eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Statistical Modelling for Recurrent Events in Sports Injury Research with Applications to Football Injury Data

---

Lore Zumeta Olaskoaga

PhD Thesis supervised by:

Dae-Jin Lee

December 2023

Statistical Modelling for Recurrent Events in Sports Injury Research with Applications to Football Injury Data

---

Lore Zumeta Olaskoaga

PhD Thesis supervised by:

Dae-Jin Lee

December 2023

This research was supported by the Spanish Ministry of Science and Innovation (MICINN) through the Severo Ochoa SEV-2017-0718 PRE2018-084007 funding and the BCAM Severo Ochoa accreditation CEX2021-001142-S/MICIN/AEI/10.13039/501100011033; by the Basque Government through the BERC 2018-2021 and BERC 2022-2025 programs, and the PRE_2021_2_0029 funding; and by the AEI/FEDER, UE through the "S3M1P4R" PID2020-115882RB-I00 project.

Izarriturriri
Amona Juanitari

Acknowledgements
Esker Onak

I want to express my gratitude to all the people who have accompanied and supported me on this journey, making this work possible. My deepest thanks to...

... my supervisor, Dr. Dae-Jin Lee, who has provided me with steady support and generously shared his knowledge and research insights with me. Mila esker, Dae-Jin, por toda la confianza depositada en mí, por tu apoyo y por todo lo que me has enseñado durante estos años.

... Jon Larruskain-i eztabaida aberasgarri guztiengatik eta hain eskuzabal agertzeagatik beti. Eskerrik asko baita Eder Bikandi, Xabi Monasterio eta Josean Lekue-ri ere. Asko ikasi dut eta Lezaman izan naizen guztietan energiaz beterik itzuli naiz.

... Max Weigert, Alex Bauer, Andreas Bender und Helmut Küchenhoff. Vielen Dank for the great opportunity to collaborate with you and for your hospitality during my research stay at the Ludwig-Maximilans-Universität in Munich.

... Saioa Iru-ri bere laguntasunagatik eta, bizpahiru ideia emanda, lan honi azal zoragarri bat egiteagatik.

... a Joaquín y a Martí por las conversaciones compartidas y por todo el apoyo recibido. Amaia, Irantzu eta Inma-ri bide honetan gertu egoteagatik eta emandako laguntzagatik.

... Diana, Jone, eta BCAM-eko bulegokide eta bulego ondoko bizilagunei, lan-giro atsegin eta laguntzaileagatik. BCAM-eko staff-ekoei jardun honetako paper-festa errazagoa egiteagatik.

... kuadrilako eta gertuko lagunei beti hor egoteagatik, emandako animoengatik eta erakutsitako interesagatik.

... Ioar-i bere babesagatik, latza izan den azken urte luze honetan ondoan egoteagatik, eta eduki nezakeen pixukide onena izateagatik. Oraintxe leherrak eta ahalak eginda nago, baina "lixto".

... bi amonei. Egin eta eman didazuen –diguzuen– guztiaren laurdena ematerik banu! Maite, aita eta ama... bihotza-bete esker zuei.

Summary

Sports injuries stand as undesirable side effects of athletic participation. They carry severe implications for athletes' health, affecting not only their overall well-being but also their professional careers by interrupting training sessions and participation in competitions. This, in turn, affects the team's overall performance and impacts the financial aspects of the sports clubs. Consequently, considerable efforts are directed towards understanding the mechanisms of sports injuries, as some may be mitigated through injury prevention programs.

Thus, partly driven by the ever-increasing amount of data now being collected, research on sports injuries has attracted significant attention across various fields, including statistics. There is an increasing tendency towards using data-based analysis methods. Appropriate statistical models can assist medical staff in monitoring athletes' health status and prescribing tailored training and prevention programs.

Yet, successfully modelling sports-related injuries remains a real challenge. Sports injuries result from the dynamic interaction of multiple risk factors. That is, their occurrence is influenced by a combination of physiological, biomechanical, psychological, environmental, and individual factors. Importantly, an athlete's risk of injury is not a fixed characteristic; it continuously changes based on the interplay of multiple risk factors. Besides, athletes can sustain multiple injuries, with subsequent ones often being affected by previous ones. Therefore, a proper statistical model should encompass the complex time-varying and recurrent nature of injuries. Recurrent Events Analysis offers a compelling approach for examining such relationships over time between time-varying exposures and outcomes.

In this regard, the focus of this dissertation is on developing statistical models for recurrent events in sports injury data, which can be directly applied in practical settings through software implementations. We develop and assess various time-to-event modelling approaches to address a range of research questions arising in real-world contexts with sports injury data. These methodological advancements are driven by interdisci-

plinary research, conducted in close collaboration with the Medical Services of Athletic Club, and are motivated by real-world applications. These applications are based on distinct football injury data sets, namely, the functional screening tests data, the external training load data, and the web-scraped football injury data. These data sets have specific characteristics and give rise to specific research questions:

- **Functional screening tests data** consist of data from professional female football players who underwent functional screening tests during a regular football season. These tests encompass a series of screening assessments that evaluate the strength, power, joint stability, movement patterns and asymmetries. The data also include non-contact lower limb time-loss injuries sustained by the players.

- **External training load data** gather information on non-contact time-loss injuries for 36 professional male football players over two consecutive seasons, along with regularly collected external training load variables.

  External training load refers to any external stimulus applied to the player that is measured independently of their internal characteristics. Such external loads elicit physiological and psychological responses in each individual, following interaction with, and variation in several other biological and environmental factors. A variety of measures, including training or competition time, distance covered, speed, power output, sprints, and more, can be used to quantify the external load.

  In this case, these variables are measured using Global Positioning System (GPS) devices integrated into vests worn by players during every match and training session.

- The web-scraped football injury data, named **"transfermarkt" data** because the data are scraped from the webpage of the same name, provide information on injuries and match sheets from the five major European male football leagues –the English Premier League, the German Bundesliga, the Spanish LaLiga, the Italian Serie A and the French Ligue 1– spanning the seasons from 2005-2006 to 2021-2022. These data serve to illustrate all the fundamental concepts and measures implemented in the statistical open-source software **R**.

The related practical questions are: (**a**) Which functional screening tests most affect injury risk in football? (**b**) How do past training exposures influence injury risk in football? and (**c**) How can we make the developed models in (**a**) and (**b**) accessible and useful? Each question is addressed in a specific chapter and is related to a particular objective.

From a statistical point of view, our focus centers on time-to-event analysis or survival analysis methods. Particularly, we investigate and assess shared frailty Cox models and flexible recurrent time-to-event models to account for the dynamic, time-varying, and recurrent nature of sports injuries.

In summary, the dissertation has three main objectives:

(**a**) To assess different variable selection methods together with shared frailty Cox models for identifying biomechanical risk factors for subsequent sports injuries. **Chapter 3** addresses this objective.

(**b**) To develop and assess a flexible recurrent time-to-event approach for modelling the effects of training load on subsequent sports injuries. **Chapter 4** addresses this objective.

(**c**) To develop software that implements the statistical methods for analyzing sports injury data proposed in this dissertation. **Chapter 5** addresses this objective.

In alignment with these objectives, **Chapter 2** introduces the common methodology that supports the subsequent development and evaluation of models. It provides a comprehensive overview of the epidemiological measures that describe and quantify injury occurrences and general regression models suitable for sports injury data. The chapter establishes a robust framework that serves as the basis for the forthcoming chapters.

Chapter 3 tackles the problem of identifying the most relevant functional screening tests and estimating player-specific injury risk over time. We employ shared frailty Cox models for this purpose.

However, in cases where the number of covariates and parameters to be estimated is large, shared frailty Cox models may encounter convergence problems. This is particularly problematic for small sample data. Therefore, it becomes essential to reduce the number of parameters to be estimated and efficiently select a subset of relevant variables associated with the risk of injury. To do so, we compare several variable selection methods for time-to-event data analysis, including regularized Cox methods such as Best Subset Selection (BeSS), Least Absolute Shrinkage and Selection Operator (Lasso), Elastic Net, Ridge regression, and Group Lasso, as well as Boosting in Cox regression.

We assess the performance of the shared frailty Cox models, which include different sets of previously selected variables, with respect to prediction accuracy. This assessment is conducted through a simulation study designed to evaluate the applicability and robustness of the discussed statistical approach, under three hypothetical controlled situa-

tions, reflecting sports injury data contexts. Throughout these scenarios, special attention is given to the impact of varying sample sizes.

Chapter 4 focuses on flexible modelling of time-varying exposures and recurrent events to analyze the effects of training load on team sports injuries, particularly in football. Players are repeatedly exposed to high competition demands that, in turn, increase the strain on their bodies and exposure to the risk of injury. This continuous exposure, manifested through training loads applied over varying time periods and with varying magnitudes, represents the cumulative stress from multiple training sessions and matches over time.

To address this, we propose the use of the Piece-wise exponential Additive Mixed Model (PAMM) with weighted cumulative exposure-type (WCE) cumulative effects. This model considers the intensity and duration of past exposures to sports participation, as well as dependencies induced by subsequent injuries. We demonstrate that the PAMM framework allows for the estimation of highly flexible models.

Recognizing that past exposures may not have an everlasting effect, we develop a method to identify a relevant time window during which past exposures have an impact. Indeed, as time passes, the effect of exposures recorded long ago may disappear. Additionally, PAMMs require data to be transformed into an appropriate format. We implement code support for an already available **R** function that performs this transformation. Our code implementation for data transformation provides support for the specific case of data including recurrent events with time-dependent covariates.

Exhaustive simulation studies assess the ability of the proposed models to simultaneously estimate both, flexible WCE-type effects and heterogeneity resulting from recurrent events. These simulation studies also evaluate the performance of the developed methods in selecting the maximum length of the time window in which past exposures are cumulatively associated with the hazard.

Chapter 5 covers the aspect of software development. Throughout this dissertation, we employ the statistical open-source software **R** as a tool for implementing and evaluating statistical models, processing and tidying the data, and presenting results in both visual and tabular formats. All these computational developments are publicly available, either as documented code repositories or packages, promoting the principles of open science and enabling complete reproducibility.

This chapter begins with an overview of existing **R** packages in the field of sports medicine. We acknowledge that there is a shortage of **R** packages designed for this field and emphasize the need for dedicated software. Next, we introduce our self-developed

R package, named **injurytools**. It is a fully documented package that facilitates the data analysis workflow and automates common tasks performed in practice with sports injury data. The chapter employs a hands-on approach to illustrate its usage, similar to the guidance provided on the package's companion website. We not only detail the technical aspects of the package but also underscore its real-world applicability. Therefore, through a comprehensive exploration of data structure principles and functionalities, we provide practitioners with powerful tools for effective sports injury data analysis.

Finally, **Chapter 6** concludes the dissertation with the main conclusions of the work and considerations for further research. The proposed statistical modelling approaches represent a fair trade-off between flexibility, accuracy, adequacy, computational efficiency, and interpretability. In the end, we enumerate all the scientific contributions derived from the work presented in this dissertation.

The statistical advancements developed in this dissertation contribute to ongoing efforts in sports injury prevention, providing insights, methodologies, and accessible software implementations for sports medicine practitioners.

Laburpena

Kirol-lesioak jarduera fisikoa egitearen albo-kaltetzat har daitezke. Eragin kaltebera dute kirolarien osasunean, ongizate orokorrari ez ezik, beraien ibilbide profesionalari ere eragiten baitie. Batetik, entrenamendu saioak eteten dituzte eta bestetik, lehiaketetan parte hartzea galarazi. Horrek, era berean, eragina du taldearen jardun orokorrean, eta baita kirol kluben finantza kontuetan ere. Ondorioz, kirol-lesioen mekanismoak ulertzea garrantzi handikoa da eta ahalegin handiak egiten ari dira bide horretan. Lesio batzuk prebentzio-programen bidez arindu edota saihets daitezke.

Horri lotuta, eta hein batean datuak biltzeko dagoen egungo erraztasunari lotuta, kirol-lesioei buruzko ikerketak arreta handia erakarri du hainbat arlotan, estatistikan barne. Gero eta joera handiagoa dago datuetan oinarritutako metodoak erabiltzeko. Izan ere, eredu estatistiko egokiak lagungarriak dira medikuentzat, kirolarien osasun-egoera monitorizatzerako eta entrenamendu- eta prebentzio-programa egokituak preskribatzerako garaian, besteak beste.

Hala ere, benetako erronka da kirol-lesioak egokiro modelizatzea. Kirol-lesioak arrisku-faktore askoren elkarrekintza dinamikoaren ondorio dira. Hau da, faktore fisiologiko, biomekaniko, psikologiko, ingurumeneko eta indibidualen arteko konbinazioek eragiten dute lesio gertaeran. Horrez gain, kirolariaren lesio-arriskua ez da ezaugarri finko bat; arrisku hori etengabe aldatzen den zerbait da, faktore askoren elkarreraginean oinarrituta. Gainera, kirolariek lesio ugari izan ditzakete, aurretiazko lesioek hurrengo lesioetan eragina izaten baitute maiz. Horrenbestez, eredu estatistiko egoki batek kontuan hartu behar ditu lesioen izaera konplexua, aldakorra eta errepikakorra. Ildo horretatik, Getaera Errekurrenteen Analisiak (*Recurrent Events Analysis*, ingelesez) metodologia baliagarria eskaintzen digu, denboran aldakorrak diren aldagai azaltzaile zein erantzun aldagai errekurrenteen arteko, denboran zeharreko, harremanak aztertzeko.

Beraz, tesi honen ardatza, kirol-lesioen datuetarako, gertaera errekurrenteetako eredu estatistikoen garapena eta eredu hauek testuinguru praktikoetan zuzenean aplikagarri egiten dituen software-aren garapena, dira.

