Computational Statistics (2026) 41:127
https://doi.org/10.1007/s00180-026-01784-w

ORIGINAL PAPER

The impact of weakness and recovery time on recurrent injuries in NBA: a dynamic event-count model with transient carryover effect

Ambra Macis[1,2] · **Marica Manisera**[1,2] · **Marco Sandri**[2] · **Paola Zuccolotto**[1,2]

Received: 16 October 2025 / Accepted: 6 July 2026

Abstract
The recurrence of injuries represents a major challenge in professional sports, with relevant financial implications for teams and considerable psychological impact on athletes, highlighting the importance of adopting sustainable risk management strategies. This paper aims to evaluate the main factors affecting repeated injuries among National Basketball Association (NBA) players within a recurrent-event analysis framework based on a dynamic event-count model. For this purpose, the study relies on three big datasets regarding players and games over a ten-year period, each characterized by complex structures. A key contribution of this work is the introduction of the *Weakness*, a variable designed to model the momentary vulnerability of players due to previous injuries and to capture a carryover effect of past injuries on future risk. This measure reflects the increased vulnerability of a player when multiple injuries occur within a short period, with the effect gradually decreasing as the time without new injuries increases. Beyond *Weakness*, both recovery time and player-specific attributes were investigated as potential factors determining the risk of relapse. Empirical results show that *Weakness* and recovery time significantly affect the risk of injury recurrence. Furthermore, age and playing position emerge as additional risk factors. The evidence highlights the potential of data-driven models to inform decision making in sports management, supporting sustainable strategies to reduce injury risk and enhance athlete availability, ultimately contributing to the long-term sustainability of their performances.

Keywords Basketball analytics · Frailty · Andersen-Gill model · Injury risk · Recurrent-event analysis · Self-exciting process

Extended author information available on the last page of the article

1 Introduction

In recent years sports analytics has become a hot topic in scientific research. Data science tools are commonly used to address various inquiries, so a lot of contributions have been published dealing with several sports (among others, see Albert et al. 2005; Albert and Koning 2007; Winston 2012; Alamar 2013; Severini 2014; Miller 2015; Passos et al. 2016; Zuccolotto et al. 2017; Groll et al. 2018, 2019; Albert et al. 2017; D'Urso et al. 2023; Carlesso et al. 2025; Cefis and Carpita 2025). With regards to basketball, numerous research studies have been carried out addressing a range of objectives, including analysing performance, investigating the effects of high-pressure game situations, determining the characteristics that distinguish winning from losing teams, tracking playing styles, measuring players' psychological attributes and their impact on match results, studying medical issues related for example to injuries, physical and physiological characteristics of players, developing optimal fitness and training strategies, and many other issues (see Zuccolotto and Manisera 2020 for a review).

A key topic in sports analytics concerns injury prevention. Recurrent injuries represent a major challenge in professional sports, and their prevention is of utmost importance considering the financial impact on teams, the psychological effects on athletes, as well as the broader need for sustainable risk management strategies. In this perspective, injuries are increasingly viewed also through the lens of sustainability. Recurrent injuries not only affect short-term performance, but may compromise athletes' long-term health and career longevity, highlighting the need for sustainable workload and recovery management (Soligard et al. 2016; Gabbett 2016). From an organizational standpoint, sustainable injury prevention strategies are crucial to balance competitive performance with player welfare, as injuries have been shown to negatively affect team performance and continuity (Hägglund et al. 2013). Thus, several studies have been carried out to study injuries from different perspectives, including the identification of injury risk factors (e.g., Wilkerson et al. 2018; Bahr and Holme 2003; Hopkins et al. 2007; Kemler et al. 2015), and the understanding of the role of perceived injury risk (e.g. Deroche et al. 2012; Travert et al. 2017; Kontos 2004).

This study aims to investigate the main risk factors of injury recurrence in basketball. From a methodological point of view, we resort to survival analysis, a class of statistical techniques that offers a powerful framework to study how the risk of occurrence of an event changes over time, with a special focus on recurrent events. Although originally developed in medical research to investigate, for example, patient survival or disease recurrence, survival analysis has proven highly versatile and is now widely applied in other fields, including sports analytics. In this context, survival models have been used to address many research questions, including the timing and recurrence of key events in athletes' careers. For example, some studies have applied these approaches to study the duration of professional basketball players' careers (Fynn and Sonnenschein 2012), the determinants of coaches' dismissal (Wangrow et al. 2018; Tozetto et al. 2019), the determinants of athlete dropout across different sports (Pion et al. 2015; Moulds et al. 2020; Smith and Weir 2022; Back et al. 2023; Difernand et al. 2025), and the persistence of Olympic success (Csurilla and Fertő 2022; Gutiérrez et al. 2011). Other research questions addressed through survival analysis include the timing of substitutions in football (Del Corral et al. 2008), the time until the first goal is scored, and the intervals between consecutive goals in football matches (Pratas et al. 2016; Nevo and Ritov 2013; Leriou and Ntzoufras 2025) and ice hockey matches (Thomas 2007), and the identification of skills that predict higher scoring performance in basketball (Macis et al. 2023; Macis 2025a). Beyond these applications, survival analysis is particularly suitable for investigating injury risk factors and the importance of recovery times in sport, as documented by several papers (Beynnon et al. 2005; Hopkins et al. 2007; Buist et al. 2010; Venturelli et al. 2011; Schneider et al. 2013; Mahmood et al. 2014; Nelson et al. 2016; Mai et al. 2017; Dekker et al. 2017; McLeod et al. 2017; Lawrence et al. 2018; Kontos et al. 2019; Howell et al. 2019; Sochacki et al. 2019; Jack et al. 2019; Ekeland et al. 2020; Lu et al. 2022; Moore et al. 2023).

Traditional survival models typically focus on the time until the first occurrence of an event. In our study, however, players may experience multiple events; therefore, we adopt a recurrent-event analysis approach, an extension of survival analysis that allows for the modelling of repeated events over time. Such approach has been widely discussed and applied, not only in the context of sports injuries (e.g., Ullah et al. 2014; Zumeta-Olaskoaga et al. 2021, 2025; Macis 2025b), but also, for example, to the analysis of the timing of corner kicks in football (Peng et al. 2025).

Specifically, this study extends the work of Macis (2025b) to evaluate the risk of injury recurrence (i.e. the occurrence of injuries after the first one) among National Basketball Association (NBA) players over ten seasons, from 2010–2011 to 2019–2020. Hereafter, 'new injury', 'recurrent injury', and 'relapse' will be used interchangeably.

More in detail, the aim is verifying whether the risk can be affected by both player features, such as the Body Mass Index (BMI), the playing position and the age, and injury-related features. Among these, the number of rest days and the recent occurrence of previous injuries are considered. In particular, since the same player may experience multiple injuries within a short time period, resulting in clustered events, we define a new variable called *Weakness*, capable of measuring the momentary vulnerability of a player, resulting from the occurrence of multiple injuries in close succession. This variable, defined as a function of the time elapsed between two consecutive injuries, is designed to capture a carryover effect (Çiğşar and Lawless 2012a; Nirmalkanna and Çiğşar 2025), i.e. a transient effect that increases the event intensity after the occurrence of certain conditions, such as a previous injury. Therefore, this variable is an internal time-dependent covariate, as it depends on the event process and varies over time; its inclusion in the model leads to a dynamic event-count model (Aalen et al. 2008).

The final data used for these analyses have been obtained by a non-trivial match and harmonization of three large databases: (*i*) data about all the injuries that occurred in the ten observed seasons; (*ii*) a play-by-play dataset of the ten analysed seasons, used to recover information about minutes played; (*iii*) a dataset containing information about players' birthdate, height, weight and position.

In particular, data on injuries have been obtained from a dataset made available with license 'CC0: Public Domain' by Randall Hopkins on the Kaggle repository. The play-by-play dataset has been provided by BigDataBall, a data provider widely recognized as a reliable source of validated and verified statistics for several sports associations, supporting numerous academic studies. Finally, information about players has been retrieved by the SQLite database shared on Kaggle by Wyatt Walsh under the 'CC BY-SA 4.0' license.

The article is structured as follows. In Sect. 2 we introduce the methodological framework for recurrent-event analysis, focusing on the model specification (Sect. 2.1), the development of the novel time-varying variable *Weakness* to capture players' momentary vulnerability (Sect. 2.2), and the estimation of individual frailty (Sect. 2.3). In Sect. 3 we describe the procedures used to match different information sources and the steps followed for obtaining the final dataset. Following, in Sect. 4 we provide a description of the data and report the results of the analysis. Finally, in Sect. 5 we draw some concluding remarks and outline the directions for future research.

2 Methodological framework

Survival analysis is a well-known statistical approach used to study the time until a specific event occurs, the so-called survival time, within a given follow-up (observation) period. However, in many studies, the exact survival time cannot be observed for all individuals due to censoring, which occurs when a subject has not experienced the event of interest by the end of the follow-up or is lost to follow-up (Collett 2015).

Classical approaches, such as the widely used Cox proportional hazards (Cox PH) regression model (Cox 1972), rely on the assumption that survival times are independent. In practice, however, this assumption may not be valid. A common example occurs when individuals experience recurrent events, leading to dependent survival times within the same subject.

