCarthy MM, Voos JE, Nguyen JT et al (2013) Injury profile in elite female basketball athletes at the women's national basketball association combine. Am J Sports Med 41(3):645–65. https://doi.org/10.1177/03635465124742
McKay GD, Goldie P, Payne WR et al (2001) Ankle injuries in basketball: injury rate and risk factors. Br J Sports Med 35(2):103–10. https://doi.org/10.1136/bjsm.35.2.103

Miller TW (2015) Sports analytics and data science: winning the game with methods and models. FT Press, Upper Saddle River

Moulds K, Abbott S, Pion J et al (2020) Sink or swim? A survival analysis of sport dropout in Australian youth swimmers. Scand J Med Sci Sports 30(11):2222–223. https://doi.org/10.1111/sms.13771

Nelson LD, Tarima S, LaRoche AA et al (2016) Preinjury somatization symptoms contribute to clinical recovery after sport-related concussion. Neurology 86(20):1856–186. https://doi.org/10.1212/wnl.0000000000002679

Nevo D, Ritov Y (2013) Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. J Quant Anal Sports 9(2):165–17. https://doi.org/10.1515/jqas-2012-0004

Nielsen RO, Bertelsen ML, Ramskov D et al (2019) Time-to-event analysis for sports injury research part 1: time-varying exposures. Br J Sports Med 53(1):61–6. https://doi.org/10.1136/bjsports-2018-099408

Nielsen RO, Bertelsen ML, Ramskov D et al (2019) Time-to-event analysis for sports injury research part 2: time-varying outcomes. Br J Sports Med 53(1):70–7. https://doi.org/10.1136/bjsports-2018-100000

Nielsen RO, Shrier I, Casals M et al (2020) Statement on methods in sport injury research from the 1st methods matter meeting, Copenhagen 2019. Br J Sports Med 54(15):941–94. https://doi.org/10.1136/bjsports-2019-101323

Passos P, Araújo D, Volossovitch A (2016) Performance analysis in team sports. Taylor & Francis, Oxfordshire

Pion J, Lenoir M, Vandorpe B et al (2015) Talent in female gymnastics: a survival analysis based upon performance characteristics. Int J Sports Med 94(11):935–94. https://doi.org/10.1055/s-0035-1548887

Pratas JM, Volossovitch A, Carita AI (2016) The effect of performance indicators on the time the first goal is scored in football matches. Int J Perform Anal Sport 16(1):347–35. https://doi.org/10.1080/24748668.2016.11868891

Rodas G, Bove T, Caparrós T et al (2019) Ankle sprain versus muscle strain injury in professional men's basketball: a 9-year prospective follow-up study. Orthop J Sports Med 7(6):232596711984903. https://doi.org/10.1177/2325967119849035

Sainani KL, Borg DN, Caldwell AR et al (2020) Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. Br J Sports M. https://doi.org/10.1136/bjsports-2020-102607

Severini TA (2014) Analytic methods in sports: using mathematics and statistics to understand data from baseball, football, basketball, and other sports. Chapman and Hall/CRC, Boca Raton

Shrier I, Steele R, Zhao M et al (2016) A multistate framework for the analysis of subsequent injury in sport (M-FASIS). Scand J Med Sci Sports 26(2):128–13. https://doi.org/10.1111/sms.12493

Smith KL, Weir PL (2022) An examination of relative age and athlete dropout in female developmental soccer. Sports 10(5):7. https://doi.org/10.3390/sports10050079

Sochacki KR, Jack RA, Hirase T et al (2019) Performance and return to sport after hip arthroscopy for femoracetabular impingement syndrome in National Hockey League players. J Hip Preserv Surg 6(3):234–24. https://doi.org/10.1093/jhps/hnz030

Su X, Fan J (2004) Multivariate survival trees: a maximum likelihood approach based on frailty models. Biometrics 60(1):93–9. https://doi.org/10.1111/j.0006-341X.2004.00139.x

Thomas AC (2007) Inter-arrival times of goals in ice hockey. J Quant Anal Sports 3(3). https://doi.org/10.2202/1559-0410.1064

Thomas JR, French KE (1986) The use of meta-analysis in exercise and sport: a tutorial. Res Q Exerc Sport 57(3):196–204. https://doi.org/10.1080/02701367.1986.10605397

Torres-Ronda L, Gámez I, Robertson S et al (2022) Epidemiology and injury trends in the National basketball association: pre-and per-COVID-19 (2017–2021). PLoS ONE 17(2):e026335. https://doi.org/10.1371/journal.pone.0263354

Tozetto AB, Carvalho HM, Rosa RS et al (2019) Coach turnover in top professional Brazilian football championship: a multilevel survival analysis. Front Psychol 10:124. https://doi.org/10.3389/fpsyg.2019.01246

Ullah S, Gabbett TJ, Finch CF (2014) Statistical modelling for recurrent events: an application to sports injuries. Br J Sports Med 48(17):1287–129. https://doi.org/10.1136/bjsports-2011-090803

Vaupel JW, Manton KG, Stallard E (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16(3):439–45. https://doi.org/10.2307/2061224

Venturelli M, Schena F, Zanolla L et al (2011) Injury risk factors in young soccer players detected by a multivariate survival model. J Sci Med Sport 14(4):293–29. https://doi.org/10.1016/j.jsams.2011.02.013

Wangrow DB, Schepker DJ, Barker VL III (2018) Power, performance, and expectations in the dismissal of NBA coaches: a survival analysis study. Sport Manage Rev 21(4):333–34. https://doi.org/10.1016/j.smr.2017.08.002

Wienke A, Arbeev K, Locatelli I, et al. (2003) A simulation study of different correlated frailty models and estimation strategies. Technical Report, MIPDR Working Paper WP

Winston WL (2012) Mathletics: how gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football. Princeton University Press, Princeton

Zuccolotto P, Manisera M (2020) Basketball data science: with applications in R. CRC Press, Boca Raton

Zuccolotto P, Manisera M, Kenett R (2017) Guest Editorial ' Statistics in sports'. Electron J Appl Stat 10(3):1–2

Zumeta-Olaskoaga L, Weigert M, Larruskain J et al (2021) Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models. AStA Adv Stat Anal 107:1–2. https://doi.org/10.1007/s10182-021-00428-2