Kirol-lesioen inguruko testuinguru eta ikerketa-galdera desberdinei erantzun asmoz, biziraupen analisiko metodologian oinarritzen diren hainbat eredu garatu eta ebaluatzen ditugu. Aurrerapen metodologiko guztiak, diziplina-arteko ikerketek bultzatu dituzte, Athletic Club-eko Zerbitzu Medikuarekin lankidetza estuan. Futboleko lesioetan oinarritutako datu-multzo desberdinek sustatu dituzte aurrerapenok, hain zuzen: *screening* test funtzionalen datuek, entrenamenduko kanpo-kargen datuek, eta web-etik karrakatu edo *eskrapeatutako* futboleko lesioen datuek. Datu hauek berariazko ezaugarriak dituzte eta berariazko ikerketa-galderak planteatuarazten dizkigute:

- **Screening test funtzionalen datuek**, futbol denboraldi batean, emakumezko futbolari profesionalek osatutako screening test funtzionalen inguruko informazioa dute. Test hauek hainbat probek osatzen dute, hala nola, indarra, artikulazioen egonkortasuna, mugimendu-patroiak eta asimetriak ebaluatzea helburu dituzten probek. Datu hauen artean daude, halaber, jokalariek sufritutako lesioak. Hau da, kontaktuzkoak ez diren, beheko gorputz-adarrei eragin dien eta denbora galera suposatu duten kirol-lesioak.

- **Entrenamenduko kanpo-kargen datuek** kontakturik gabeko eta denbora-galera suposatu duten lesioei buruzko informazioa dute. Bi denboralditan zehar, entrenamendu eta partida bakoitzean, 36 gizonezko futboalari profesionalengandik jasotako kanpo-kargek osatzen dituzte datuok.

  Entrenamenduko kanpo-karga jokalariari aplikatzen zaion kanpoko edozein estimulu da, haren barne-ezaugarriak edozein direla ere neurtzen dena. Kanpo-karga hauek erantzun fisiologiko eta psikologikoak sortzen dituzte indibiduo bakoitzean, beste faktore biologiko eta ingurumeneko batzuen arteko elkarrekintzaren eta aldaketaren ondoren. Hori kuantifikatzeko, entrenamendu- edo norgehiagoka-denbora, egindako distantziak, abiadura, potentzia eta sprintak bezalako neurriak erabil daitezke, besteak beste.

  Kasu honetan, jokalariek partida- eta entrenamendu-saio bakoitzero janzten dituzten txalekoetan integratutako Kokapen Sistema Globaleko (GPS) gailuak erabiliz neurtzen dira aldagai horiek.

- Web-etik eskrapeatutako futbol lesioen datuek, **"transfermarkt" data** deitu dioguna –datuak izen bereko web-orritik eskrapeatzen baitira–, 2005-2006 eta 2021-2022 denboraldien bitarteko, Europako bost futbol-liga nagusietako –Ingalaterrako Premier League, Alemaniako Bundesliga, Espainiako LaLiga, Italiako A Seriea eta Frantziako Ligue 1–, gizonezkoen futbol-taldeetako lesioei eta partida-fitxei bu-

ruzko informazioa dute. Datu hauek balio dute kode irekiko software estatistikoan pausatutako oinarrizko kontzeptu eta neurri guztiak erakustarazteko.

Datu hauei lotuta, honako galderak egin ditzake batek: (**a**) Zein screening proba funtzionalek eragiten diote gehien futboleko lesio-arriskuari? (**b**) Nola eragiten diete iraganeko entrenamendu esposizioek futboleko lesioak izateko arriskuari? eta (**c**) Nola egin ditzakegu erabilgarri eta eskuragarri (**a**) eta (**b**)-n garatutako ereduak? Galdera bakoitza kapitulu jakin batean lantzen da, eta helburu jakin batekin lotuta dago.

Metodologia estatistikoari dagokionean, biziraupen analisiko metodoetan zentratzen gara. Zehazki, *shared frailty Cox* ereduak eta *flexible recurrent time-to-event* ereduak ikertzen eta ebaluatzen ditugu, eredu hauek aintzat hartzen baitituzte kirol-lesioen izaera dinamiko, errekurrente eta denbora aldakorra.

Hori horrela, tesi honek hiru helburu nagusi ditu:

(**a**) Aldagaiak hautatzeko metodo desberdinak ebaluatzea, *shared frailty Cox* ereduekin batera, (ondorengo) kirol-lesioen arrisku-faktore biomekanikoak identifikatzeko. **3. Kapituluan** lantzen dugu helburu hau.

(**b**) *Recurrent time-to-event* arloko eredu flexible bat garatzea eta ebaluatzea, entrenamenduko kanpo-kargek (ondorengo) kirol-lesioetan duten eragina modelizatzeko. **4. Kapituluan** lantzen dugu helburu hau.

(**c**) Kirol-lesioen datuen azterketarako, tesi honetan proposatutako eredu estatistikoak kodean inplementatzea eta softwarea garatzea. **5. Kapituluan** lantzen dugu helburu hau.

Helburu horiekin bat, **2. Kapituluan** oinarrizko metodologia aurkezten dugu, ondorengo ereduen garapenari eta ebaluazioari bide egingo dion metodologia. Lesioen gertaera deskribatzen eta kuantifikatzen duten neurri epidemiologikoak aurkezten ditugu, eta kirol-lesioen datuetarako egokiak diren erregresio orokorreko ereduak azaltzen ditugu. Hots, hurrengo kapituluetarako oinarri izango den marko sendo bat ezartzen du kapitulu honek.

3. Kapituluan, screening test funtzional garrantzitsuenak identifikatzearen eta denboran zeharreko, jokalari bakoitzari dagokion, lesio arriskua estimatzearen inguruko problema aztertzen dugu. *Shared frailty Cox* ereduak erabiltzen ditugu helburu honi heltzeko.

Hala ere, estimatu beharreko aldagai eta parametro kopurua handia den kasuetan, *shared frailty Cox* ereduek konbergentzia-arazoak izan ditzakete. Bereziki arazo-

tsua da hori datuen lagin tamaina txikia denean. Beraz, funtsezkoa da estimatu beharreko parametro-kopurua murriztea eta lesio-arriskuari lotutako aldagai garrantzitsuen azpimultzo bat eraginkortasunez hautatzea. Horretarako, aldagaiak hautatzeko hainbat metodo erabili ditugu, *time-to-event* arloan kokatzen direnak. Hala nola: Cox-en metodo erregularizatuak, *Best Subset Selection* (BeSS), *Least Absolute Shrinkage and Selection Operator* (Lasso), *Elastic Net*, *Ridge regression*, eta *Group Lasso*; eta *Boosting*-a Cox-en erregresioan.

Aurretik hautatutako aldagai-multzoak dituzten *shared frailty Cox* ereduen portaera ebaluatzen dugu aurresateko gaitasunarekiko. Ebaluazio hori egiteko –eztabaidatutako eredu estatistikoen aplikagarritasuna eta sendotasuna ebaluatzeko– simulazio azterketa bat egiten dugu eta kirol-lesioen datuen testuinguruak islatzen dituzten hiru agertoki hipotetiko planteatzen ditugu horretarako. Agertoki horietan, lagin-tamaina desberdinen eraginari aparteko arreta jartzen diogu.

4. Kapituluan, denboran aldakorrak diren esposizioen eta gertaera errekurrenteen arteko modelizazio flexiblea aztertzen dugu, entrenamenduko kanpo-kargek taldeko kirol-lesioetan duten eragina aztertzeko, futboleko lesioetan bereziki. Kirolariek lehia-eskakizun handiei aurre egin behar izaten diete etengabe, eta honek areagotu egiten ditu beraien gorputzeko muskuluen gainkargak eta lesioak sufritzeko arriskua. Entrenamendu-karga hau, intentsitate eta magnitude desberdinetan aplika dakioke jokalariari. Karga hau, finean, entrenamendu-saio eta partida ugariren ondorioz, jokalari bati ezartzen zaion estres akumulatua da.

Hortaz, *Piece-wise exponential Additive Mixed Model* (PAMM) deituriko ereduen markoa erabiltzea proposatzen dugu, *weighted cumulative exposure* motako (WCE) efektu akumulatuak dituena. Eredu honek kontuan hartzen ditu aurretiazko kirol esposizioen intentsitate eta iraupenak, eta ondoz-ondoko lesioek indibiduo bereko datuetan sortzen dituzten dependentziak. Kapituluan zehar argi ikusten da, eredu oso flexibleak zenbatesteko aukera ematen duen markoa dela, PAMM ereduen markoa.

Bestalde, iraganeko esposizioek ez dute zertan betiereko efektu bat eduki. Hau da, denbora pasa ahala, aspaldi erregistratutako esposizioen eragina desagertu egin daiteke. Hori kontuan izanik, metodo bat garatzen dugu iraganeko esposizioen eragin-eremua edo eragin-leihoa identifikatzeko. Aitzitik, PAMM eredu bat doitzeko, formatu egoki batera transformatu behar dira datuak. Bada, transformazio hori egiten duen **R**-ko funtzio bati euskarri berri bat gehitzen diogu. Gure kodeak, kasu konkretu baterako datuen –gertaera errekurrenteak eta denboran aldakorrak diren aldagaiak dituzten datuen– transformazioa ahalbidetzen du.

Simulazio-azterketa sakonen bitartez, proposatutako ereduen gaitasuna ebaluatzen

dugu. Zehatz esanda, WCE motako efektu flexibleak eta gertaera errekurrenteek sortutako heterogeneotasuna aldi berean estimatzeko gaitasuna. Era berean, iraganeko esposizioeen eragin-eremuaren gehienezko luzeera identifikatzeko metodoen gaitasuna ebaluatzen dugu.

5. Kapitulua software-aren garapena aurkezteari eskaintzen diogu. Tesi osoan zehar, kode irekiko **R** software estatistikoa erabiltzen dugu eredu estatistikoak inplementatu eta ebaluatzeko, datuak txukundu eta prozesatzeko, eta emaitzak formatu grafiko zein tabularrean aurkezteko. Garapen konputazional hauek guztiak eskuragarri daude, bai errepositorio publiko gisa, baita dokumentatutako kode-pakete gisa ere, zientzia irekiaren printzipioak sustatuz eta erreproduzibilitatea ahalbidetuz.

Kapitulu honetan, kirol medikuntzaren alorrerako dauden **R**-ko paketeen ikuspegi orokor bat aurkezten dugu. Azpimarra egiten dugu, bateko, alor honetarako diseinatutako **R**-ko paketeen eskasian, eta besteko, software espezifiko baten beharrean. Ondoren, eta aipatutako behar horri erantzun asmoz, guk garatutako **R**-ko paketea aurkezten dugu, **injurytools** izena jarri dioguna. Paketeak berariazko funtzio eta tresna egokiak eskaintzen ditu, kirol-lesioen datu analisia errazten du eta ohiko zenbait zeregin automatizatzen ditu. Kapituluko atal honek tutorial traza hartzen du. Paketearen alderdi teknikoa zehazteaz gain, bere erabilera eta aplikagarritasuna erakusten dugu, paketeak berak dakarren webgune osagarrian egiten denaren antzera. Beraz, datuen egituraren printzipioak eta funtzionalitateak sakon aztertuz, tresna baliotsuak eskaintzen dizkiegu erabiltzaile eta profesionalei, kirol-lesioei buruzko datuak eraginkortasunez azter ditzaten.

Azkenik, **6. Kapituluak** tesiari itxiera ematen dio. Bertan, lanaren ondorio nagusiak eta etorkizuneko ikerketetarako gogoetak plazaratzen ditugu. Proposaturiko modelizazio estatistikoek oreka egokia erakusten dute flexibilitatearen, zehaztasunaren, egokitasunaren, eraginkortasun konputazionalaren eta interpretagarritasunaren artean. Azkenburuan, tesi honetatik eratorritako ekarpen zientifiko guztiak zerrendatzen ditugu.

Tesi honetan garatutako aurrerapen estatistikoak, kirol-lesioen prebentzioan egiten ari diren ahaleginen beste ekarpen bat dira. Aurrerapen hauek kirol-medikuntzako profesionalei bideratutako metodologia egokia, software eskuragarria eta ideia berriak eskaintzen dituzte.

Contents

Part I.
Introduction and Background

Chapter 1

Introduction

1.1 Outline

This dissertation focuses on the statistical analysis of sports injury data, addressing situations where the risk of injury occurrence varies over time, numerous exposure variables are observed –some of which change over time– and repeated measurements are involved. It contributes to the field of time-to-event or survival analysis. The main goal is to develop practical and accessible statistical modelling approaches for sports injury data, which can be directly applied in sports medicine through software implementations for practitioners. These methodological advancements are driven by interdisciplinary research, conducted in close collaboration with the Medical Services of Athletic Club, and are motivated by real-world applications in sports injury prevention science. Specifically, these developments are based on football injury data, stemming from diverse contexts and raising various research questions. All contributions include flexible implementations of the methodological approaches in the statistical open-source software **R** (R Core Team, 2023), available either as documented code repositories or packages, promoting the principles of open science and enabling complete reproducibility.

This introductory chapter provides a brief overview of the research questions addressed and motivates their statistical relevance. It concludes by outlining the three proposed objectives and describing the organization of the remainder of the dissertation.

1.2 Motivation and scope

Injury prevention has been declared a priority in 21st-century disease prevention (Dorney et al., 2020). It is crucial not only for individual well-being but also for public health.

When it comes to sports, injuries are undesirable side effects of sports participation (Van Mechelen et al., 1992). **Sports injuries** have serious consequences for athletes' health, affecting not only their well-being but also their professional careers by disrupting training and competition participation. Consequently, these injuries affect overall team performance (Hägglund et al., 2013) and have significant implications for club finances (Ekstrand, 2013; Lutter et al., 2022). In this sense, considerable efforts are directed towards understanding the underlying mechanisms of injuries, as some can be mitigated through injury prevention programs.

Thus, partly driven by the ever-increasing amount of data now being collected, research on sports injuries has attracted significant attention across various fields, including Statistics. There is a growing trend towards using data-based analysis methods. Indeed, the application of suitable statistical models can assist medical staff in monitoring athletes' health status and in prescribing tailored training and prevention programs, offering meaningful insights into injury risk. However, several publications have cautioned that statistical errors are common in this research area and deserve special attention (Nevill et al., 2007; Nielsen et al., 2018; Kim and Lee, 2019). During the "*2019 Methods Matter Meeting*", international researchers with expertise in research methods in sports science highlighted pertinent statistical and epidemiological issues for consideration in the injury modelling process (Nielsen et al., 2020). Recent publications by Sainani et al. (2021) call for increased statistical collaboration in sports medicine and sports injury prevention research, while Casals and Finch (2017) emphasizes the importance of sports biostatisticians as essential members of sports science and medicine teams for injury prevention.