To account for this dependence structure, extensions of classical survival models have been developed within the framework of recurrent-event analysis. In this context, the event history of an individual can be represented by a counting process $\{N(t), 0 \leq t\}$, which records the cumulative number of events observed until time t:

$$N(t) = \sum_{k=1}^{\infty} I(T_k \leq t)\,,$$

where T_k denotes the time of the k^{th} event.

Models for recurrent events aim to estimate the intensity function, i.e. the instantaneous probability of an event occurring at time t, conditional on the process history $\mathcal{H}(t)$ (Cook and Lawless 2007):

$$\lambda(t|\mathcal{H}(t)) = lim_{\Delta t \to 0} \frac{Pr\{\Delta N(t) = 1|\mathcal{H}(t)\}}{\Delta t}, \tag{1}$$

where $\Delta N(t)$ represents the number of events in the interval $[t, t + \Delta t)$ and $\mathcal{H}(t) = \{N(s) : 0 \leq s < t\}$ represents the history of the process up to time t.

2.1 The Andersen-Gill model

Different time scales may be chosen, leading to different model formulations (Ullah et al. 2014). In this work, we consider the so-called *calendar time*, which relies on a unified timescale for all events, defined with respect to a fixed time origin, and ensures that risk periods for a given player do not overlap.

Within this framework, we use the Andersen-Gill (A–G) model (Andersen and Gill 1982), a semiparametric regression model, that specifies the intensity of the counting process $\lambda(t|\mathcal{H}(t))$ as a function of the observed covariates $\boldsymbol{x}_i(t)$, which may be *fixed* or *time-varying*. Moreover, covariates can be classified as *external* or *internal*. An external covariate assumes values independently of the recurrent event process, while internal covariates depend on the event history of the process itself (e.g., number of previous events).

The classical A-G model is specified as:

$$\lambda_i(t|\mathcal{H}(t)) = Y_i(t) \exp(\beta^T \boldsymbol{x}_i(t))\lambda_0(t), \tag{2}$$

where $Y_i(t)$ is the at-risk indicator, equal to 1 if the i^{th} individual is under observation, and thus at risk at time t, and 0 otherwise; $\boldsymbol{\beta}$ is the vector of regression coefficients associated with the p observed covariates; and $\lambda_0(t)$ is the baseline intensity function. In the A–G formulation, $\lambda_0(t)$ is common to all individuals, left unspecified, and assumed not to follow any particular parametric form (Cook and Lawless 2007).

The coefficients can be interpreted in terms of intensity ratio, by considering their exponential: $\exp(\beta_k)$, with $k = 1, ..., p$, denotes the multiplicative effect of a one-unit increase in the k^{th} numerical covariate, holding all other covariates constant (Cook and Lawless 2007). In particular, the intensity ratio represents the ratio of the event intensity for two individuals differing in a given covariate, all else being equal. If the intensity ratio is greater than one, a one-unit increase in the covariate is associated with a higher risk of event occurrence; if it equals one, there is no effect; if it is lower than one, it is associated with a lower risk. For categorical covariates, the same interpretation applies, comparing a given category with the reference category rather than considering a one-unit increase.

2.2 The *Weakness* variable

Determining the appropriate timing for return to play is a central issue for coaches and medical staff in the management of injured players, as it involves a trade-off between two competing objectives: allowing adequate recovery to minimize the risk of reinjury, and reintegrating the player into competition as early as reasonably possible. A premature return may increase the short-term risk of a subsequent injury, thereby contributing to injuries occurring in close succession, whereas an excessively delayed return may unnecessarily prolong absence from competition and reduce team availability.

In the analysis of recurrent events, the temporary increase in the risk of a subsequent event after a previous one, a form of short-term event dependence, is commonly referred to as a carryover effect, that is, a transient increase in event intensity driven by the individual's event history (Çiğşar and Lawless 2012b; Nirmalkanna and Çiğşar 2025).

One possible way to model carryover effects is by means of an internal time-dependent covariate, i.e. a covariate whose value at time t depends on the subject's event history up to that time. In this spirit, Çiğşar and Lawless (2012a) considered a formulation based on a binary time-dependent covariate that takes value 1 only within a prespecified interval following the occurrence of an event, usually referred to as the *risk window* Δ, and 0 otherwise:

$$X(t) = I(N(t^{-}) > 0)I(B(t) \leq \Delta)\,,$$

where $I\{\cdot\}$ denotes the indicator function, $N(t^{-})$ is the number of events observed prior to time t, and $B(t) = t - T_{N(t-)}$ is the *backward recurrence time*, namely the elapsed time since the most recent event preceding t.

Building on this class of internal time-dependent formulations, the central contribution of this paper is the introduction, within the A–G framework, of a new time-varying internal covariate, termed *Weakness*, specifically designed to model carryover effects in recurrent events. The rationale underlying *Weakness* is that the occurrence of a previous injury may induce a temporary state of increased vulnerability, thereby increasing the risk of a subsequent event when injuries occur in close succession. More specifically, the proposed covariate is constructed to retain memory of recent injuries when the time elapsed between consecutive events is short, while this memory progressively fades, and ultimately decays towards zero, as the gap between events increases. Accordingly, *Weakness* yields a nonlinear autoregressive representation of the carryover effect, with its value at time t recursively determined by its value at time $t-1$ and by nonlinear excitation and decay components reflecting the subject's injury history. In this setting, the inclusion of an internal time-varying covariate such as *Weakness* places the resulting A–G specification within the class of dynamic models (Aalen et al. 2008), which can be interpreted as a *modulated Poisson process* (Cook and Lawless 2007).

To capture the idea that the effect of a previous injury should depend on the time elapsed since its occurrence, we first introduce a nonlinear weighting function of the backward recurrence time. This function is intended to assign greater weight to shorter backward recurrence times and progressively smaller weight to longer ones, thereby reflecting the gradual fading of injury-related vulnerability over time. Specifically, we consider the following logistic weighting function:

$$f(B(t)) = \frac{e^{\alpha(B(t)-q)}}{1 + e^{\alpha(B(t)-q)}} \,,$$

where $B(t)$ denotes the backward recurrence time, the parameter α controls the steepness of the function, and q is the value such that $f(B(t)) = 0.5$. In practice, q may be fixed according to the time horizon within which a new injury is still likely to be regarded as a relapse of the previous one. When $\alpha < 0$, $f(B(t))$ is decreasing in $B(t)$, with larger absolute values of α corresponding to a sharper decline. Moreover, the condition $\alpha \leq -\log\big((1-\epsilon)/\epsilon\big)/q$ ensures that $f(0) \geq 1-\epsilon$, for arbitrarily small $\epsilon \in (0, 1)$. Two examples of $f(B(t))$ are shown in Fig. 1.

The *Weakness* for the i^{th} player at time t is then defined recursively as

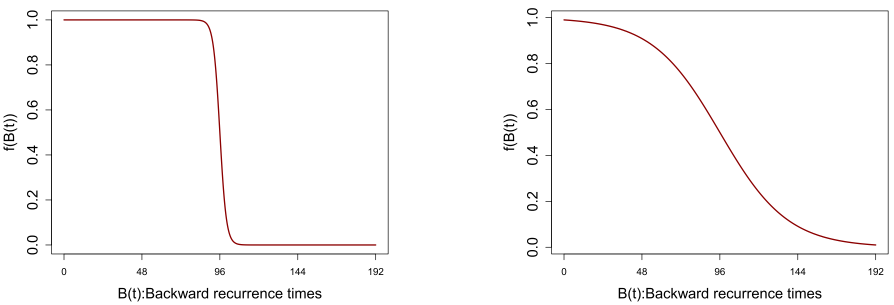
$$W_i(t) = \left[W_i\big(T_{i,N_i(t-)}\big) + f(B_i(t))\right] \cdot f(B_i(t)) \,,$$

where $B_i(t)$ is the backward recurrence time for the i^{th} player, and $T_{i,N_i(t-)}$ denotes the time of the most recent injury observed before time t.

For the purpose of constructing the covariate to be included in the A–G model, *Weakness* is evaluated only at the times of new injury events.

In this case, the backward recurrence time coincides with the elapsed time between two consecutive injuries, namely $B_i(T_{i,N_i(t)}) = \Delta T_{N_i(t)}$, where $\Delta T_{N_i(t)}$ denotes the time between consecutive injuries, expressed in minutes played. Accordingly, the value of *Weakness* at each event time is determined by both the player's previous level of vulnerability and the time elapsed since the most recent injury.

This construction has a natural interpretation. When injuries occur in close succession, that is, when $\Delta T_{N_i(t)} \approx 0$, the function $f(\cdot)$ is close to 1, and *Weakness* is approximately equal to the cumulative number of injuries suffered over that period. By contrast, when a new injury occurs after a long interval, for example when $\Delta T_{N_i(t)} \gg q$, $f(\cdot)$ approaches 0, and *Weakness* becomes approximately 0. Conceptually, this means that injuries occurring close in time are more likely to be related to one another, so that the player's temporary vulnerability increases with the accumulation of relapses. Conversely, when two consecutive injuries are separated by a long period, it is plausible that they are unrelated, and the vulner-



(a) $\alpha = -0.5$, $q = 96$ (b) $\alpha = -\frac{\log\frac{1-\epsilon}{\epsilon}}{q}$, $q = 96$, $\epsilon = 0.01$

Fig. 1 Logistic function $f(B(t))$ that weights the time elapsed since the most recent event prior to t

ability induced by the previous injury can reasonably be regarded as negligible. A practical illustration of the behaviour of *Weakness* is provided in Sect. 3 (Fig. 3 and Table 1).