Ruddy et al. (2019) provide an overview of existing strategies to monitor and model the occurrence and duration of sports injuries, encompassing both classical statistical and machine learning models. They also highlight several limitations of these models due to the unique characteristics of sports injury data. Recent perspectives in sports medicine and injury prevention suggest that sports injuries result from the dynamic interaction of multiple risk factors, making them "*complex*" phenomena (Bolling et al., 2018). Therefore, an appropriate statistical model should encompass the complex, time-varying and recurrent nature of injuries: a player's injury susceptibility may change over time, and a player can sustain multiple injuries, with subsequent injuries often influenced by previous ones (Hägglund et al., 2006; De Visser et al., 2012).

This dissertation focuses on the dynamic, time-varying nature of injuries. It advocates for the use of time-to-event methods and develops and assesses various approaches to address a range of research questions that arise in real-world contexts with sports injury data.

Use of some technical terms

In this dissertation, certain terms related to sports medicine are used interchangeably. The following remarks are provided to clarify their usage.

***Remark* 1.1** Unless otherwise specified, the term "*injury*" refers to "*sports injuries*". In general, the term "*players*" is used, but it may be interchangeably referred to as "*athletes*" in some contexts.

***Remark* 1.2** The term "*sports injury*" encompasses a wide range of definitions. When the terms "*injury*" or "*sports injuries*" are used, they generally refer to non-contact injuries that result in time loss. Injuries resulting from contact with another player or object are not considered. Additionally, injuries that do not result in a loss of time, meaning those allowing the player to fully participate in future match play or training sessions despite requiring medical attention, are also excluded.

***Remark* 1.3** The term "*subsequent injury*" is used to describe injuries occurring in the same player. This term is preferred over "*recurrent injury*", which is typically defined as an injury of the "same" type. The use of "*subsequent injury*" adheres to a broader definition. It does not only refer to the same index injury but also: to an injury in the same location, though not precisely the same type; an injury in another location but the same type; or an entirely different injury (Fuller et al., 2007).

Readers are also referred to the "*Consensus Statement on Injury Definitions and Data Collection Procedures in Studies of Football (soccer) Injuries*" by Fuller et al. (2006) for a comprehensive guide on standardized injury definitions and data collection procedures in football-related injury research.

1.2.1 Motivating data sets

In the following section, a brief overview is provided of the various data sets that have motivated the statistical developments in this dissertation. These include: (a) functional screening tests data, (b) external training load data, and (c) "transfermarkt" data. Each data set is intricately linked to a specific objective, which will be outlined in the next Section 1.3. Further details about these data will be presented in their respective chapters.

(a) Functional screening tests data

These observational data were recorded during the 2017-2018 football season by the medical staff of Athletic Club, specifically focusing on the 22-player professional female foot-

ball team. Throughout that season, players conducted regular screening tests consisting of various medical evaluations. These evaluations included functional movement tests to assess biomechanical factors and muscle imbalances, anthropometric measurements, range of motion assessments, dynamometries, core strength evaluations, drop jump mechanics, and countermovement jumps (see Figure 1.1). The purpose of these tests was to evaluate movement patterns and identify any asymmetries, thereby providing insights into mechanical restrictions and potential injury risks.

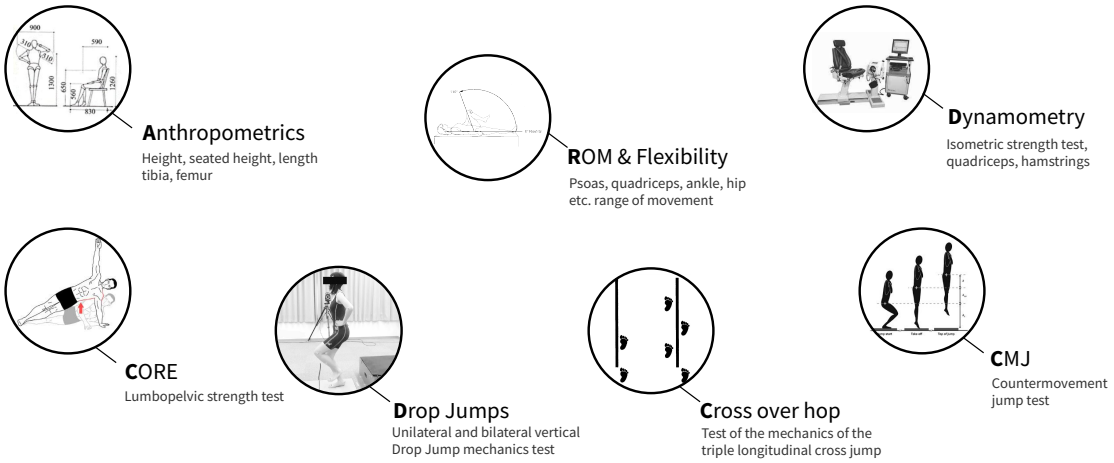


Figure 1.1: An illustration of the types of functional screening tests conducted to evaluate movement patterns and asymmetries.

This series of tests resulted in a high-dimensional data setting, with the tests being repeated at three different moments during the season. The main goal is to identify those tests that most significantly influence the risk of non-contact lower limb injuries and to estimate how the risk changes over time among different players.

(b) External training load data

These observational data, collected during the 2017-2018 and 2018-2019 football seasons, include information on non-contact time-loss injuries and external training load variables for the 36 professional male football players at Athletic Club.

The external training load is defined as any external stimulus applied to a player, measured independently of the player's internal characteristics. Such external loads trigger physiological and psychological responses in each individual, following interaction with, and variation in, several other biological and environmental factors (Soligard et al., 2016). A variety of measures, such as training or competition time, distance covered, speed, power output, sprints, and more, can be used to quantify the external load.

In this case, the external training load variables were regularly tracked during each match and training session using Global Positioning System (GPS) devices (see Figure 1.2). Consequently, the status of both the explanatory and outcome variables vary over time.

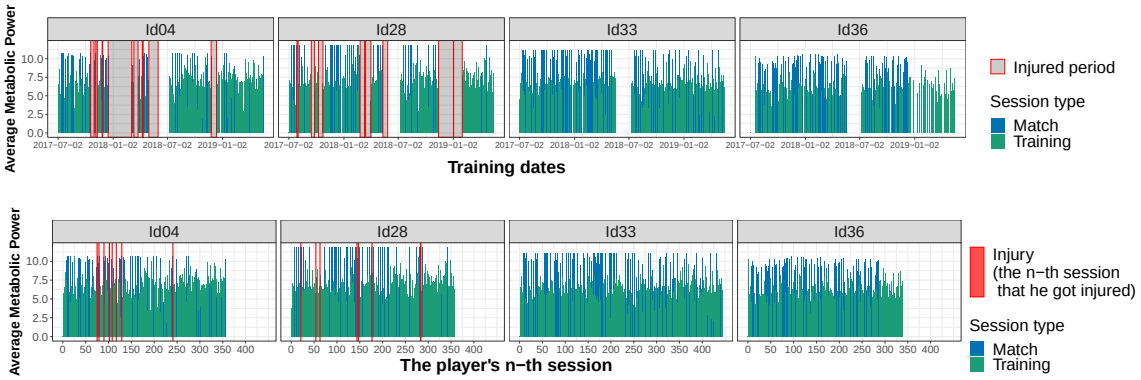


Figure 1.2: Longitudinal profiles of the *Average Metabolic Power* training load variable for four players. Each bar in the graph corresponds to a match (coloured in blue) or a training session (in green). Periods during which players were injured are marked with grey areas, delineated by red vertical lines. Top: the scale represents calendar dates. Bottom: the scale represents the session number of each player.

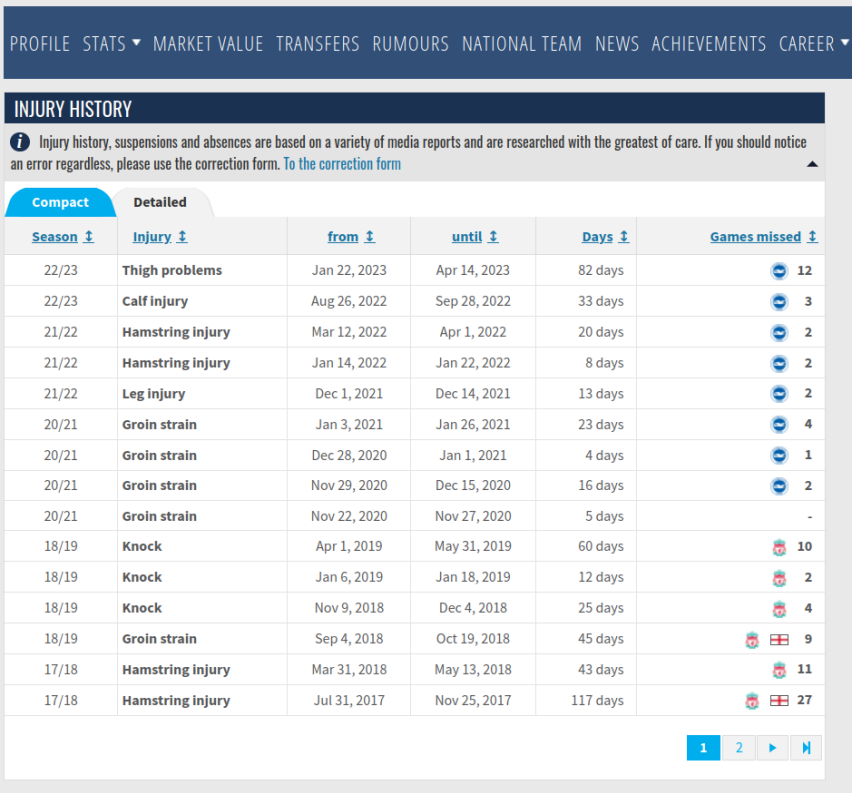
These GPS devices, integrated into vests worn by players, are primarily used for monitoring and collecting data on their physical performance and movements. They provide real-time data on a variety of performance metrics, including distance covered, speed, accelerations, decelerations, change of direction, metabolic power, jumps and impacts, and distance covered at different intensities, among others.

The aim is to assess the potential relationship between external training load variables and non-contact time-loss injuries; namely, how the training load, from multiple training sessions and matches, cumulatively affects, over a period of time, a player's risk of a (subsequent) football injury.

(c) "transfermarkt" data

These data are observational, web-scraped data from the five major European male football leagues –the English Premier League, the German Bundesliga, the Spanish LaLiga, the Italian Serie A and the French Ligue 1– spanning the seasons from 2005-2006 to 2021-2022. The data comprise injury and match sheet data obtained through web scraping from the popular German website https://www.transfermarkt.com/. It is one of the largest sports websites and provides football information such as scores, results, statistics, trans-

fer news, and fixtures, as well as football players' injury histories (see Figure 1.3).



PROFILE STATS MARKET VALUE TRANSFERS RUMOURS NATIONAL TEAM NEWS ACHIEVEMENTS CAREER

INJURY HISTORY

Injury history, suspensions and absences are based on a variety of media reports and are researched with the greatest of care. If you should notice an error regardless, please use the correction form. To the correction form

Compact | Detailed

Season	Injury	from	until	Days	Games missed
22/23	Thigh problems	Jan 22, 2023	Apr 14, 2023	82 days	12
22/23	Calf injury	Aug 26, 2022	Sep 28, 2022	33 days	3
21/22	Hamstring injury	Mar 12, 2022	Apr 1, 2022	20 days	2
21/22	Hamstring injury	Jan 14, 2022	Jan 22, 2022	8 days	2
21/22	Leg injury	Dec 1, 2021	Dec 14, 2021	13 days	2
20/21	Groin strain	Jan 3, 2021	Jan 26, 2021	23 days	4
20/21	Groin strain	Dec 28, 2020	Jan 1, 2021	4 days	1
20/21	Groin strain	Nov 29, 2020	Dec 15, 2020	16 days	2
20/21	Groin strain	Nov 22, 2020	Nov 27, 2020	5 days	-
18/19	Knock	Apr 1, 2019	May 31, 2019	60 days	10
18/19	Knock	Jan 6, 2019	Jan 18, 2019	12 days	2
18/19	Knock	Nov 9, 2018	Dec 4, 2018	25 days	4
18/19	Groin strain	Sep 4, 2018	Oct 19, 2018	45 days	9
17/18	Hamstring injury	Mar 31, 2018	May 13, 2018	43 days	11
17/18	Hamstring injury	Jul 31, 2017	Nov 25, 2017	117 days	27

1 2

Figure 1.3: An example of the web scraped injury data, available on the "transfermarkt" webpage: https://www.transfermarkt.com/adam-lallana/verletzungen/spieler/43530.

As these data are publicly available, they are particularly useful for illustrating all the fundamental concepts and measures implemented in the statistical software **R**.

1.3 Objectives

This dissertation pursues three main objectives:

(a) To assess different variable selection methods together with shared frailty Cox models for identifying biomechanical risk factors for subsequent sports injuries (Chapter 3).

(b) To develop and assess a flexible recurrent time-to-event approach for modelling the effects of training load on subsequent sports injuries (Chapter 4).

(c) To develop software that implements the statistical methods for analyzing sports injury data proposed in this dissertation (Chapter 5).

Each specific objective is developed in three independent chapters in Part II, using a common methodology presented in the following Chapter 2.

More precisely, Chapter 2 introduces the fundamental statistical modelling approaches for sports injury prevention research proposed throughout this dissertation. Part II covers the individual contributions that are the core of this dissertation, comprising: Chapter 3, dedicated to analyzing and evaluating regularized Cox and shared frailty models in the context of sports injury data, characterized by a high number of covariates and a low number of events; Chapter 4, which proposes and evaluates a flexible methodology for investigating the association between (past) training load and (subsequent) injuries; and Chapter 5, which emphasizes the necessity for dedicated software and introduces the self-developed **injurytools R** package. The dissertation concludes with Chapter 6 summarizing the main findings and suggesting directions for future research. In addition, each chapter begins by outlining its primary research contributions. At the end of Chapter 6, all scientific contributions derived from this dissertation are collectively enumerated.