The proposed *Weakness* covariate differs from standard carryover formulations in three main respects. First, it incorporates an autoregressive component through the recursive term $W_i(T_{i,N_i(t^-)})$, so that the current level of *Weakness* depends on its most recent past value. Second, it includes a self-exciting component through the additive contribution of $f(B_i(t))$, which allows a new injury to increase the current level of vulnerability. Third, it embeds a decay mechanism through the multiplicative effect of $f(B_i(t))$, which ensures that the impact of previous injuries decreases smoothly as the time since the last event increases. Taken together, these features make *Weakness* a nonlinear autoregressive representation of the carryover effect.

2.3 The frailty component

The A–G model described above, including the proposed internal time-varying covariate *Weakness* and the other observed covariates, may still fail to fully account for unobserved heterogeneity across players. Such residual heterogeneity may induce additional within-subject dependence among recurrent injury events. To account for this feature, subject-specific latent random effects, usually referred to as individual *frailties*, can be incorporated into the model (Vaupel et al. 1979). In this setting, the frailty term represents a player-specific propensity to experience injuries that is not explained by the observed covariates.

The corresponding intensity function, conditional on the individual frailty and on the observed covariates and event history available just before time t, can be written as

$$\lambda_i(t \mid Z_i, \mathcal{H}_i(t)) = Y_i(t)\, Z_i\, \lambda_0(t) \exp\{\beta^T \boldsymbol{x}_i(t)\}, \tag{3}$$

where $Y_i(t)$ is the at-risk indicator, $\lambda_0(t)$ is the baseline intensity, $\boldsymbol{x}_i(t)$ denotes the vector of observed covariates, including *Weakness*, and Z_i is the frailty term for the i^{th} player.

The frailty is usually assumed to be a positive random variable with mean equal to 1, so that the baseline intensity retains its population-level interpretation, and variance θ, which quantifies the degree of unobserved heterogeneity. In the proposed framework, frailty and *Weakness* represent two distinct sources of heterogeneity in injury risk. The frailty term accounts for stable, unobserved differences across players in their underlying propensity to experience injuries. These differences may arise from individual characteristics that are not measured, or are only imperfectly captured by the observed covariates. Thus, for the same observed covariate profile and follow-up time, a player with higher frailty is expected to experience more recurrent injuries than a player with lower frailty (Balan and Putter 2020). Some deeper insights about the meaning of the frailty component are addressed in Macis (2025b), where a practical explanation of how the frailty should be interpreted is provided.

By contrast, *Weakness* represents a dynamic, internal time-varying component of risk. It is designed to capture a transient state of increased vulnerability induced by recent injuries, and therefore reflects how the player's own injury history may temporarily modify the risk of a subsequent event.

This distinction is particularly important when carryover effects are of interest. If stable unobserved heterogeneity is ignored, the model may incorrectly attribute part of the excess recurrence risk due to unmeasured player-specific characteristics to event-history dependence. As a result, the carryover effect may be overestimated (Nirmalkanna and Çiğşar 2025). The inclusion of a frailty component therefore allows latent inter-individual heterogeneity to be accounted for jointly with the observed dynamic event-history structure represented by *Weakness*. In this sense, frailty and *Weakness* are complementary rather than alternative components: frailty captures persistent differences between players, whereas *Weakness* captures transient within-player changes in vulnerability following recent injuries.

Different distributions can be assumed for the frailty term. A common choice is the Gamma distribution, typically parametrized to have mean 1 and variance θ, where θ quantifies the degree of unobserved heterogeneity across subjects. Other commonly used choices include the inverse Gaussian, positive stable, and log-normal distributions (Hougaard 2000; Cook and Lawless 2007; Balan and Putter 2020). These distributions differ in the way they represent between-subject variability and in the resulting marginal dependence structure among recurrent events from the same subject. In applied recurrent-event models, the Gamma and log-normal specifications are among the most frequently used, partly because they are widely available in standard statistical software.

Several R implementations are available for recurrent-event models with frailty terms, including `survival`, `frailtypack`, `frailtyEM`, and `coxme`. These packages differ in their modelling focus, supported frailty distributions, and estimation procedures. In particular, `frailtypack` and `frailtyEM` provide more specialized tools for frailty modelling, whereas `coxme` is oriented toward Cox mixed-effects models with Gaussian random effects. In the present work, we used the `coxph` function from the `survival` package (Therneau and Lumley 2015), because it provides a direct and flexible implementation of the Andersen–Gill counting-process formulation, accommodates interval-specific time-varying covariates, and allows the inclusion of a player-level frailty term within the same modelling framework. This was particularly convenient for the proposed analysis, in which *Weakness* was coded as a time-varying covariate and the frailty term was specified at the player level.

3 Risk of recurrent injuries among NBA players

The modelling framework described in the previous sections was applied to the analysis of recurrent injuries among NBA basketball players observed over ten consecutive seasons, from 2010–2011 to 2019–2020.

In this application, a key methodological choice concerns the definition of the time scale. Instead of indexing follow-up in elapsed calendar days, we used cumulative playing time, defined as the total number of minutes spent on court by each player.

This choice reflects the fact that, in professional basketball, exposure to injury risk is more directly related to time actively spent in competition than to the mere passage of chronological time. Indeed, NBA basketball games sometimes take place on dates that are very close to each other, and two injuries occurring within a short calendar interval may have different interpretations depending on how many minutes the player actually played between them (Macis 2025b). Accordingly, in the present analysis, the recurrent-event process was indexed by exposure time rather than chronological time.

Study time was therefore defined as cumulative playing time and was not reset after each injury event. For each player, the origin of follow-up was set at the return to play after the first recorded injury. Subsequent injury events were then represented on the cumulative playing-time scale, so that each event time corresponded to the total number of minutes played by that player since the start of follow-up. This formulation allowed the Andersen–Gill counting-process structure to be implemented using cumulative minutes played as the time scale.

Fig. 2 provides an example of the resulting data structure for four players (A, B, C, D) who experienced three, four, zero, and two relapses, respectively, during follow-up. The follow-up period starts at time 0, corresponding to each player's return to play after their first recorded injury. Three players ended the study censored, namely A, B, and C, whereas player D experienced an injury at the end of the observation period. For example, player A experienced a first relapse after 1400 minutes of play. After returning to play, he accumulated an additional 1600 minutes before sustaining a second recurrent injury, corresponding to 3000 cumulative minutes from the start of follow-up. He then played a further 1800 minutes, reaching 4800 cumulative minutes at the time of the third new injury. Finally, he reached 9345 cumulative minutes by the end of follow-up, corresponding to a censored observation.

To further clarify how the proposed internal time-varying covariate operates in the present data structure, Fig. 3 provides an example of the behaviour of *Weakness* for one player. The example illustrates how $W(t)$ changes according to the elapsed playing time between consecutive injuries and how it translates the player's recent injury history into a dynamic measure of temporary vulnerability. The player experienced seven relapses after the first recorded injury and was then censored after having accumulated 2000 minutes of playing time. The white diamond in Fig. 3 denotes the first recorded injury, the black squares indicate the observed recurrent injuries and their corresponding event times, and the white dot denotes the end of follow-up without any further injury events.

In this example, several injuries occurred in close succession, whereas others were separated by longer intervals of playing time. This pattern allows the main features of *Weakness* to be visualized: $W(t)$ increases when recurrent injuries occur after short playing-time intervals and decreases when the elapsed playing time since the previous injury becomes longer.

Figure 3 combines three graphical elements that should be interpreted jointly. First, the black squares indicate the observed injury events and their corresponding event times on the cumulative playing-time scale. Second, the red points indicate the observed values of *Weakness* assigned to the player at the beginning of each at-risk interval. These are the values used as covariate values in the A–G model. Third, the blue dashed curves show how $W(t)$ would evolve as a function of playing time after each injury, under the hypothetical scenario that a subsequent injury occurred earlier than actually observed. These curves are therefore useful for visualizing the decay mechanism of *Weakness* between consecutive events.

The figure shows that $W(t)$ increases when injuries occur in close succession and decreases when the elapsed playing time between consecutive injuries is longer. At the beginning of follow-up, the player has a value of *Weakness* equal to 0.205, reflecting the first injury occurring after 100 minutes from the start of the first playing season (grey region of Fig. 3). The first four relapse then occur after relatively short intervals of 117, 53, 45, and 96 minutes, leading to a rapid increase in *Weakness*. In contrast, the subsequent injuries occurring after 989 and 360 minutes are preceded by longer playing-time intervals, so that *Weakness* progressively decays towards zero. Finally, the last observed recurrent injury occurs after 40 minutes, resulting in a renewed increase in *Weakness*, with an observed value equal to 0.876.