Chapter 2

Statistical modelling approaches for sports injury data

Statistics play a crucial role in sports injury research, offering valuable and essential methods for extracting meaningful information from injury events. This research field frequently employs statistical methodologies to address a variety of important questions, which include the frequency of injuries, their severity, and associations with potential risk factors (e.g., athletes' biomechanics, physiological markers, training loads), among other topics. Yet, successfully modelling sports-related injuries remains a real challenge due to their complex and multifactorial nature (Van Mechelen et al., 1992; Meeuwisse, 1994). In this context, the following statement by Phillips (2000) holds particular relevance:

> *"Sports injuries occur when athletes are exposed to their given sport and they occur under specific conditions, at a known time and place."*

In the following sections, we introduce measures to describe injury occurrence. Then, we focus on modelling injuries and present two general regression models that describe how potentially related variables can explain the event of injury when injuries are viewed as either (a) count data (e.g., injury incidence modelling) or (b) time-to-event data (e.g., injury hazard modelling) for analysis. Finally, we present a methodological framework that links both approaches, allowing for the estimation of highly flexible models.

2.1 Measures of injury occurrence

We adapt key epidemiological measures, such as rates and prevalence –which quantify the frequency and distribution of diseases within a population– for the context of sports injuries.

A rate is a measure that consists of a denominator and a numerator over a period of time. The denominator can represent various time metrics (e.g. the number of minutes trained and played, the number of matches played, or calendar days). A rate reflects the speed at which new injury events occur. On the other hand, prevalence denotes the proportion of a population that is injured at a given point in time. It can be interpreted as the probability that, at time t, a randomly selected player from the population will have the injury.

More precisely, after defining the specific injury under study, we define the *injury incidence rate*, the *injury burden rate* and the *prevalence*.

Definition 2.1. *Injury incidence rate* is the number of new injury cases (I) per unit of player-exposure time, i.e.,

$$I_r = \frac{I}{\Delta T},$$

where ΔT is the total time under risk of the study population.

Definition 2.2. *Injury burden rate* is the number of days lost (n_d) per unit of player-exposure time, i.e.,

$$I_{br} = \frac{n_d}{\Delta T},$$

where ΔT is the total time under risk of the study population.

Definition 2.3. *Prevalence*, or *period prevalence*, is the number of players that have reported the injury, divided by the total player population at risk at any time during the specified period of time (ΔT time window), i.e.,

$$P = \frac{X}{N},$$

where X is the number of injury cases and N is the total number of players in the study at any point in the time window ΔT. X includes players who already had the injury at the start of the time period and those who suffered it during that period.

Assuming the number of incidence cases (I or n_d) follows a Poisson distribution, the computation of confidence intervals for rates can be done using the Poisson or the normal distribution based on the central limit theorem.

Let I_r (I_{br}) be the underlying true incidence rate (burden rate), whose estimator is

$$\widehat{I}_r = \frac{I}{\Delta T}.$$

It can be assumed that I (n_d), the number of incident cases (days lost) throughout the total time under risk, ΔT, follows a Poisson distribution with parameter $I_r \cdot \Delta T$. Hence, the expected value and the variance of I_r are I_r and $I_r/\Delta T$, respectively. This leads to the approximate confidence interval under large sample conditions:

$$\text{CI}(I_r; 1-\alpha) = \hat{I}_r \pm z_{1-\alpha/2} \cdot \sqrt{\hat{I}_r/\Delta T},$$

where $1-\alpha$ is the confidence level and $z_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the standard normal distribution.

With regards to prevalence, P, given a sample of independent observations and assuming that X follows a binomial distribution with parameters n and P, there are various ways to compute a confidence interval for P, which include: using the binomial distribution (exact interval, Clopper and Pearson 1934), using Jeffreys interval (Bayesian approach, Jeffreys 1946; Brown et al. 2001) or using the normal distribution. The latter, also known as Wald interval or asymptotic interval, is based on the central limit theorem and calculated as:

$$\text{CI}(P; 1-\alpha) = \hat{P} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{P}(1-\hat{P})/N},$$

where $1-\alpha$ is the confidence level and $z_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the standard normal distribution.

Based on this, statistical inference can be carried out by comparing estimates from two different populations or at different time points using methods such as the exact binomial test, the test of equal or given proportions, or the Wald test for two incidence rates, among others. We refer the reader to Chapter 5 for examples of these measures in practical applications.

We list some remarks to keep in mind:

***Remark* 2.1** Rates, either the injury incidence rate or injury burden rate (I_r or I_{br}), are not ratios and they are not interpreted as a probability. Their unit is (person-time)$^{-1}$, e.g. per 1000h of player-exposure, per player-season etc.

***Remark* 2.2** Rates, either I_r or I_{br}, can be studied in cohort studies, but not in case-control or cross-sectional studies.

***Remark* 2.3** Injury prevalence can be estimated in cross-sectional studies, but not in cohort or case-control studies.

***Remark* 2.4** Injury prevalence depends on injury duration: the longer the duration, the higher the prevalence.

***Remark* 2.5** Injury incidence rate (likelihood) and injury burden rate (severity) should be reported and assessed in conjunction rather than in isolation (Bahr et al., 2018).

***Remark* 2.6** Before computing any of those values, it is important to clearly define the time scale for each measure. Incidence-based measures, which provide a standardized time window for the population at risk (e.g., injuries per hour) are preferred over measures in which the time at risk varies among individuals (e.g., injuries per athletic exposure, injuries per number of matches, see Stovitz and Shrier (2012)). Using measures with standardized time scales facilitates the comparison of statistics across different cohorts and sports (Waldén et al., 2023).

2.2 Regression models for injury data

In the following, we comprehensively present various regression modelling approaches, analyzing the outcome variable "*injury*" from two perspectives: as count data and as time-to-event data.

2.2.1 Injuries as count data

Let Y_l be a random variable representing the number of injuries (or the number of days lost due to injury) sustained by player l at time period Δt_l, which he or she has been exposed to the risk of injury[1]. Consider $\boldsymbol{X} = (1, X_1, \ldots, X_p)$ as a set of predictors which can include both continuous or categorical variables.

To analyze the possible relationship between this set of predictor variables $\boldsymbol{X}$ and the response variable $\boldsymbol{Y}$, let us first assume that Y_l follows a Poisson distribution with mean $\mu_l = \lambda_l \cdot \Delta t_l$, where λ_l corresponds to either the previously defined I_r or I_{br}. We also assume a linear relationship between the predictors and some function of the expected outcome. Then, we model a Poisson generalised linear model (Poisson GLM), also known as a log-linear model, as,

$$\eta_l = g\left(\mathbb{E}(Y_l|\boldsymbol{X}_l)\right) = \boldsymbol{X}_l'\boldsymbol{\beta} + \log\left(\Delta t_l\right), \qquad l = 1, \ldots, L. \tag{2.1}$$

where $\boldsymbol{X}_l'$ is the l^{th} row-vector of covariates of player l, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of unknown regression coefficients, $g(\cdot)$ is the *link* function, in this case, $g(\mathbb{E}(Y_l|\boldsymbol{X}_l)) = g(\mu_l) = \log(\mu_l)$, and $\log(\Delta t_l)$ is the *offset* term.

[1] By convention, we have opted to use the letter Y. It represents either previously defined I or n_d.

The vector of regression coefficients $\boldsymbol{\beta}$ of model (2.1) are usually estimated by the maximum likelihood method, for which the log-likelihood, $\boldsymbol{l}(\cdot)$, is given by:

$$\begin{aligned}\boldsymbol{l}\left(\boldsymbol{\beta}; l\right) &= \log\left(\prod_{l=1}^{L} f(Y_l)\right) = \sum_{l=1}^{L} \log\left(f(Y_l)\right) = \\ &= \sum_{l=1}^{L} \log\left(\lambda_l^{y_l} \exp(-\lambda_l)/y_l!\right) = \sum_{l=1}^{L} \left(y_l \log \lambda_l - \lambda_l - \log(y_l!)\right) \overset{\log(\lambda_l)=\sum_i x_{l,i}\beta_i + \log(\Delta t_l)}{=} \\ &= \sum_{l=1}^{L} \left(y_l \left(\sum_{i=1}^{p} x_{l,i}\beta_i + \log(\Delta t_l) \right) - \exp\left(\sum_{i=1}^{p} x_{l,i}\beta_i + \log(\Delta t_l) \right) - \log(y_l!) \right).\end{aligned}$$

Then, to find out the vector $\boldsymbol{\beta}$ that maximizes the log-likelihood function, one must solve the following equation, setting the score function U_i equal to zero:

$$U_i = \frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_i} = 0 \quad \Leftrightarrow \quad \boldsymbol{x}_i \left(\boldsymbol{y} - \exp\left(\boldsymbol{x}_i \beta_i\right) \Delta \boldsymbol{t}\right) = 0, \quad \forall i \in \{1, \ldots, p\}. \tag{2.2}$$

Generally, to solve the given equation and obtain the maximum likelihood estimators $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, an iterative weighted least squares algorithm is employed, utilizing a numerical method procedure such as the Newton-Raphson technique.

It is also worth noting that one can equivalently derive the solutions in Eq. (2.2) by taking advantage of the fact that the Poisson distribution belongs to the exponential family.

Besides, to account for the fact that a player might sustain multiple injuries in different time periods –thereby introducing some dependency in the data from the same individual–, we assume that, conditional on a random effect b_l, Y_l follows a Poisson distribution. This leads us to model a Poisson generalised linear mixed model (Poisson GLMM), also known as the log-linear mixed effects model,

$$\eta_l = g\left(\mathbb{E}(Y_l | \boldsymbol{X}_l, b_l)\right) = \boldsymbol{X}_l' \boldsymbol{\beta} + b_l + \log\left(\Delta t_l\right), \qquad l = 1, \ldots L, \tag{2.3}$$

where $b_l \sim N\left(0, \sigma_b^2\right)$ and the remaining terms are defined in the same way as in model (2.1).

Similarly, the maximum likelihood method can be used to estimate the regression coefficients in model (2.3). The estimation is based on the marginal likelihood where the random effects are integrated out. A penalized iteratively re-weighted least squares algorithm can be used to solve the maximization problem in combination with numerical calculation methods such as the Laplace method for integral approximation, the penalized quasi-likelihood, the adaptive Gauss-Hermite quadrature or Monte Carlo methods. We refer to McCullagh and Nelder (1989) and McCulloch et al. (2003) for more details.

More importantly, the regression setting allows us to compute incidence rates (IR) and incidence rate ratios (IRR) given a predictor or a pair of sets of predictors. To do so, it is enough to take the exponent of the estimated $\widehat{\beta}_i$ coefficient. Besides, the random effect b_l in Eq. (2.3) informs us about the inherent susceptibility of a player to get injured.

Example 2.1. For the sake of simplicity, let's assume that X_1 is a binary variable (e.g., indicating having suffered a previous injury). The IR for a player having condition X_1 (i.e., when X_1 is true), and all other variables equal to zero or at their reference values, is:

$$\text{IR} := \mathbb{E}\left(Y_l | (x_1 = 1, x_2 = 0, \ldots, x_p = 0)\right) = \exp\left(\beta_0 + \beta_1 x_1 + \log\left(\Delta t_l\right)\right),$$

and the IRR of a player having condition X_1 compared to not having it, while holding all else equal, is:

$$\text{IRR} := \frac{\mathbb{E}\left(Y_l | (x_1 = 1, x_2, \ldots, x_p)\right)}{\mathbb{E}\left(Y_l | (x_1 = 0, x_2, \ldots, x_p)\right)} = \exp\left(\beta_1\right).$$

Confidence intervals of these measures can be computed on the *link* scale, based on the standard error of $\widehat{\beta}_i$ and on Student's t-distribution. The exponential of the interval endpoints is taken to interpret the results on the response scale.

However, Poisson regression models frequently face specific challenges, such as *overdispersion*, where the variance of the data is greater than the mean, and zero inflation, which refers to an excess of zero counts.

In the context of sports injuries, these issues often arise, as the distribution of the "injury-related" variable –the number of injuries or the number of days lost due to injury– typically exhibits right-skewness, and (fortunately) a significant presence of zero values (Shrier et al., 2009). When overdispersion is present, an alternative to Poisson regression is the Negative Binomial (NB) regression model. Additionally, in cases where there are many zeros in the data, *the zero-inflated Negative Binomial model* is suggested (Lambert, 1992; Yau et al., 2003).

The zero-inflated negative binomial model

The zero-inflated negative binomial (ZINB) distribution is a mixture of distributions expressed as,

$$\text{ZINB} \sim \begin{cases} 0, & \text{with prob. } p \qquad \text{non-susceptible population,} \\ \text{NB}(y; r, t), & \text{with prob. } 1-p \qquad \text{susceptible population,} \end{cases}$$

or equivalently,

$$P(Y=0) = p + (1-p)\cdot \text{NB}(0;r,t) = p + (1-p)t^r, \quad 0<p<1,$$
$$P(Y=y) = (1-p)\cdot \binom{y+r-1}{r-1} t^r (1-t)^y, \quad y=1,2,\ldots \text{ and } 0<p<1.$$

When considering explanatory variables, the parameters of a zero-inflated negative binomial mixed model are modelled as,

$$\begin{aligned} \text{logit}(p_l) &= \xi_l = \boldsymbol{X}_l'\boldsymbol{\gamma} + u_l, \\ \log(\lambda_l) &= \eta_l = \boldsymbol{X}_l'\boldsymbol{\beta} + b_l, \end{aligned} \tag{2.4}$$

where p_l and λ_l parameters[2] are linearly related to the covariates through the *link* functions; u and b are player-related random effects normally distributed as $N(\mathbf{0}, \sigma_u^2)$ and $N(\mathbf{0}, \sigma_b^2)$, respectively; and we assume same covariates for both submodels, although this does not necessarily have to be the case.