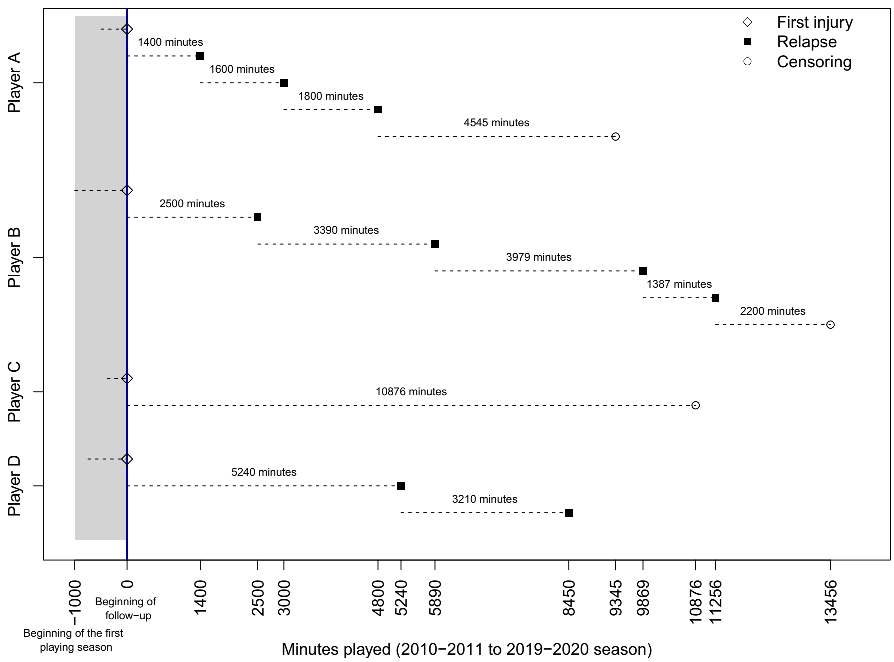


Fig. 2 Data structure for four players, with white diamonds representing the first injury, black squares representing relapses and white dots indicating censored observations

Table 1 reports some examples of the values corresponding to the curves displayed in Fig. 3. These values illustrate how *Weakness* would change if a relapse occurred at different times after the previous injury. For example, at the beginning of follow-up the player has a value of *Weakness* equal to 0.205; if the first recurrent injury had occurred after 10 minutes of play, $W(t)$ would have increased to approximately 1.169, whereas if it had occurred after 20 minutes, $W(t)$ would have been equal to 1.149. The remaining values can be interpreted in the same way, as hypothetical values of *Weakness* evaluated at different elapsed playing times since the previous injury.

Because *Weakness* is an internal time-varying covariate, its value is determined by the previous injury history of the player. However, when it is included in the A–G model, the value used to model the risk at a given time must be based only on information already available before that time. For this reason, in each at-risk interval from T_{start} to T_{stop}, *Weakness* is assigned the value known at T_{start} , that is, before observing whether the interval will end with a relapse or with censoring. The injury event occurring at *Tstop*, when present, is then used to update *Weakness* only for the following interval. This convention preserves the temporal ordering between covariate values and event occurrence and avoids using future information to explain the event that closes the current interval.

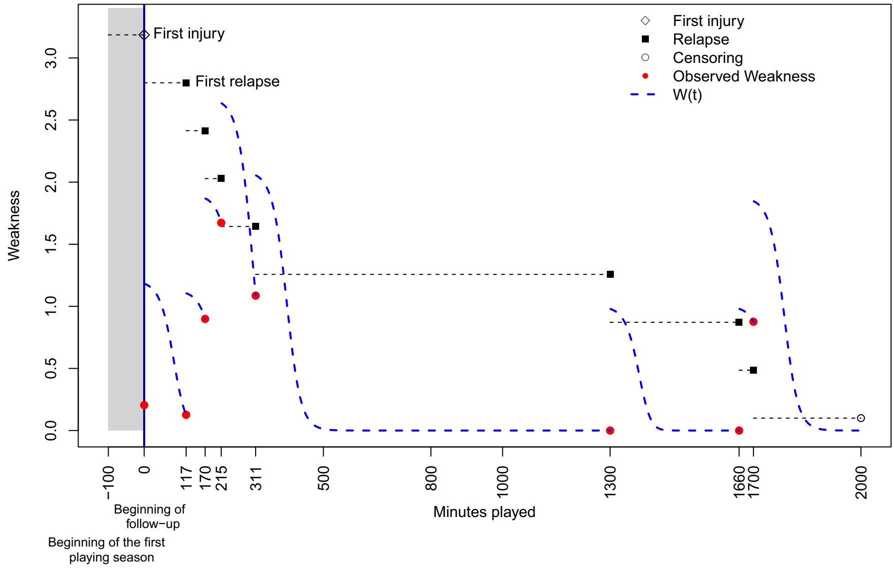


Fig. 3 Example of *Weakness* function for a given player. Example for a player who suffered seven relapses (black squares) after the first injury (white diamond). The player ends the study without new relapses, and therefore 2000 is a censoring time (white dot). Red dots denote the values of *Weakness* at the beginning of each interval. Blue dashed lines indicate the values of *Weakness* function $W(t)$. For the evaluation of the weighting function the following values have been used: $q = 96$, $\epsilon = 0.01$ and $\alpha = -\frac{\log \frac{1-\epsilon}{\epsilon}}{q}$.

Table 1 Values of the Weakness function shown in Fig. 3

Event type	$N(t^-)$	$T_{N(t^-)}$	t	$B(t)$	$f(B(t))$	$W(T_{N(t^-)})$	$W(t)$
1^{st} injury	–	−100	0	–	–	0*	0.205
	0	0	0^+	0	$1-\epsilon=0.99$	0.205	1.183
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	20	20	0.974	0.205	1.149
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	117^-	117	0.268	0.205	0.127
1^{st} relapse ($T=117$)	0	0	117	117	0.268	0.205	0.127
	1	117	117^+	0	$1-\epsilon=0.99$	0.127	1.106
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	117	137	20	0.974	0.127	1.073
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	117	170^-	53	0.849	0.127	0.899
2^{nd} relapse ($T=170$)	1	117	170	53	0.849	0.127	0.899
	2	170	170^+	0	$1-\epsilon=0.99$	0.899	1.870
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2	170	190	20	0.974	0.899	1.825
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2	170	215^-	45	0.920	0.899	1.673
3^{rd} relapse ($T=215$)	2	170	215	45	0.920	0.899	1.673
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4^{th} relapse ($T=311$)	3	215	311	96	0.500	1.673	1.086
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5^{th} relapse ($T=1300$)	4	311	1300	989	0.000	1.086	0.000
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6^{th} relapse ($T=1660$)	5	1300	1660	360	0.000	0.000	0.000
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7^{th} relapse ($T=1700$)	6	1660	1700	40	0.936	0.000	0.876
	7	1700	1700^+	0	$1-\epsilon=0.99$	0.876	1.847
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	7	1700	1720	20	0.974	0.876	1.803
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	7	1700	2000^-	300	0.000	0.876	0.000
End of follow up (T=2000)	7	1700	2000	300	0.000	0.876	0.000

* This value is set equal to 0 by assumption.

The construction of *Weakness* requires the specification of the parameters governing the logistic weighting function, namely α, q, and ϵ. In the present case study, ϵ was fixed at 0.01, so that the weighting function was constrained to take values close to one immediately after an injury. The parameter q was set equal to 96 minutes, corresponding to the time point at which the assigned weight is equal to 0.5. Alternative values of q were also explored in sensitivity analyses, as reported below.

The parameter α controls both the direction and the steepness of the logistic weighting function. Since the weight $f(B(t))$ should decrease as the playing time elapsed since that injury increases, only negative values of α were considered. Within this range, larger absolute values of α correspond to a steeper decrease, whereas values closer to zero produce a more gradual decay. This parameter was selected empirically. Given the values of q and ϵ, different plausible negative values of α were evaluated by fitting separate A–G frailty models and comparing their performance using the Bayesian Information Criterion (BIC) and the cross-validated Concordance Index. Across the explored values, the selected value was $\hat{\alpha} = -\log((1-\epsilon)/\epsilon)/q$. This value is the least negative value compatible with the constraint $f(0) \geq 1-\epsilon$, and therefore corresponds to the most gradual decay of the weighting function among the values satisfying this constraint.

3.1 Data preparation

The analysis is based on a dataset we call `Data`, built by combining information from three different big databases. The first information source is a dataset, `Inj`, in which all the injuries sustained by NBA players over the 2010–2011 to 2019–2020 seasons are recorded[1]. This dataset includes information about the date of the event (which may indicate either the onset or the end of the sick leave), the team of the involved player, the name of the player, and some notes providing a brief description of the event, for a total of 27,105 events recorded. The second data source is the dataset `PbP`, containing play-by-play records of all the games played in the analysed seasons (nearly six million match events recorded, covering more than 12,000 games, made available by BigDataBall[2]). Lastly, we retrieved information about players from the dataset $\texttt{Inf}_{\texttt{Pl}}$, a SQLite database including records of more than 64,000 games, 4800 players, and 30 teams[3].