Given that we are interested in the overall effects of risk factors, rather than in the specific effects (i.e., $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ corresponding to each submodel), we need to work out the expressions for the coefficients associated with the covariates in model (2.4). For more details on how to derive the overall incidence rates (IR) and incidence rate ratios (IRR) in a ZINB model, please refer to Appendix A and to Preisser et al. (2012).

Example 2.2. For simplicity, let's suppose that X_1 is a binary variable (e.g., having a previous injury). Then, the IR of a player having condition X_1, with all other variables equal to zero or at their reference values, is,

$$\text{IR} := \mathbb{E}\left(Y_l | (x_1 = 1, 0, \ldots, 0)\right) = \frac{\exp\left(\beta_0 + x_1\beta_1\right)}{1 + \exp\left(\gamma_0 + x_1\gamma_1\right)},$$

and the IRR of having the condition X_1 compared to not having it, holding all else equal, is,

$$\text{IRR} := \frac{\mathbb{E}\left(Y_l | (x_1 = 1, x_2, \ldots, x_p)\right)}{\mathbb{E}\left(Y_l | (x_1 = 0, x_2, \ldots, x_p)\right)} = \exp\left(\beta_1\right) \frac{1 + \exp\left(\gamma_0\right)}{1 + \exp\left(\gamma_0 + \gamma_1\right)}.$$

Regarding coefficient estimation, one can employ either the expectation-maximization (EM) algorithm or the Newton–Raphson method to derive the maximum likelihood estimates. The likelihood function of this model, which can be factorized into two terms, is not explicitly denoted here. See Yau et al. (2003) and Min and Agresti (2005) for details.

[2]The parameter p_l refers to the probability that individuals are from the non-susceptible population and the parameter λ_l to the mean of the negative binomial distribution for the susceptible population.

In order to calculate the confidence intervals of the previous measures, the model-based semi-parametric bootstrap can be used.

The use of ZINB models in the field of sports medicine has not yet become widespread, even though ZINB models often provide a much better fit for injury count data compared to the Poisson distribution. One reason for the limited popularity of ZINB models may be the challenges in interpretation. It may prove difficult to understand that the composition of two respective subpopulations is a theoretical and mathematical construct. Some recent publications that apply this model class to study risk factors of football injuries include Rommers et al. (2020) and Monasterio et al. (2023a,b).

Limitations

When it comes to modelling the relationship between exposure variables and injury risk, the models described in this section often fail to account for the changing nature of many risk factors or the time-varying nature of the outcome variable (risk of injury). A player's risk of injury is not a fixed characteristic, players continuously change their susceptibility to injury based on the interplay of multiple risk factors. As such, Time-to-Event analysis offers a compelling alternative approach for examining the relationships over time between time-varying exposures and time-varying outcomes (Nielsen et al., 2016, 2019).

2.2.2 Injuries as time-to-event data

This approach – encompassing those methods employed in Time-to-Event analysis or Survival analysis field– is applicable when players are followed over the course of time, such as in a prospective cohort study or randomized trial. The outcome of interest is broadly defined as the time until the occurrence of an injury (event) and data analysis is regularly performed before or without complete knowledge of all injury event times. For example, a study might be finished with players not experiencing the injury or players may drop out of the study, resulting in incomplete observations, known as censoring.

The important concepts here are: (i) *time origin*, (ii) *time scale* and (iii) *censoring*. The *time origin* refers to the point at which we start observing or following a player in the study, which in general, in this dissertation, is the time of inclusion (baseline) into the study. The *time scale* denotes the variable used to identify the "time at risk", i.e. the time period at which the players are at risk of sustaining an injury. For example, this could be minutes or hours of exposure, calendar days, weeks or sports season. *Censoring* occurs when the information available for some players is incomplete. This may happen because the injury event occurs before a player enters the study, or because the study ends before

the injury event takes place. In some cases, the only known information is that the injury occurred within a specific time frame. There are various types of censoring, which depend on the monitoring approach adopted in the study. In this dissertation, we assume right-censoring: if an injury event is not observed, it is only known that the actual time of the injury event is later than a certain value. Reasons for not observing the injury event may include the end of the study, a player's transfer to another team, or a player quitting sports for reasons unrelated to the injury of interest.

All in all, it is crucial to accurately define these three concepts: time origin, time scale and censoring.

Definition 2.4. Let T_l be the time until player l suffers a (predefined) injury. Then, T is a non-negative random variable that can be characterized by either of the following functions:

(i) The *hazard function*, $\lambda(t)$,

$$\lambda(t) := \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.5}$$

(ii) The *cumulative hazard function*, $\Lambda(t)$,

$$\Lambda(t) := \int_0^t \lambda(u)\,\mathrm{d}u\,. \tag{2.6}$$

(iii) The *survival function*, $S(t)$ and the *cumulative distribution function*, $F(t)$,

$$F(t) := 1 - S(t) := P(T \leq t) = 1 - \exp\left(-\Lambda(t)\right). \tag{2.7}$$

In addition, the following relationships are satisfied, among the previously defined functions:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)} = \frac{f(t)}{S(t)} = \frac{-\,\mathrm{d}S(t)\,/\,\mathrm{d}(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}\left(\log(S(t))\right),$$

and thus,

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right), \quad \text{for all } t \geq 0.$$

One may consider $\lambda(t)\,\mathrm{d}t$ as the instantaneous risk of occurring the injury event in the interval $[t, t + \mathrm{d}t)$, knowing that it has not yet occurred for that time. Moreover, due to the dynamic nature of survival data, a characterization of the distribution by the hazard function is very convenient. In fact, the hazard function does not change when conditioning, it is already conditioned on survival time.

Observable data

In the presence of right-censoring, let $T_1, T_2, \ldots, T_n$ be a sample of (partially observed) times and $C_1, C_2, \ldots, C_n$ random censoring. We assume that C_l is independent of T_l for all $l = 1, 2, \ldots, L$, or at least that the distribution of survival times T provides no information about the distribution of censorship times C and vice versa, i.e. non-informative censoring. Then, the observable data is $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_L, \delta_L)$ where

$$Y_l = \min\{T_l, C_l\}, \qquad \delta_l = \mathbb{1}_{\{T_l \leq C_l\}} = \begin{cases} 1, & T_l \leq C_l, \\ 0, & T_l > C_l. \end{cases}$$

The random variable δ_l is the no-censorship indicator, although it is usually known as the censorship indicator.

***Remark* 2.7** In the presence of censoring the hazard remains "undisturbed". For this reason, it is said that survival analysis is hazard-based. That is: what is the probability of observing the actual event time in the small time interval $[t, t + \mathrm{d}t)$, conditional on the fact that neither event nor censoring has happened before t?

The interval $[t, t + \mathrm{d}t)$ is so small that, assuming T and C to be different, at most one is in $[t, t + \mathrm{d}t)$: if the event occurs in $[t, t + \mathrm{d}t)$, it will be observed (still supposing $Y = \min\{T, C\} \geq t$). Because C and T are independent, the probability that the event occurs in $[t, t + \mathrm{d}t)$, conditional on $Y \geq t$, is the same as in the absence of censoring,

$$\lambda(t) \cdot \mathrm{d}t = P\left(T \in [t, t + \mathrm{d}t) \mid T \geq t\right) = P\left(T \in [t, t + \mathrm{d}t), T \leq C \mid \min\{T, C\} \geq t\right),$$

as a consequence, we may estimate the instantaneous hazard function from censored data.

Following, we show that using product integration results in an estimation of the survival function.

Since $\mathrm{d}\Lambda(u) = \lambda(u)\,\mathrm{d}u = P\left(T \in [u, u + \mathrm{d}u) \mid T \geq u\right)$, we may write,

$$1 - \mathrm{d}\Lambda(u) = P\left(T \geq u + \mathrm{d}u \mid T \geq u\right). \tag{2.8}$$

The survival function should then be an infinite product over conditional probabilities of Eq. (2.8). We call such an infinite product, a product integral and write $\prod$. So,

$$S(t) = \prod_0^t (1 - \mathrm{d}\Lambda(u))) \tag{2.9}$$

$$\approx \prod_{k=1}^{K} (1 - \Delta\Lambda(t_k)) \approx \prod_{k=1}^{K} P(T > t_k \mid T > t_{k-1}), \tag{2.10}$$

where $0 = t_0 < t_1 < t_2 < \ldots < t_{K-1} < t_K = t$ partitions the time interval $[0, t]$ in K (small) intervals and $\Delta\Lambda(t_k) = \Lambda(t_k) - \Lambda(t_{k-1})$. Now, the right-hand side of Eq. (2.7) can simply be seen as a solution of the product integral in Eq. (2.9). The product integral itself shows up with the Kaplan-Meier estimator of the survival function.

***Remark* 2.8** In Eq. (2.9), when $\Lambda(t)$ is absolutely continuous, using that for small $\mathrm{d}u$, $\exp(-\lambda(u)\,\mathrm{d}u) \approx 1 - \lambda(u)\,\mathrm{d}u$, we have that:

$$S(t) = \prod_0^t (1 - \mathrm{d}\Lambda(u)) = \prod_0^t (1 - \lambda(u)\,\mathrm{d}u) = \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp(-\Lambda(t)).$$

The Kaplan-Meier estimator of $S(t)$ is obtained by estimating the $\Delta\Lambda(t)$. The latter can broadly be expressed as,

$$\Delta\hat{\Lambda}(t_k) = \frac{d(t_k)}{n(t_k)}, \tag{2.11}$$

where $d(t_k)$ denotes the number of injury events that have occurred within $(t_{k-1}, t_k]$ and $n(t_k)$ the number of players at risk just prior to t_k. If $0 < t_1 < t_2 < \ldots < t_K \leq t$ is the ordered sequence of the observed injury event times, then, plugging Eq. (2.11) into the product integral,

$$\hat{S}_{\text{KM}}(t) = \prod_{k=1}^{K} \left(1 - \Delta\hat{\Lambda}(t_k)\right) = \prod_{k=1}^{K} \left(1 - \frac{d(t_k)}{n(t_k)}\right). \tag{2.12}$$

Figure 2.1 illustrates the estimated survival probabilities using the Kaplan-Meier method in Eq. (2.12). More specifically, it illustrates the probabilities of remaining free from a first-time football injury[3], based on different time scales: panel (a) minutes of exposure, (b) hours of exposure, and (c) number of training sessions and matches completed. It becomes clear that the result can differ depending on the time scale used.

In Time-to-Event data analysis, there are different parametric, semi-parametric, and non-parametric regression models available to incorporate covariate information into the statistical model. They are generally built by specifying the class of hazard function, $\lambda(t)$. Here, we describe *the Cox proportional hazards model* and *the shared frailty Cox model*, as these are the main models we employ throughout this dissertation.

The Cox proportional hazards model

The Cox proportional hazards model, introduced in the influential publication by Cox (1972), is often referred to as semi-parametric because it is comprised of a parametric part

[3]We use the so-called "external training load data" for this illustrative Figure 2.1.

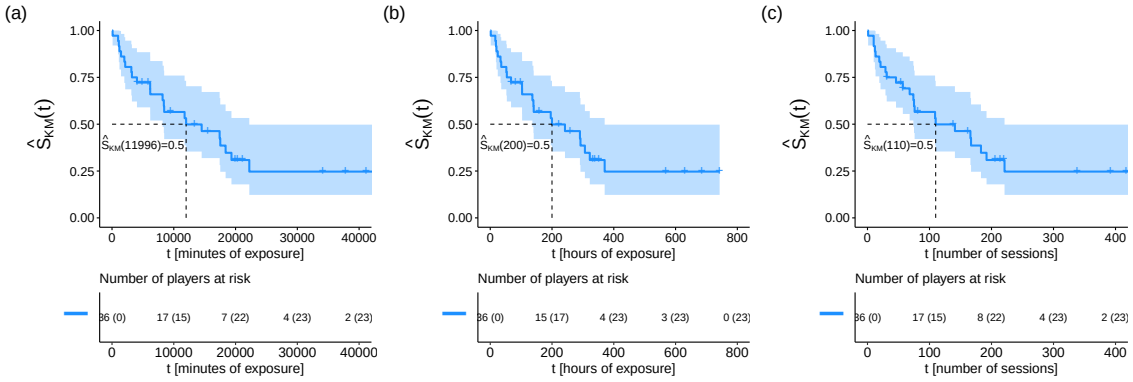


Figure 2.1: Estimation of Kaplan-Meier curves for time-to-first injury outcome, displayed as a function of time scale: panel (a) minutes of exposure, (b) hours of exposure, and (c) number of training sessions and matches completed.

(covariate effects) and a non-parametric part (baseline hazard). Specifically, the hazard function for a player l, with covariate values $\boldsymbol{z}_l(t) = (z_{1l}(t), \ldots, z_{pl}(t))'$, is modelled as,

$$\begin{aligned} \lambda_l(t \mid \boldsymbol{z}(t)) &= \lambda_0(t) \exp\left(\boldsymbol{z}_l'(t)\boldsymbol{\beta}\right) = \\ &= \lambda_0(t) \exp\left(z_{1l}(t)\beta_1 + \ldots + z_{pl}(t)\beta_p\right), \quad l = 1, \ldots, L,\ t \geq 0, \end{aligned} \tag{2.13}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $\lambda_0(t)$ is the baseline hazard. That is, $\lambda_0(t)$ is a hazard function of a player with $\boldsymbol{Z} = \mathbf{0}$, assumed common for all players, which is left unspecified and estimated nonparametrically, e.g., by the Nelson-Aalen estimator or its variants.

To estimate the parameters $\boldsymbol{\beta}$ of Cox proportional hazard model in Eq. (2.13), the partial likelihood is maximized:

$$\mathcal{L}_P(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left(\frac{\exp(\boldsymbol{z}_i'(t)\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{z}_{j_i}'(t)\boldsymbol{\beta}\right)} \right)^{\delta_i},$$

where N is the total number of observations (e.g. $N = L$, the number of total players, when studying time-to-first injury; or $N = \sum_{l=1}^{L} n_l$, when studying time-to-subsequent injuries, where n_l is the total number of injury events that a player has been at risk of), δ_i the censorship indicator and $R(t_i)$ is the risk set, i.e., the set of players who are at risk at time t_i. Note that the baseline hazard cancels out.

The shared frailty model

The shared frailty model (Hougaard, 1995; McGilchrist and Aisbett, 1991) is a frailty model, which allows for dependence between several survival times through a frailty term

that is shared by all the survival times pertaining to a player or, in general, to a cluster. This way, the survival times of a player who sustains multiple injuries –times that are related to each other– have the same level of frailty attached to them. In this context, frailty models are random effects models, expressed as,

$$\begin{aligned}\lambda_l(t \mid \boldsymbol{z}(t), b_l) &= \lambda_0(t) \exp\left(\boldsymbol{z}_l'(t)\boldsymbol{\beta}\right) \alpha_l = \\ &= \lambda_0(t) \exp\left(z_{1l}(t)\beta_1 + \ldots + z_{pl}(t)\beta_p + b_l\right), \quad b \sim N(\mathbf{0}, \boldsymbol{D}),\end{aligned} \tag{2.14}$$

for $l = 1, \ldots, L, t \geq 0$ and where $\boldsymbol{\beta}$ is a vector of regression coefficients, $\lambda_0(t)$ is the baseline hazard and $\alpha_l = \exp(\beta_l)$ is the frailty term associated with player l. In this dissertation, we assume that it follows a log-normal distribution.

Hence, the frailty measures the specific risk level for a cluster or a player's recurrent time-to-event process, and, given α, the survival times are assumed to be independent.

Both Cox proportional hazards and shared frailty regression models have a multiplicative structure. In fact, the effect of a covariate Z_i, $i = 1, \ldots, p$, is described by factors of proportionality, $\exp(\beta_i)$, which is a commonly used effect measure known as the hazard ratio (HR):

$$\text{HR} = \frac{\lambda(t \mid \boldsymbol{z}' = (0, \ldots, 0, z_i, 0, \ldots, 0))}{\lambda(t \mid \boldsymbol{z}' = (0, \ldots, 0))} = \exp(\beta_i).$$

In our context, it relates the hazard at moment t of a player with profile $\boldsymbol{z}$, $\lambda(t \mid \boldsymbol{z})$, with the hazard of a player with profile $\boldsymbol{z} = \mathbf{0}$ at the same time, $\lambda_0(t)$, keeping all other covariates equal. The HR does not depend on t. Thus, the effect of association remains constant over time, hence the name proportional hazards.

Other interesting extensions of the Cox model include joint models (Tsiatis et al., 1995), where one fits a stochastic model both for the covariate processes and for how these influence the hazard rates; and multi-state models (Putter et al., 2007), to analyze the evolution of a process of interest (e.g., time-to different event types (states) such as injury, recovery, second injury etc.). Furthermore, we will shift our attention to another powerful alternative approach.

2.3 Piece-wise Exponential Additive Mixed Model

The Piece-wise Exponential Additive Mixed Model (PAMM, Bender et al. 2018) is a model class that allows for the estimation of very flexible survival models including a variety type of covariate effects: time-varying non-linear covariate effects, cumulative effects of time-varying covariates and random effects. It is the semi-parametric extension of the Piece-wise Exponential Model (PEM, Holford 1980; Whitehead 1980; Friedman

et al. 1982). Both require a specific representation of the data. The data transformation is part of the modelling process, and this makes the survival analysis tasks become Poisson regression tasks.

The representation of the data involves partitioning the follow-up period into a finite number of intervals and assuming that hazards are piece-wise constant in each of these intervals. In the following section, we show that, under certain assumptions, PEMs are essentially Poisson generalized linear models with likelihoods proportional to the (partial) likelihood of a corresponding Cox model.

Quoting Carstensen (2005), the conceptual idea behind this fact is explained as:

> "*if survival studies are viewed in the light of the demographic tradition, the basic observation is not one time to event (or censoring) for each individual, but rather many small pieces of follow up from each individual* [...] *Modelling of rates rather than time to response becomes the focus; the basic response is now a 0/1 outcome in each interval, albeit not independent, but with a likelihood which is a product across intervals*".

Equivalence between the Cox and Poisson model

Let us define the follow-up time as $(0, t_{\max}]$ and the cut points that partition the study follow-up time into J intervals as κ_j, $j = 0, \ldots, J$, where $0 = \kappa_0 < \kappa_1 < \ldots < \kappa_J = t_{\max}$.

In the context of PEM, assuming that the risk is constant in each interval j, i.e., $\lambda_0(t) = \lambda_{0j}$ for all $t \in (\kappa_{j-1}.\kappa_j]$, the Cox model in Eq. (2.13) simplifies to:

$$\lambda_0(t) = \lambda_{0j} \exp\big(\boldsymbol{x}_l'\boldsymbol{\beta}\big), \quad \forall t \in (\kappa_{j-1}, \kappa_j],\ j = 1, \ldots, J, \tag{2.15}$$

which does not depend on t, given the interval j.

Then, if the time-to-event data are structured in a certain way: with event indicators δ_{lj} and *offsets* o_{lj} for all intervals j in which player l is at risk; the likelihood of a Poisson regression model,

$$\mathbb{E}(\delta_{lj}|\boldsymbol{x}_l) = \exp\big(\log(\lambda_j) + \boldsymbol{x}_l'\boldsymbol{\beta} + o_{lj}\big),$$

is proved to be proportional to model in Eq. (2.15) –see section A.2 in the Appendix A where this equivalence is demonstrated.

Consequently, the two models are equivalent with respect to the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$.

$$\lambda_l(t|\boldsymbol{x}_l) = \frac{\mathbb{E}(\delta_{lj}|\boldsymbol{x}_l)}{t_{lj}}, \quad \text{where } t_{lj} = \exp(o_{lj}).$$

Specifically, the Cox proportional hazards model and the Poisson model provide the same estimates if there are no ties in the data (i.e., no subjects experiencing the event at the

same time) and if all unique event (and censoring) times are used as interval cut points to transform the data to the PEM format.

Despite this equivalence between the Cox proportional hazards and the Poisson model, and although the PEM representation is not new, research in the field of survival analysis has mostly been centered around the Cox model and its extensions. Part of the reason for the predominance of the Cox model, in contrast to PEM, has been its computational efficiency, especially for increasing J, and the availability of the Cox routine in standard statistical software. Today, however, there is an **R** package, called **pammtools** (Bender and Scheipl, 2018), that facilitates all the steps involved –such as data transformation, model fitting, and visualization– in the analyses of PEMs and PAMMs.

Data transformation

As already mentioned, time-to-event data requires to be transformed in a particular way. Given intervals $(\kappa_{j-1}, \kappa_j]$ and observed survival times y_l, for each time interval j that player l is under risk, $j = 1, \ldots, J$ and $l = 1, \ldots, L$, the transformation is done by creating (a) an event-specific indicator δ_{lj} and (b) an offset variable $o_{lj} = \log(t_{lj})$. Formally:

(a) δ_{lj} is 1 if both $y_l \in (\kappa_{j-1}, \kappa_j]$ and $y_l = T_l$. Otherwise, δ_{lj} is 0.

(b) $o_{lj} = \log(t_{lj})$ denotes the time player l is under risk in interval j in the logarithmic scale and $t_{lj} = \min(y_l - \kappa_{j-1}, \kappa_j - \kappa_{j-1})$.

Table 2.1: Left: Data in the "standard" time-to-event format for two players, $l \in \{1, 2\}$. Player 1 has been injured at $y_1 = 4$, whereas player 2 has been censored at $y_2 = 15$. Right: Data in piece-wise exponential format with one row per interval in which a player was in the risk set, and intervals are defined by the cut points $0, 5, 10, 15, 20$.

l	y_l	δ_l
1	4	1
2	15	0

l	j	$(\kappa_{j-1}, \kappa_j]$	δ_{lj}	t_{lj}	$o_{lj} = \log(t_{lj})$
1	1	(0, 5]	1	4	$\log(4) = 1.4$
2	1	(0, 5]	0	5	$\log(5) = 1.6$
2	2	(5, 10]	0	5	$\log(5) = 1.6$
2	3	(10, 15]	0	5	$\log(5) = 1.6$

We briefly illustrate this preprocessing step, from a standard time-to-event data format to a piece-wise exponential data format, in Table 2.1. See also the "data-transformation" vignette of the **pammtools R** package.

General formulation of PAMMs

If the follow-up partition is chosen carefully, PEMs and PAMMs enable researchers to benefit from the methodological and algorithmic advancements developed for generalized additive mixed models (GAMMs, Wood 2017).

Figure 2.2 illustrates the basic idea of PEM and PAMM for time-to-event data by applying it to survival times drawn from a Gompertz distribution. To estimate the true underlying Gompertz hazard rate (Figure 2.2, panel (a)), the follow-up is partitioned into a fixed number of intervals (here $J = 4$) with interval cut-points $\kappa_0 = 0 < \ldots < \kappa_J = 20$ (Figure 2.2, panel (b)) and a constant hazard is estimated for each interval (Figure 2.2, panel (c)). Thus, the name piece-wise exponential, because the hazard rate of an exponential distribution is constant over time.

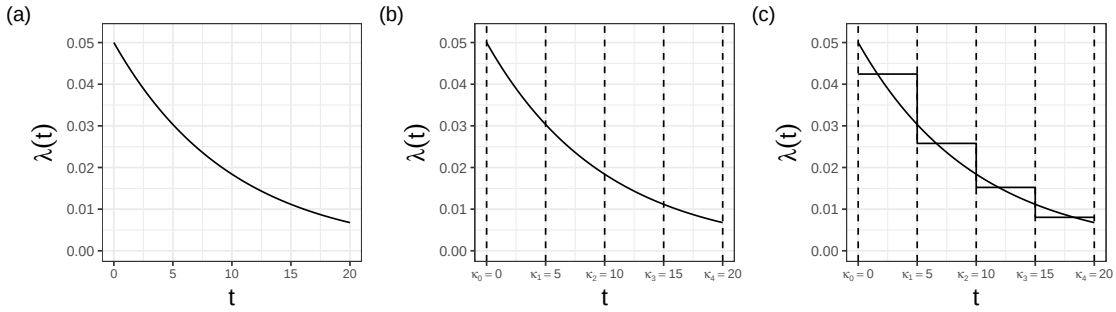


Figure 2.2: *Reproduced and slightly modified from Bender et al. (2018).* (a): Hazard rate of a Gompertz distribution; (b): partitioning of the follow-up into $J = 4$ intervals; (c): estimate of the hazard rate via interval-specific piece-wise constant hazards, obtained by fitting a PEM to the data.

While the approximation in Figure 2.2 may seem crude, with a sufficient number of cut-points, PEM and PAMM estimates closely correspond (or are even equivalent to) Cox regression estimates, as previously shown.

The difference between a PEM and a PAMM lies in their respective approaches for the estimation of the baseline hazard and other smooth, time-varying effects. PAMM, in contrast to PEM, flexibly models the baseline hazard and other time-varying effects using penalized splines. This way, it resolves the arbitrary choice of the cut-points that partition the follow-up time, and thus avoids overfitting and instability issues (Bender et al., 2018). In practice, one can simply use a relatively large number of cut-points and use spline basis functions evaluated e.g. at κ_j, the right end of each interval, and penalize the wiggliness of the estimate via penalized splines, e.g. via P-splines (Eilers and Marx, 1996) based on differences of neighbouring basis coefficients (refer to Bender 2018 for an empirical

discussion on the placing of cut points).

In this regard, a general PAMM of the hazard rate of the i-th injury (event) of the l-th player with covariate vector $\boldsymbol{z}_l(t)$, is given by:

$$\lambda_{i_l}\left(t|\boldsymbol{z}_l(t), b_l\right) = \lambda_0(t)\exp\left(\sum_{k=1}^{p} f_k(z_{lk}(t), t_j) + b_l\right) = \\ = \exp\left(\beta_0 + f_0(t_j) + \sum_{k=1}^{p} f_k(z_{lk}(t), t_j) + b_l\right), \quad \forall t \in (\kappa_{j-1}, \kappa_j], \tag{2.16}$$

where each element represents:

- t_j. In each interval j, t_j is a constant time, e.g. $t_j = \kappa_j$, the right end of the interval, or $t_j = \frac{\kappa_j + \kappa_{j-1}}{2}$, the midpoint of the interval, so that the hazard functions continue to be of the PEM family.
- $f_0(t_j)$, smooth log-baseline hazard rate and $\exp(\beta_0 + f_0(t_j))$, baseline hazard rate. Penalized splines are used to estimate $f_0(t_j)$, so it can flexibly recover the shape of the baseline hazard. For example, baseline hazard may change rapidly at the beginning of the study, and have a less steep growth thereafter. It can be expressed as the linear combination of B-spline basis functions, $B_m(t_j)$: $f_0(t_j) = \sum_{m=1}^{M} \gamma_{0m} B_m(t_j)$.
- $\sum_{k=1}^{p} f_k(z_{lk}(t), t_j)$. Very general types of effects for each covariate $z_{lk}(t)$. It can denote anything from a linear time-constant effect, linear time-varying effect, to smooth time-varying effect or cumulative effects. See Bender et al. (2018) for a comprehensive overview of the possible effect specifications in PAMM.
- b_l. Random intercept term associated to player l, $l = 1, \ldots, L$.

While any method that can optimize Poisson likelihood with offset can be used for PEM estimation, for the PAMM model class –embedded in the context of GAMMs– all the current methods and highly developed software implementations for GAMMs can be transferred.

In Chapter 4 we deepen on the PAMM model class for the context of recurrent time-to-event data and cumulative effects.

Part II.
Main Research Contributions

Chapter 3

Time-to-event modelling and variable selection for recurrent football injury data

Contributing article

Zumeta-Olaskoaga, L., Weigert, M., Larruskain, J., Bikandi, E., Setuain, I., Lekue, J., Küchenhoff, H. & Lee, D.-J. (2023). Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models. *AStA Advances in Statistical Analysis*, 107(1-2), 101-126.