To combine these three datasets, some preliminary cleaning and alignment operations were necessary on the `Inj` dataset:

- deletion of sick leaves unrelated to injuries (e.g. vertigo, COVID-19, migraines, etc.), as these conditions are not directly related to physical injuries;
- alignment of dates for injuries officially recorded during the interval between games to the immediately preceding match played (recovered from `PbP`). This assumption was necessary when an injury was recorded in the injury list on a date different from that of a match (e.g., injuries sustained during training sessions). In such cases, the assumption was considered valid, given that the outcome of interest was minutes played in official matches;
- treatment of extended sick leaves without player's return to the line-up as continuations of the previous injury. Indeed, in these cases, in the absence of evidence of a new injury event, it was reasonable to assume that the extension is related to the injury already recorded.

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

[2] www.bigdataball.com

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

In addition, a consistency check was performed due to small differences in the way the same player name is sometimes recorded in the three different databases (*e.g* 'Juan Hernangomez' and 'Juancho Hernangomez', or 'Kevin Knox' and 'Kevin Knox II'). After the check, the players' names were made consistent across the three databases `Inj`, `PbP` and $\mathtt{Inf}_{\mathtt{Pl}}$, ensuring a correct and reliable merge of the datasets.

All major steps undertaken to create the `Data` dataset are explained below and summarized in the pseudo-code Algorithm 1:

1. The first step consisted of computing the total minutes played by each player in each match using `PbP` data, and recording this information in an auxiliary dataset, called `MPl`.
2. Next, information on injury occurrence was added to each observation in the `MPl` dataset by aligning it with the `Inj` dataset. Based on this merge, a binary event indicator was created to denote whether the player had an injury recorded on that date or shortly thereafter. As a result, the `MPl` dataset now contains one row per player per match, including information about minutes played and injury occurrence.
3. The next step consisted of retrieving player-specific characteristics, such as height, weight, date of birth, and playing position, from the $\mathtt{Inf}_{\mathtt{Pl}}$ dataset, and merging them into the `MPl` dataset. Moreover, each player's BMI was evaluated using the available height and weight information.
4. Then, for each injury event, the cumulative minutes played between consecutive injuries were computed. For the first injury, the total minutes played from the beginning of the observation period were considered. Additionally, the cumulative minutes played from the last injury (or from the start of the observation period if no injury occurred) to the end of the observation period were computed. For each row in the dataset, two time variables were defined: T_{start}, representing the cumulative minutes played at the start of the interval (0 for the first time interval), and T_{stop}, representing the cumulative minutes played at the end of the interval, corresponding to an injury occurrence. Consequently, `MPl` now contains multiple rows per player, corresponding to each injury event, plus an additional row representing the final observation period.
5. For each resulting time interval, the player's age, the number of days at rest, and *Weakness* variable (see Sect. 2.2) were calculated at the beginning of the interval.
6. Finally, since the focus of this analysis is on injury relapses, players who never sustained an injury were excluded from the dataset. Furthermore, for players who sustained at least one injury, the interval corresponding to the first injury was removed. Then, the observation period effectively starts at 0 when the player returns to the court after the first recorded injury.

To stand in the A–G framework, the final dataset, denoted as `Data`, is therefore organized in a counting-process format, i.e. a long form structure, and contains data about all the players who sustained at least one injury during the ten analysed seasons. More in detail, for each player there are as many rows as the number of injury relapses, plus, possibly, an additional row recording the data about the time from the last injury to the end of the follow-up. These rows represent the player-specific at-risk intervals. For each interval, the starting time T_{start} was defined as the cumulative number of minutes played at the beginning of the interval, whereas the stopping time T_{stop} was defined as the cumulative number of minutes played at the time of the next injury event or censoring. The event indicator was equal to 1 when the interval ended with a recurrent injury and 0 when the interval ended with censoring. Therefore, censored observations correspond to players who did not experience any relapses during the observation period or who were not injured at the end of follow-up.

In this structure, each player could contribute several rows to the dataset, corresponding to successive at-risk intervals between return to play and the next injury event, or between return to play and the end of observation. Thus, recurrent injuries were represented as multiple event times on the same cumulative playing-time scale. Importantly, because the time scale was not reset after each event, T_{start} and T_{stop} retained their interpretation as cumulative minutes played since the origin of follow-up.

The recorded variables in the dataset include the timepoints T_{start} and T_{stop} defining each player-specific at-risk interval, the injury indicator, and many covariates, including the player's main playing position, the player's BMI, the age - evaluated at the beginning of each time interval -, *Weakness*, and the number of rest days after the previous injury. In particular, this last variable measures the number of days between the injury date and the date of the player's return to the court. In some cases, this measure may also include days off due to the end of the playing season, thereby capturing the total number of days in which the player was effectively resting and not participating in official matches, rather than considering only days prescribed for recovery. It is also important to note that, under this time-scale formulation, age, *Weakness*, and the number of rest days are all treated as time-varying covariates, since their values may change over time, particularly at the beginning of each at-risk interval.

Require: `Inj` ← the injury dataset, as described above
Require: `PbP` ← the play-by-play records, after players' names harmonization
Require: $\texttt{Inf}_{\texttt{Pl}}$ ← the dataset with additional information about players, after players' names harmonization

for each match in the dataset `PbP` and each player who played in that match **do**
  compute the total minutes played
  `MPl` ← the dataset with one row for each player in each match, with variables the date of the match and the total minutes played
end for
for each row in the dataset `MPl` **do**
  by joining by match date, retrieve information from dataset `Inj` regarding the occurrence (1) or non-occurrence (0) of an injury
end for
for each player in the dataset `MPl` **do**
  retrieve and record information about the player (height, weight, birth date, playing position) from dataset $\texttt{Inf}_{\texttt{Pl}}$
  evaluate the player's BMI as the ratio between weight and the square of height.
  extract all the played matches, with the 0-1 variable describing occurrence or non-occurrence of the injury
  for each occurrence of an injury **do**
    cumulate the minutes played between consecutive injury events (for the first injury: the minutes played since the beginning of the observation time)
    compute age of the player at the beginning of each time interval
    compute the number of rest days for each injury, recording the information at the beginning of each subsequent time interval.
    compute *Weakness* variable at the beginning of each subsequent time interval.
  end for
  cumulate the minutes played from the last injury (if no injury has occurred: from the beginning of the observation time) to the end of the observation time
  compute age of the player at the beginning of the last time interval
  compute *Weakness* variable at the beginning of the last time interval.
end for
`MPl` ← the dataset with each row corresponding to a player's injury event, plus an additional row representing the final observation period. Each injury event is associated with two timepoints, T_{start} and T_{stop}, representing the cumulative minutes played at the beginning and at the end of each interval.
retain only players with at least two injury events
remove the interval corresponding to the first injury
set the time origin to 0 at the player's return to court after the first injury
return dataset `Data`, in a life-table form, with each row describing an injury or the censoring, with variables the event date, the player, the minutes played since the previous event of the same player (or since the beginning of the observation time), the BMI, age at the beginning of each time interval, players' playing position, the number of days at rest, and *Weakness* variable.

Algorithm 1 Data preparation

4 Results

In this Section, the description of the sample and the results of the fitted dynamic event-count model are presented.

4.1 Data description

During the follow-up, i.e. the period ranging from the 2010–2011 season to the 2019–2020 season, a total of 1066 players have been observed, and, overall, 8367 relapses have been recorded. More in detail, in the overall sample, the mean number of observed events (new injuries) was equal to 7.8, with a standard deviation (SD) equal to 8.5.

It emerges that, among the sample, 89 (8.3%) players did not suffer any relapses, while the remaining 977 players (91.7%) experienced one or more new injuries after the first one. Among these, 258 players ended the study still injured. Moreover, only 10% of the players suffered a single relapse, while 81.7% of the players ($n = 870$) sustained at least two new injuries.

The players have been classified according to their main playing position on the court: 508 (47.7%) were guards, 296 (27.8%) were forwards, and 262 (24.6%) were centers.

On average, throughout the observation period, the observed players played 4844.1 minutes (SD= 5867.5).

Since the number of relapses may also be influenced by the number of minutes played, the injury incidence rate per 100 player-match exposures has also been evaluated (Lee and Zumeta-Olaskoaga 2022):

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

where $I_{r,i}$ is the injury incidence rate of the i^{th} player, n_inj_i is the number of relapses suffered, and ΔT_i denotes the total minutes played by the player during the follow-up period.

The mean injury incidence rate in the sample was equal to 56.8, corresponding to almost 57 new injuries per 100 players per 48 minutes of play. Table 2 shows some summary statistics, after grouping players into four classes defined by the quartiles of the injury incidence rate. In particular, conditional column frequencies of playing position and partial means of BMI in the four groups are reported. Results show that, compared with the overall statistics, a larger proportion of centers had higher injury incidence rates, whereas a greater proportion of guards had lower rates. For example, it can be seen that in the third and fourth group nearly 30% of players are centers, while this percentage decreases to about 20% in the first two groups with lower injury incidence rates. Therefore, given the same number of minutes played, a higher percentage of guards would be expected to sustain fewer injuries than forwards and centers.