Code repository

`https://github.com/lzumeta/TimeToEvent-InjurySim`

3.1 Context

In this chapter, we focus on lower-limb injuries that frequently occur in women's football –one of the fastest-growing sports worldwide. Lower-limb injuries are of great concern due to their severity and given their high incidence in women football players (Crossley et al., 2020). During a regular season, the medical staff –which includes medical doctors, physiotherapists, and strength and conditioning coaches, among others– conducts regular screening tests for various purposes, which include injury prevention, rehabilitation, and fitness conditioning. The tests consist of a series of medical evaluations, such as functional movement tests that assess biomechanical factors and muscle imbalance. All tests are

designed to monitor the players' health status, identify those predisposed to injury, and consequently, optimize both the player's and the team's performance. In addition, these tests commonly include the quantification of inter-limb asymmetries, which may help in identifying players at a higher risk of lower-limb injuries. Some studies have indicated that bilateral strength asymmetry could be a significant risk factor for musculoskeletal injuries (Croisier et al., 2002, 2003; Knapik et al., 1991; Hewett et al., 2005). However, the scientific evidence supporting the efficacy of these screening tests remains limited (see McCall et al., 2015; Bahr, 2016, for a review).

An important aspect is that a high number of functional tests are made in each evaluation, see Figure 1.1 in Chapter 1. The number of these functional screening tests requires special attention in the modelling process, especially when the interest –from a practical perspective– lies in sparse and interpretable models. Besides, and as already mentioned before, a player's injury susceptibility may change over time, and she may also suffer more than one injury. Hence, to adequately account for all these relevant aspects, we consider variable selection methods and shared frailty Cox models in this chapter.

Related work

Recurrent events models, such as shared frailty Cox models, are widely used in many biomedical studies, but their application in sports injury research has been insufficiently explored (Nielsen et al., 2016). Applications of frailty models in the field of sports injury include studies that identify risk factors for contact injuries, including subsequent injuries, in professional rugby league players (Gabbett et al., 2012); analyse the training load and shoulder injuries in a large youth handball cohort (Møller et al., 2017); study the genetic association with hamstring injuries in soccer players (Larruskain et al., 2018). Furthermore, in recent times, researchers have increasingly applied machine learning approaches, primarily using classification techniques where a binary outcome (injured/non-injured) is predicted (e.g. Rossi et al., 2018). While machine learning methods are appealing and powerful tools in many applications, they typically require large sample sizes for training and hyperparameter tuning. Besides, most classical machine learning methods do not explicitly account for recurrent events or easily handle imbalance classes, such as when there are very few injuries or injured players compared to ready-to-play.

In the application of shared frailty Cox models, challenges persist, particularly when a large number of predictors and numerous parameters increase the complexity, potentially leading to convergence issues in the estimation of frailty terms (McGilchrist and Aisbett, 1991; Therneau et al., 2003). This is particularly problematic for small sample data, as is often the case with sports injury data. Moreover, the data are frequently limited to

individual teams or a small sample of players, resulting in a small total number of injuries from a statistical point of view. Therefore, it becomes essential to reduce the number of parameters to be estimated and to efficiently select a subset of relevant variables associated with the risk of injury.

Aim of the work

The aim of the work is to assess the adequacy and performance of a family of statistical methods for time-to-event data analysis in the context of injury data, focusing on regularization techniques and Cox regression. Our objective is to compare the performance of frailty models that include different sets of previously selected variables, with respect to prediction accuracy.

The work has two major components: (i) an empirical analysis using real data from a single team with 22 players to compare the performance of different approaches, and (ii) a simulation study that systematically evaluates all considered variable selection methods across three different scenarios and varying data sizes.

Outline

In the following section 3.2, we present the methods used, i.e. different regularized Cox models to perform variable selection and shared frailty Cox models to fit the data with a reduced number of variables. Then, in section 3.3, we describe the data that motivated the work and the results obtained from the analysis of these data. In section 3.4, we explain the simulation study carried out and finally, in section 3.5, we conclude with a general discussion.

3.2 Methods

We follow a two-step strategy to manage the large number of potential covariates from the functional screening tests data. In the first step, we utilize various variable selection techniques based on regularized Cox models that do not explicitly account for repeated measures. Next, we fit shared frailty Cox models using the most relevant variables –those comprising a reduced number of variables selected by each method in the first step.

Notation

We define the primary outcome variable as the exposure time in minutes for a player until the occurrence of an injury, denoted as a non-negative random variable T and the censorship as a random variable C. The observed data are then composed by the set $\{(Y_l, \delta_l, X_l)$,

$l = 1, \ldots, N\}$, where $Y_l = \min\{T_l, C_l\}$ and $\delta_l = \mathbb{I}\{T_l \leq C_l\}$ is the censorship indicator and N the total number of observations. We assume that censoring is non-informative and that given $\boldsymbol{x}_l$, y_l and δ_l are independent.

3.2.1 Regularized Cox methods

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

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

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

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

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

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

where $\lambda \geq 0$ and $\alpha \in (0, 1)$ are the regularization tuning parameters and $pl(\boldsymbol{\beta})$ is the Cox partial log-likelihood to be maximized subject to a constraint, that is a penalty function to be multiple of a L_1 or L_2-norm, or a L_0-seminorm. For Group Lasso, the vector of coefficients is partitioned into G groups of size n_g, i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_G')'$. We use the type of (functional screening) test as the grouping factor, see Table B1 in Appendix B; and then, all these G groups are equally penalized. For the sake of simplicity, we only consider Elastic Net with $\alpha = 0.5$. We estimate the best regularization parameter λ by 10-fold cross-validation, for which we use the same cross-validation splits across all models to enable a fair comparison of their performance.

It is worth noticing that, although the Ridge regression technique itself is not a variable selection method, we include it as a regularized method for the comparisons. Hence, the estimated coefficients' 95% confidence intervals are generated via bootstrap, and we check whether the interval includes zero or not. To determine variable selection, we consider

that the variables are selected, when their corresponding coefficients' 95% confidence intervals do not include zero (Chatterjee and Lahiri, 2010; Sartori, 2011).

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

3.2.2 The shared frailty model

We fit the occurrence of non-contact lower-limb injuries by shared frailty Cox model (Hougaard, 1995; McGilchrist and Aisbett, 1991). Such a model considers the dependence, that observations within the same player possibly share, by including a player-specific random effect that acts on the baseline hazard in a multiplicative way. The frailty term accounts for unobserved heterogeneity, as observations within each player may be correlated; and individual characteristics –variables that differentiate players from one another– may often remain unobserved or, in some cases, be unmeasurable.

Let's specify the total number of players by L, where the l-th player has n_l observations (the maximum injury number that player l has been at risk of) indexed by i_l (the i-th injury that player l has been at risk of), so that the repeated measures are explicitly accounted for the data observed, i.e. $\{(Y_{i_l}, \delta_{i_l}, X_{i_l}),\ i_l = 1, \ldots, n_l \text{ and } l = 1, \ldots, L\}$. We denote N as the total number of observations, which is the sum of the number of observations for each player: $N = \sum_{l=1}^{L} n_l$. To deal with recurrent events, we consider the so-called gap time approach (Kelly and Lim, 2000; Ullah et al., 2014). This gap time approach determines the risk interval of each player, in such a way that a new risk interval is set every time the player has totally recovered from an injury and starts to train. Thus, each recurrent event is represented by a separate interval and once an injury has occurred the player is "at-risk" from the starting point of the previous injury recovery, where the time is reset to zero. For a visual representation, see Figure 3.1.

Technically, each observation of the data set corresponds to a single player: some players had not been injured at all during the follow-up, and contributed to censored survival

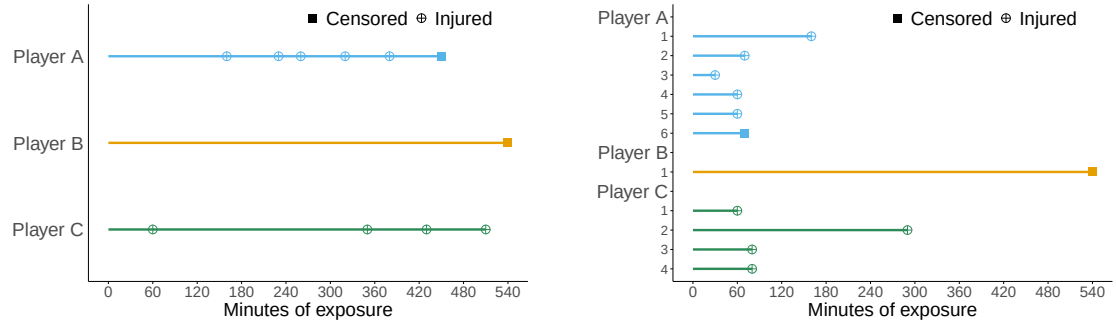


Figure 3.1: Illustration of the risk interval formulation. Left: an example of recurrent time-to-event data representation. Right: the gap time approach representation where each time to an event or censoring is a separate risk interval.

times; others, sustained at least an injury and are, thus, represented by one or multiple survival times.

The hazard rate, at time t, for the i-th observation (representing the number of injury events) of the l-th player is given by:

$$\begin{aligned} \lambda_{i_l}(t|\alpha_l, \boldsymbol{x_{i_l}}) &= \alpha_l \lambda_0(t) \exp\left(\boldsymbol{x}'_{i_l}\boldsymbol{\beta}\right) = \\ &= \lambda_0(t) \exp\left(\boldsymbol{x}'_{i_l}\boldsymbol{\beta} + \boldsymbol{z}'_{i_l} b_l\right), \qquad i_l = 1, \ldots, n_l,\ l = 1, \ldots, L, \end{aligned} \tag{3.1}$$

where λ_0 is the unspecified baseline hazard, p the number of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ the vector of coefficients, $\boldsymbol{x_{i_l}}$ the corresponding row of this observation in the design matrix $\boldsymbol{X}$ and α_l, or $b_l = \ln(\alpha_l)$, the player's frailty term, where matrix $\boldsymbol{Z}$ is a $N \times L$ sparse matrix such that $\boldsymbol{z}_{i_l} = 1$, when i_l-th observation corresponds to player l and 0 otherwise. We use penalized partial likelihood to estimate the regression coefficients and the frailty terms (Ripatti and Palmgren, 2000). As stated by Gasparini et al. (2019), the choice of a particular parametric frailty distribution has minimal impact on the estimation and testing of regression coefficients. We assume the log-normal distribution for the α_l frailty term, i.e. the Gaussian distribution for $b_l = \ln(\alpha_l)$, given the fact that the models fitted to functional screening tests data gave best fits with this distribution according to the Akaike information criterion.

Equivalently, the marginal survival function, i.e. the probability of a player not sustaining an injury at time t, given the covariates, can be derived from Eq. (3.1), integrating out the frailty term from the conditional survival probability as,

$$\begin{aligned} S(t|\boldsymbol{x}) &= \int_{-\infty}^{\infty} S(t|\boldsymbol{b}, \boldsymbol{x}) g(\boldsymbol{b}) \mathrm{d}\boldsymbol{b} = \\ &= \int_{-\infty}^{\infty} S_0(t)^{\exp(\boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{b})} g(\boldsymbol{b}) \mathrm{d}\boldsymbol{b}, \end{aligned} \tag{3.2}$$

where $\boldsymbol{b} = (b_1, \ldots, b_L)$ represents the vector of the frailties and $g(\cdot)$ its density function.

Based on the survival function, it is possible to predict players' injury probabilities for times $t > 0$. In particular, the marginal approach of the survival function in Eq. (3.2), i.e. a population-averaged probability, includes predictions for new players which have not been part of the data used to fit the model. On the contrary, the conditional survival probability approach does not allow estimating predictions of new players –in this case player-specific predictions– since their frailties are unknown. The chosen gap time approach for recurrent events allows predictions of future survival times following an injury, provided that the survival time to be predicted falls within the range of recorded event times used to fit the model.

3.2.3 Evaluation of frailty models

We assess the predictive performance of frailty models through the Brier Score (BS) and the Integrated Brier Score (IBS), i.e. the area under the BS curve (Gerds and Schumacher, 2006; Graf et al., 1999). The Brier Score (BS) is a time-dependent predictive measure used to assess a model's overall performance (Steyerberg et al., 2010). It is commonly employed in survival analysis because it copes with the fact that risk prediction in this field is expressed in terms of probabilities.

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

$$\text{BS}(t|\hat{S}(t|x)) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} \frac{(0-\hat{S}(t|x_i))^2}{\hat{G}(t_i)} & t_i \leq t,\ \delta_i = 1 \\ \frac{(1-\hat{S}(t|x_i))^2}{\hat{G}(t)} & t_i > t \\ 0 & t_i \leq t,\ \delta_i = 0. \end{cases} \tag{3.3}$$

The BS ranges from 0 to 1, with smaller values corresponding to a better prediction. Should random guessing be employed, survival probabilities of 0.5 would be assigned and a BS of 0.25 would be obtained for a random guess.

The IBS is calculated as an overall measure of the model's performance across all available time points:

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

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

Due to the lack of external validation data, and to avoid overfitting, we use the so-called "bootstrap .632+" approach (Efron and Tibshirani, 1997). This method strikes a balance between the apparent BS estimate and the bootstrap BS estimate. It has been demonstrated to provide accurate estimates (Binder and Schumacher, 2008). For additional details on this estimation method, please refer to section B.2 in Appendix B.

3.3 Application to functional screening tests data

Data

We analyse functional screening tests data from a female football team comprising 22 players, which was prospectively followed during the 2017-2018 season. The data include records of players' exposure –specifically, the time spent training and playing matches, measured in minutes–, as well as time-loss non-contact lower-limb injuries (Fuller et al., 2006), recorded by the club's medical staff. Lower-limb non-contact injuries were recorded when a player was unable to participate in a future training session or match due to a physical complaint resulting from football training or match play, and was considered injured until the medical staff cleared the player for full participation in training and match play (Fuller et al., 2006). Players completed biomechanical and functional conditioning screening tests at three different moments during the season: in the preseason, mid-season and at the end of the season. From 200 measured variables, a total of 28 variables were selectively included based on medical experts' criteria. For a detailed list of these variables, please see Table B1 in Appendix B. These variables comprise anthropometric data, as well as results from biomechanical functional tests, assessed as bilateral strength asymmetries of the lower limbs, defined by Impellizzeri et al. (2007).