Table 2 Summary statistics of the sample by injury incidence rate

	< 4.47	[4.47; 9.55)	[9.55; 25.13)	≥ 25.13	Overall
Playing position					
Center, n (%)	58 (21.7%)	53 (19.9%)	79 (29.7%)	72 (27.0%)	262 (24.6%)
Forward, n (%)	75 (28.1%)	71 (26.7%)	73 (27.4%)	77 (28.8%)	296 (27.8%)
Guard, n (%)	134 (50.2%)	142 (53.4%)	114 (42.9%)	118 (44.2%)	508 (47.7%)
BMI, Mean (SD)	24.9 (1.8)	25.0 (1.6)	24.9 (1.7)	24.8 (1.9)	24.9 (1.7)

Data referring to NBA players observed between the 2010–2011 and 2019–2020 seasons

4.2 Model estimation

The final aim of this study is to evaluate the risk of recurrent injuries. Different approaches can be used to model this kind of data (Ullah et al. 2014); here, we used the Andersen-Gill model with frailty described in (3).

The aim is to evaluate whether variables related to player characteristics (BMI, age at the beginning of each time interval, and playing position) and variables describing the injuries sustained (number of rest days after the previous injury and *Weakness*) affect the player's risk of relapse. As in Macis (2025b), a transformation of the BMI has been considered to account for possible non-linearities:

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

The first fitted model also included three interaction terms (between the position and, respectively, the BMI_{transf} and the *Weakness* and between the number of rest days and the *Weakness*). However, since the first two interaction terms were not statistically significant, a model including only the main effects and the interaction between the number of rest days and the *Weakness* has been considered; this model was also preferable according to the BIC value.

The final model therefore includes three piecewise-constant time-varying covariates recorded at the beginning of each time interval - namely, *Weakness*, the number of rest days after the previous injury, and the age of the player (expressed in years) - two time-independent covariates, i.e. the transformed BMI (BMI_{transf}) and the player's playing position (with three categories: guard, forward and center), and one interaction term between *Weakness* and the number of rest days. As already mentioned, since *Weakness* is an internal time-varying covariate, as its value depends on the past history of the recurrent event process and evolves over time, its inclusion into the A–G model yields a dynamic event-count model that, conditional on the frailty, can be interpreted as a modulated Poisson process. Then, the fitted model, aimed at estimating the conditional intensity of relapse at playing time t, measured in minutes of on-court playing time, given the covariates, the frailty, and the history of the process, is specified as follows:

$$\begin{aligned}\lambda_i(t|Z_i, \mathcal{H}_i(t)) = Z_i\lambda_0(t)\exp\{&\beta_1 W(t) + \beta_2 \text{Rest\_days}(t) + \beta_3 \text{Age}(t) + \\ &+ \beta_4 \text{BMI}_{\text{transf}} + \beta_5 \text{Position} + \beta_6 \text{Rest\_days}(t) \times W(t)\}.\end{aligned}$$

For the following analyses, a significance level of $\alpha = 0.05$ has been used. Results are shown in Table 3.

As expected, the *Weakness* is positively associated with the intensity of new injuries ($\exp\{\beta_1\} > 1$): an increase in the number of consecutive injuries occurring in quick succession leads to a higher rate of occurrence of new injuries. Furthermore, the number of rest days affects the intensity in interaction with the *Weakness*, suggesting that, for a given value of *Weakness*, a longer recovery period reduces the effect of *Weakness* on the intensity. For what concerns players' characteristics, age is positively associated with the relapse occurrence rate, indicating that injuries occur more frequently among older players over time. Moreover, in line with Macis

Table 3 Results from the Andersen-Gill model with frailty

	βestimate [95%CI]	Hazard ratio [95%CI]	p-value
Weakness	0.1399 [0.1150; 0.1647]	1.1501 [1.1219; 1.1791]	< 0.001
Rest days	$-0.0003[-0.0007; 0.0001]$	0.9997 [0.9993; 1.0001]	0.099
Age	0.0361 [0.0234; 0.0488]	1.0368 [1.0237; 1.0500]	< 0.001
BMI$_{\mathbf{transf}}$	$0.0381[-0.0283; 0.1045]$	1.0388 [0.9721; 1.1101]	0.260
Playing position			
Guard	ref	ref	ref
Forward	$0.0670[-0.0649; 0.1990]$	1.0693 [0.9371; 1.2202]	0.320
Center	0.1587 [0.0238; 0.2937]	1.1720 [1.0241; 1.3413]	0.021
Rest days×Weakness	$-0.0005[-0.0008; -0.0002]$	0.9995 [0.9992; 0.9998]	< 0.001

CI: Confidence Interval

(2025b), centers exhibited a higher risk of recurrent injury than guards, whereas no comparable difference was observed for forwards. Then, differently from the findings reported in Macis (2025b), BMI is not statistically significant.

Focusing on the frailty component, it is assumed to follow a Gamma distribution, with an estimated variance of 0.48 (95% confidence interval: [0.43; 0.55]). The presence of residual unobserved heterogeneity was evaluated using the Commenges–Andersen score test, applied to the A–G model without frailty (Commenges and Andersen 1995). The test assesses the null hypothesis of homogeneity across players, corresponding to the absence of a residual frailty component after accounting for the observed covariates, including *Weakness*. Rejection of this null hypothesis provides evidence of unexplained player-specific heterogeneity and supports the inclusion of a frailty term in the recurrent-event model. The Commenges-Andersen heterogeneity test shows that the random effect is highly significant ($p < 0.001$), confirming that its inclusion in the model is appropriate.

Figure 4 shows an example illustrating the effect of *Weakness* and the number of rest days on the intensity of new injuries. Specifically, the intensity was estimated at a fixed time point, $t = 240$ minutes, for two anonymized players, denoted here as player A and player B, taking into account their values of age, BMI_{transf}, position, and estimated individual frailty. Estimates were obtained as a function of *Weakness*, ranging from 0 to 13, and rest days, considering two different recovery times, namely 7 and 30 days.

This intensity is a measure of the instantaneous rate of relapse at time $t = 240$ minutes for the two players, after a sequence of injuries determining a given value of *Weakness* and after either 7 rest days, represented by the solid lines, or 30 rest days, represented by the dashed lines, since the previous injury. The red and blue lines show the estimated intensity for player A and player B, respectively, for different values of *Weakness*. In both cases, the intensity increases as *Weakness* increases and, for a fixed value of *Weakness*, decreases as the number of rest days increases.

Moreover, player A shows a higher intensity than player B, due to differences in individual frailty, equal to 2.95 for player A and 0.22 for player B, and in the observed covariates, namely age, position, and BMI_{transf}. Regarding the covariates, the effects of BMI and playing position appear similar across players, with guards and forwards exhibiting comparable effects. The main difference concerns age, which is higher for

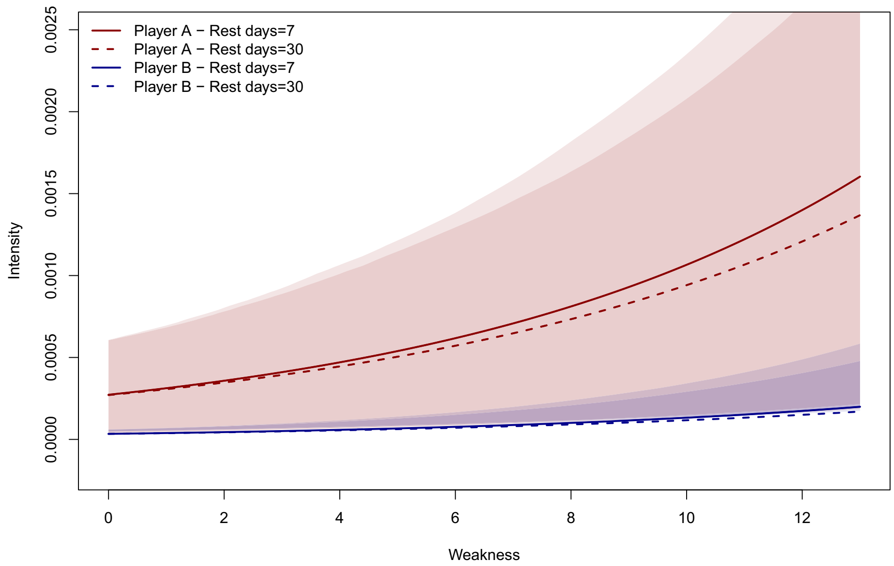


Fig. 4 Estimated intensity at $t = 240$ minutes as a function of *Weakness*, according to length of rest, for two anonymized players. The curves are shown with corresponding 95% bootstrap confidence bands. Player A: Age = 22.75, BMI_{transf} = 0.08, Position = Guard, Frailty = 2.95; Player B: Age = 34.41, BMI_{transf} = 0.07, Position = Forward, Frailty = 0.22

player B. However, as shown in Fig. 4, despite his higher age, the lower frailty leads to a lower intensity for player B.

The 95% bootstrap confidence bands slightly overlap; however, the aim was not to test for differences between the curves, but rather to show the effect of *Weakness* and length of rest on the intensity.

Finally, a sensitivity analysis was carried out to test the robustness of the chosen q value in the definition of the *Weakness*. Different values of q have been considered ($q = 48, q = 144, q = 192$), and no substantial differences in the results were observed: all previously identified relationships were still valid.

5 Conclusions

In this paper, we investigated recurrent injuries among NBA players over the 2010–2011 to 2019–2020 seasons within a recurrent-event modelling framework. The proposed specification combines three key components: the Andersen–Gill counting-process formulation for recurrent events, a nonlinear autoregressive internal covariate, *Weakness*, designed to represent dynamic carryover effects after previous injuries, and a frailty term accounting for residual heterogeneity among players.

The dataset used for the present analysis was obtained by matching and harmonizing injury records, play-by-play data, and player-level characteristics. This step required careful data cleaning and alignment because the three sources differed in observational unit and temporal resolution. To improve transparency and reproducibility, we provided a pseudo-code summarizing the key steps used to construct the recurrent-event data structure required for model estimation. This procedure may provide a useful practical reference for researchers interested in applying recurrent-event methods to basketball injury data or related sports contexts.

The final aim was to evaluate the risk of recurrent injuries by considering both players-level characteristics (BMI, playing position, age), and injury-related features, such as the number of rest days after each injury, and the number of injuries occurring in rapid succession. Within this framework, one of its main contributions is the creation of *Weakness*, an internal time-varying covariate designed to model a temporary carryover effect after previous injuries. Unlike standard carryover formulation (Çiğşar and Lawless 2012a; Nirmalkanna and Çiğşar 2025), *Weakness* provides a nonlinear autoregressive representation of injury-related vulnerability, since its current value depends both on its previous recorded value and on the playing time elapsed since the most recent injury. Through its self-exciting and decaying components, *Weakness* increases when injuries occur in close succession and decreases when the time elapsed since the previous injury becomes longer. In practical terms, it can be interpreted as a measure of temporary vulnerability induced by clusters of injuries occurring within a short period, under the assumption that injuries close in time may be related to the initial event.

Our proposal is conceptually related to the idea underlying Hawkes processes (Hawkes 1971b, a), which have also been used in sports injury modelling (Worrall et al. 2025) to model the injury rate. In a Hawkes process, the conditional intensity is specified directly as a self-exciting function of the full event history, so that previous events increase the subsequent event rate through a predefined excitation kernel. In our approach, by contrast, the A–G intensity is not formulated as a Hawkes intensity. Instead, the effect of previous injuries is summarized through the internal time-varying covariate *Weakness*, which enters the model as part of the observed covariate history. Interestingly, our specification has many similarities with Hawkes processes. Indeed, the trajectory of $W(t)$, when mapped onto the intensity scale, i.e., after exponentiation and weighting by its regression coefficient, resembles the shape of the intensity of a Hawkes process with a parametric exponential triggering kernel (as that shown in Figure 8 of Worrall et al. 2025), displaying similar excitation and decay behaviour. Thus, the proposed specification captures a form of history-dependent self-excitation within a standard recurrent-event regression framework, rather than requiring a separate point-process model.

It is also important to distinguish *Weakness* from the frailty term. *Weakness* is an internal time-varying covariate that represents a transient state of vulnerability induced by injuries occurring in close succession, and its value changes over time according to the player's recent injury history. Frailty, by contrast, is a subject-specific latent random effect that captures persistent differences in injury propensity across players not explained by the observed covariates. Thus, the two components are complementary: *Weakness* models dynamic within-player variation in risk after recent injuries, whereas frailty accounts for residual between-player heterogeneity in susceptibility to recurrent injuries (Macis 2025b).

The proposed frailty-extended dynamic A–G model provided interesting evidence on the factors associated with recurrent-injury intensity. Unlike the results reported in

Macis (2025b), where BMI was associated with the risk of sustaining an injury, in the present study BMI was not significantly associated with the intensity of injury recurrence. By contrast, the role of playing position was broadly consistent with previous evidence: centers showed a higher recurrent-injury intensity than guards, whereas forwards did not differ substantially from guards. Age was also associated with injuries recurrence, with older players showing a higher instantaneous rate of recurrent injury.

The results also support the role of the proposed *Weakness* covariate. Injuries occurring in close succession were significantly associated with increased recurrent-injury intensity, indicating that recent injury history contributes to the short-term dynamics of recurrence. The number of rest days further modified this association through its interaction with *Weakness*: for a given level of *Weakness*, a longer rest period corresponded to a lower instantaneous rate of subsequent injury. This finding is consistent with the interpretation of *Weakness* as a transient state of vulnerability whose effect may be attenuated by a longer rest period.

The present study is subject to some limitations. First, some player characteristics, such as BMI and playing position, were treated as time-fixed covariates because repeated measurements were not available in the player-level dataset $\texttt{Inf}_{\texttt{Pl}}$. Although these characteristics may vary over a player's career, the available data provided only one value per player; therefore, these variables were included in the analysis as constant over the observation period. Second, the proposed definition of *Weakness* does not incorporate information on injury severity. In the present analysis, the dynamic carryover effect was modelled as a function of the previous value of *Weakness* and of the playing time elapsed since the most recent injury. However, the risk of subsequent injuries may also depend on the severity of the preceding injury, which could affect both the duration and the magnitude of the vulnerability state. Further investigation could involve the development of a new variable capturing this feature. Third, the parameters governing the decay of *Weakness* were not estimated directly within the inferential model. In this study, they were selected through model comparison, whereas future research may consider more flexible formulations in which the decay parameters are estimated jointly with the other model parameters. Such extensions could also allow for player-specific decay rates, thereby accounting for heterogeneity across players in the duration of the vulnerability induced by previous injuries.

Future research may extend the present framework in several directions. From a methodological perspective, alternative approaches for recurrent-event analysis, including machine learning-based methods, could be considered to explore nonlinear effects, complex interactions, and prediction-oriented extensions.

From an applied perspective, further work may assess how dynamic measures of post-injury vulnerability can support injury prevention and return-to-play decisions. The management of injured players necessarily involves competitive and economic considerations; however, an early return to play may increase the risk of recurrent injury, particularly when residual vulnerability remains high. Therefore, future applications of this framework may help inform decisions on player reintegration by balancing the immediate needs of the team with the foreseeable risk of subsequent injury.

Acknowledgements The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions. They also wish to thank the S-training project, coordinated by Christophe Ley and Andreas Groll, for their useful and valuable suggestions. 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). The authors further acknowledge BigDataBall, a data provider leveraging computer vision technologies to enhance and extend sports datasets with unique metrics, for providing the play-by-play data used in this study.

Author contributions AM, MM, MS, PZ: conceptualization, methodology, formal analysis and investigation. AM, MS: software. AM: data curation. AM: writing (original draft and revised version). MM, MS, PZ: review and editing. MM and PZ: funding acquisition. PZ: resources. MM, MS and PZ: supervision.

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.

Data availability The final dataset has been obtained after merging three big datasets. Two datasets have been retrieved by public repositories (https://www.kaggle.com/datasets/ghopkins/nba-injuries-2010-2018, https://www.kaggle.com/datasets/wyattowalsh/basketball), while the third one is the exclusive property of BigDataBall and we do not have the rights to disclose it.

Declarations

Conflict of interest The authors have no conflict of interest to disclose.

References

Aalen OO, Borgan Ø, Gjessing HK (2008) Survival and event history analysis: a process point of view. Springer

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

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

Albert J, Bennett J, Cochran JJ (2005) Anthology of statistics in sports, vol 16. SIAM

Albert J, Glickman ME, Swartz TB et al (2017) Handbook of Statistical Methods and Analyses in Sports. CRC Press

Andersen PK, Gill RD (1982) Cox’s regression model for counting processes: a large sample study. The Annals of Statistics pp 1100–1120. https://doi.org/10.1214/aos/1176345976

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

Bahr R, Holme I (2003) Risk factors for sports injuries—a methodological approach. Br J Sports Med 37(5):384–392. https://doi.org/10.1136/bjsm.37.5.384

Balan TA, Putter H (2020) A tutorial on frailty models. Stat Methods Med Res 29(11):3424–3454. https://doi.org/10.1177/0962280220921889
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):1485–1491. https://doi.org/10.1177/0363546505275490
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):598–604. https://doi.org/10.1136/bjsm.2007.044677
Çiğşar C, Lawless JF (2012a) Assessing transient carryover effects in recurrent event processes, with application to chronic health conditions. The Annals of Applied Statistics pp 1641–1663. https://doi.org/10.1214/12-AOAS560
Çiğşar C, Lawless JF (2012b) Assessing transient carryover effects in recurrent event processes, with application to chronic health conditions. The Annals of Applied Statistics pp 1641–1663. https://doi.org/10.1214/12-AOAS560
Carlesso ML, Cappozzo A, Manisera M et al (2025) Scoring probability maps in the basketball court with Indicator Kriging estimation. Comput Statistics 40(4):1731–1751. https://doi.org/10.1007/s00180-024-01564-4
Cefis M, Carpita M (2025) On the cta-pls test for hierarchical models: an application to the football player's performance. Computational Statistics 40(4). https://doi.org/10.1007/s00180-024-01566-2
Collett D (2015) Modelling survival data in medical research. CRC Press
Commenges D, Andersen PK (1995) Score test of homogeneity for survival data. Lifetime Data Anal 1(2):145–156. https://doi.org/10.1007/BF00985764
Cook RJ, Lawless JF (2007) The statistical analysis of recurrent events. Springer
Cox DR (1972) Regression Models and Life-Tables. J Roy Stat Soc: Ser B (Methodol) 34(2):187–220 (http://www.jstor.org/stable/2985181)
Csurilla G, Fertő I (2022) How long does a medal win last? Survival analysis of the duration of Olympic success. Appl Econ 54(43):5006–5020. https://doi.org/10.1080/00036846.2022.2039370
D'Urso P, De Giovanni L, Swartz T (2023) Big data and data science in sport. Ann Oper Res 325(1):1–7. https://doi.org/10.1007/s10479-023-05344-z
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. The Journal of Bone & Joint Surgery 99(11):897–904. 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):160–172. https://doi.org/10.1177/1527002507308309
Deroche T, Stephan Y, Woodman T et al (2012) Psychological mediators of the sport injury—perceived risk relationship. Risk Analysis An International Journal 32(1):113–121. https://doi.org/10.1111/j.1539-6924.2011.01646.x
Difernand A, Mallet A, De Larochelambert Q et al (2025) A survival analysis of dropout among French swimmers. Frontiers in Sports and Active Living 7:1509306. https://doi.org/10.3389/fspor.2025.1509306
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):33–37. https://doi.org/10.1136/bjsports-2018-100020
Fynn KD, Sonnenschein M (2012) An analysis of the career length of professional basketball players. The Macalester Review 2(2):3
Gabbett TJ (2016) The training—injury prevention paradox: should athletes be training smarter and harder? Br J Sports Med 50(5):273–280. https://doi.org/10.1136/bjsports-2015-095788
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):3–4. 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):115–128. https://doi.org/10.1093/imaman/dpq005
Hägglund M, Waldén M, Magnusson H et al (2013) Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. Br J Sports Med 47(12):738–742. https://doi.org/10.1136/bjsports-2013-092771

Hawkes AG (1971) Point spectra of some mutually exciting point processes. J Roy Stat Soc: Ser B (Methodol) 33(3):438–443. https://doi.org/10.1111/j.2517-6161.1971.tb01530.x

Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58(1):83–90. https://doi.org/10.1093/biomet/58.1.83

Hopkins WG, Marshall SW, Quarrie KL et al (2007) Risk factors and risk statistics for sports injuries. Clin J Sport Med 17(3):208–210. https://doi.org/10.1097/JSM.0b013e3180592a68

Hougaard H (2000) Analysis of Multivariate Survival Data. Springer

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):54–61. https://doi.org/10.3171/2018.11.PEDS18626

Jack RA, Sochacki KR, Hirase T, et al (2019) Performance and return to sport after hip arthroscopic surgery in major league baseball players. Orthopaedic Journal of Sports Medicine 7(2). https://doi.org/10.1177/2325967119825835

Kemler E, Vriend I, Paulis W et al (2015) Is overweight a risk factor for sports injuries in children, adolescents, and young adults? Scandinavian journal of medicine & science in sports 25(2):259–264. https://doi.org/10.1111/sms.12180

Kontos AP (2004) Perceived risk, risk taking, estimation of ability and injury among adolescent sport participants. J Pediatr Psychol 29(6):447–455. https://doi.org/10.1093/jpepsy/jsh048

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):394–401. 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). https://doi.org/10.1371/journal.pone.0196062

Lee DJ, Zumeta-Olaskoaga L (2022) Can we really predict injuries in team sports? Boletın de Estadıstica e Investigación Operativa BEIO p. p 149

Leriou I, Ntzoufras I (2025) Survival modeling of goal arrival times in english premier league. Comput Statistics 40(4):2109–2133. https://doi.org/10.1007/s00180-024-01589-9

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 (2025) Basketball players performance measurement with algorithmic survival data analysis. AStA Advances in Statiistical Analysis 109:529–555. https://doi.org/10.1007/s10182-025-00533-6

Macis A (2025) The role of the frailty in the evaluation of injury risk factors for national basketball association players. Comput Statistics 40:1985–2003. https://doi.org/10.1007/s00180-024-01556-4

Macis A, Manisera M, Zuccolotto P (2023) A survival analysis to discover which skills determine a higher scoring in basketball. Statistica Applicata-Italian Journal of Applied Statistics 35(2): https://doi.org/10.26398/IJAS.0035-009

Mahmood A, Ullah S, Finch C (2014) Application of survival models in sports injury prevention research: a systematic review. Br J Sports Med 48(7):630–630. https://doi.org/10.1136/bjsports-2014-093494.190

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):2226–2232. https://doi.org/10.1177/0363546517704834

McLeod TCV, Lewis JH, Whelihan K et al (2017) Rest and return to activity after sport-related concussion: a systematic review of the literature. J Athl Train 52(3):262–287. https://doi.org/10.4085/1052-6050-51.6.06

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

Moore IS, Bitchell CL, Vicary D et al (2023) Concussion increases within-player injury risk in male professional rugby union. Br J Sports Med 57(7):395–400. https://doi.org/10.1136/bjsports-2021-105238

Moulds K, Abbott S, Pion J et al (2020) Sink or swim? A survival analysis of sport dropout in Australian youth swimmers. Scandinavian Journal of Medicine & Science in Sports 30(11):2222–2233. 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):1856–1863. 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. Journal of Quantitative Analysis in Sports 9(2):165–177. https://doi.org/10.1515/jqas-2012-0004

Nirmalkanna K, Çiğşar C (2025) Analysis of recurrent event processes with dynamic models for event counts. Stat Biosci 17(2):321–365. https://doi.org/10.1007/s12561-024-09432-x

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

Peng KK, Hu XJ, Swartz TB (2025) On the time of corner kicks in soccer: an analysis of event history data. Computational Statistics 40(4). https://doi.org/10.1007/s00180-024-01567-1

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):935–940. 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):347–354. https://doi.org/10.1080/24748668.2016.11868891

Schneider KJ, Iverson GL, Emery CA et al (2013) The effects of rest and treatment following sport-related concussion: a systematic review of the literature. Br J Sports Med 47(5):304–307. https://doi.org/10.1136/bjsports-2013-092190

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,

Smith KL, Weir PL (2022) An Examination of Relative Age and Athlete Dropout in Female Developmental Soccer. Sports 10(5):79. 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. Journal of Hip Preservation Surgery 6(3):234–240. https://doi.org/10.1093/jhps/hnz030

Soligard T, Schwellnus M, Alonso JM et al (2016) How much is too much?(part 1) international olympic committee consensus statement on load in sport and risk of injury. Br J Sports Med 50(17):1030–1041. https://doi.org/10.1136/bjsports-2016-096581

Therneau TM, Lumley T (2015) Package survival. version 32-10 R Top Doc 128(10):28–33

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

Tozetto AB, Carvalho HM, Rosa RS et al (2019) Coach turnover in top professional Brazilian football championship: A multilevel survival analysis. Front Psychol 10:1246. https://doi.org/10.3389/fpsyg.2019.01246

Travert M, Maïano C, Griffet J (2017) Understanding injuries in sports: self-reported injury and perceived risk of injury among adolescents. Eur Rev Appl Psychol 67(6):291–298. https://doi.org/10.1016/j.erap.2017.10.002

Ullah S, Gabbett TJ, Finch CF (2014) Statistical modelling for recurrent events: an application to sports injuries. Br J Sports Med 48(17):1287–1293. 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):439–454. 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):293–298. 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 Management Review 21(4):333–346. https://doi.org/10.1016/j.smr.2017.08.002

Wilkerson GB, Gupta A, Colston MA (2018) Mitigating sports injury risks using internet of things and analytics approaches. Risk Anal 38(7):1348–1360. https://doi.org/10.1111/risa.12984

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

Worrall J, Wu PPY, Toohey LA et al (2025) Sports Injury is Self-exciting: Modelling AFL Injuries Using Hawkes Processes. SN Computer Science 6(5):398. DOI:https://doi.org/10.1007/s42979-025-03939-w

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

Zuccolotto P, Manisera M, Kenett R (2017) Guest Editorial ‘ Statistics in sports’. Electronic Journal of Applied Statistical Analysis 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 Advances in Statistical Analysis pp 1–26. https://doi.org/10.1007/s10182-021-00428-2

Zumeta-Olaskoaga L, Bender A, Lee DJ (2025) Flexible modelling of time-varying exposures and recurrent events to analyse training load effects in team sports injuries. J R Stat Soc: Ser C: Appl Stat 74(2):391–405. https://doi.org/10.1093/jrsssc/qlae059

Authors and Affiliations

Ambra Macis[1,2] · Marica Manisera[1,2] · Marco Sandri[2] · Paola Zuccolotto[1,2]

✉ Ambra Macis
ambra.macis@unibs.it

Marica Manisera
marica.manisera@unibs.it

Marco Sandri
sandri.marco@gmail.com

Paola Zuccolotto
paola.zuccolotto@unibs.it

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

2 Big and Open Data Innovation Laboratory, University of Brescia, Contrada S. Chiara, 50, Brescia 25122, Italy