Concerning outcome variable T, we consider that players' follow-up started at the beginning of the season, i.e. when the first screening test was conducted, and continued until mid-season. At this moment, when new covariates are collected, we reset the time origin of the primary outcome to zero. Consequently, if a player does not sustain an injury during the first (or second) half of the season, we consider the exposure time as censored at the moment of the second screening (or at the season's end). Refer to Figure 3.2 for a better insight.

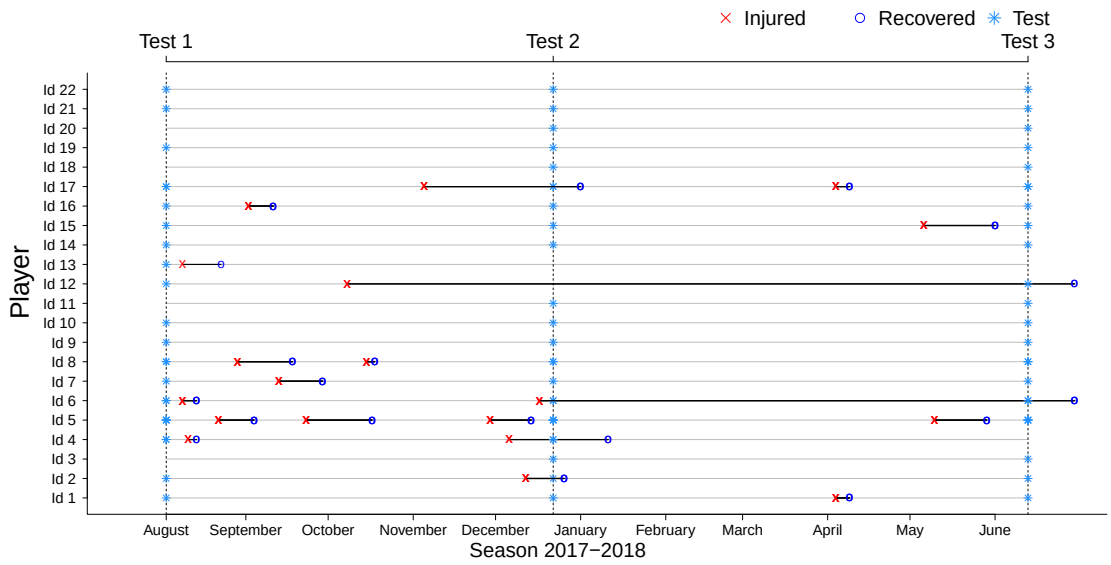


Figure 3.2: Horizontal timeline of football players using functional screening tests data. The red cross indicates the exact moment of the injury occurrence, the blue circle denotes the moment when the player is fully recovered and the bold black line represents time lost due to injury. The three vertical lines correspond to the moment when the screening tests were performed.

Figure 3.2 presents a comprehensive overview of each player's injuries, the number of days lost until full recovery and return to competition, as well as the screening tests each player completed. Data from the third series of screening tests were not utilized, as players' follow-up concluded at that point.

Results

A total of 12 players sustained lower-limb injuries, with the team experiencing 19 injuries in total: seven players were injured once, four players twice, and one player sustained injuries on four occasions. Meanwhile, 45% of the players remained injury-free. The median exposure time for a single player was 13,302 minutes and the cumulative exposure time for the entire team was approximately 250,000 minutes. The team's injury incidence rate was 4.56 injuries per 1000 hours of exposure, and the injury burden was 178.67 days lost due to injury per 1000 hours of total exposure.

The results from the variable selection techniques highlight the unique characteristics of each method. Group Lasso tends to select more variables –all variables within the same group– whereas all other techniques, BeSS, Lasso, Elastic Net, Ridge regression and Cox Boosting, are more restrictive in selecting relevant variables. Figure 3.3 graphically

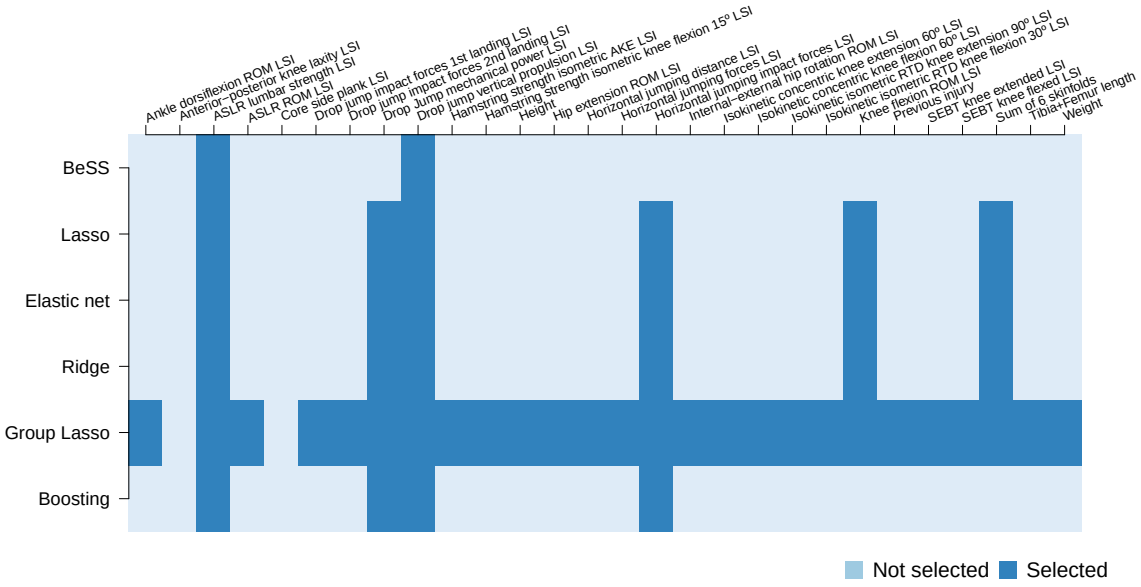


Figure 3.3: Summary of the selected variables (in dark blue) by each variable selection method considered.

displays a summary of the variables that were selected and those that were not by each of the methods. Regarding model sparsity, BeSS yields the most parsimonious model, estimating only 2 out of the 28 variables to be non-zero coefficients. Lasso, Elastic Net and Ridge regression, though using different penalizations, select the same variables. *'ASLR lumbar strength LSI'* and *'Drop jump vertical propulsion LSI'* are the only variables selected by all considered methods, followed by *'Horizontal jumping impact forces LSI'* and *'Drop jump mechanical power LSI'*, which are selected by five of them.

In Figure 3.4, we present the effects of the selected variables on injury risk, accompanied by their 95% confidence intervals, excluding those from the Group Lasso method. The frailty term proves to be significant in all models, emphasizing the need to incorporate such a multiplicative random effect. Additionally, in Figure 3.5, we show the predictive performance of each shared frailty Cox model. Generally, apart from the model based on Group Lasso-selected variables, there is minimal variation in the prediction error curves across the different regularization methods. The model fitted with BeSS-selected variables shows superior predictive performance, with the model employing Cox Boosting-selected variables ranking closely behind. Models based on variables selected by Lasso, Elastic Net, and Ridge regression demonstrate marginally improved performance over the model without any covariate information or frailty term, i.e. the Kaplan-Meier curve derived from all data observations. Conversely, the model based on Group Lasso-selected

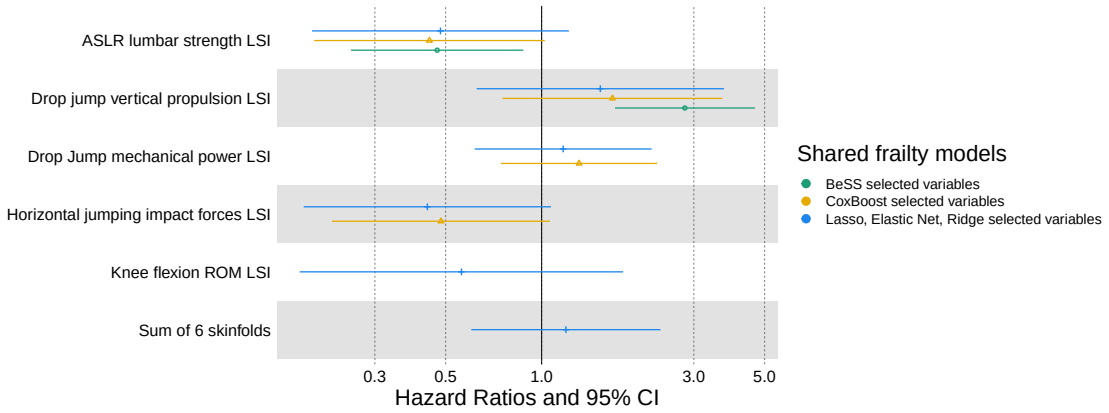


Figure 3.4: Hazard ratios and 95% confidence intervals of the fitted shared frailty Cox models with the set of variables selected by BeSS, Lasso, Elastic Net, Ridge regression and Cox Boosting. Log scale is used for the x-axis.

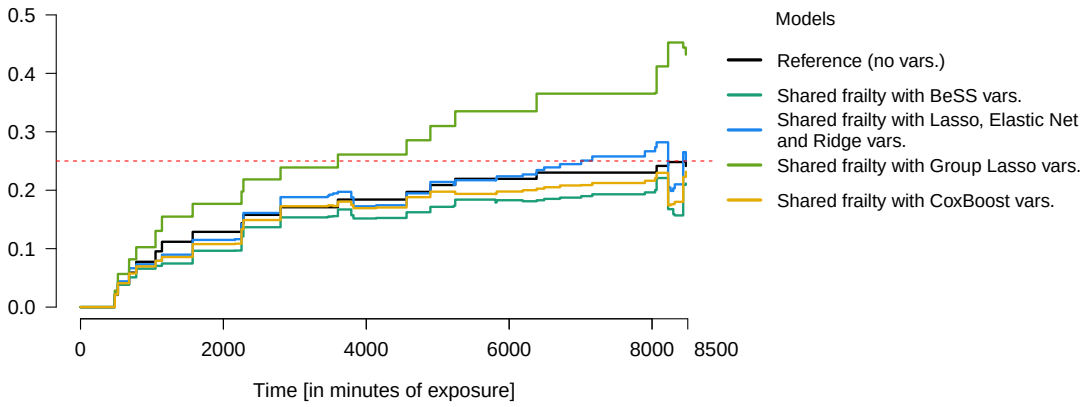


Figure 3.5: Comparison of Brier Score estimates using the bootstrap .632+ method with 30 bootstrap samples for four Gaussian frailty models, each fitted with variable sets selected by BeSS, Lasso (Elastic Net and Ridge regression), Group Lasso, and Cox Boosting. A Kaplan-Meier curve serves as the reference model, without considering any variables or frailty terms.

variables exhibits the least favourable performance. This discrepancy may be attributed to the fact that the variables chosen by the other regularization methods span several groups of screening tests. In general, the prediction errors at early follow-up times are low and comparable across all models. However, at later time points, the prediction error increases since less information is available.

3.4 Simulation study

Due to limitations caused by the characteristics of functional screening tests data, we conduct a simulation study to explore several hypothetical controlled situations. The objective is to evaluate the applicability and robustness of the statistical approaches discussed, by establishing three hypothetical controlled situations reflective of sports injury data contexts. In these scenarios, we pay special attention to the impact of varying sample sizes.

3.4.1 Simulation design

The simulation procedure is summarized in three steps, following, in part, the structured strategy proposed by Morris et al. (2019) for planning simulation studies:

Step 1: Generation of the data

The underlying data-generating process is designed to closely emulate the original functional screening tests data, characterized by a time-to-event outcome with many censored observations and a high number of covariates.

We consider three different scenarios: (i) augmenting the original application data by resampling through bootstrap and adding random noise; and generating time-to-event observations that arise from covariates that share (ii) a weak correlation and (iii) a high correlation. The detailed explanation of these three scenarios is provided in the following *"Parameters defining simulation setting"* section.

In this regard, we employ a modified version of the random spline method, proposed by Harden and Kropko (2019), to simulate the true underlying data-generating process. This method does not assume any distributional form for the baseline hazard function and thus, it matches the Cox model's inherent flexibility. It requires initially determining the number of points –or knots– to be drawn to fit a cubic spline for the baseline hazard function. The method is slightly modified in a way to include a multiplicative random effect, i.e. a frailty term (see section B.3 in Appendix B).

Step 2: Fitting the models

We fit the six regularized Cox methods described in the previous Section 3.2 and select small sets of variables. Afterwards, for each of the six sets of selected variables, we fit shared frailty Cox models to the data, accounting for the players' unobserved variability. We assume that the frailty term follows a Gaussian distribution, in accordance with the functional screening tests data analysis.

Step 3: Performance Measures

We repeat the previous two steps N_{sim} times, for a given prespecified configuration. Finally, we assess the models by a number of different measures that evaluate both, (i) the model performance and (ii) the predictive accuracy of the final shared frailty models. We compare the model performance by assessing how well the estimated models represent the underlying true model. Thus, we evaluate the selection of significant variables through the measures presented in Table 3.1; additionally, we quantify the differences between the true and estimated coefficients by the mean squared error (MSE), defined as,

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

Table 3.1: Summary of measures used to evaluate the performance of variable selection methods.

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

† The value one would obtain if the variable selection method always found the correct model. For the first cell, i.e. the average model size, it refers to each one of the settings.

On the other hand, we evaluate the predictive accuracy of the shared frailty models using the BS and the IBS. This evaluation is repeated for each N_{sim} replica. To summarize the overall predictive accuracy, we report the medians of the IBS in the [0,1000] and [0,3500] time intervals.

Parameters defining simulation settings

In this section, we provide detailed descriptions of the three simulation scenarios. See Table 3.2 and Table 3.3 for a summary. Table 3.2 shows fixed parameters common to all settings, while Table 3.3 lists parameters unique to each setting, such as the true vector of coefficients, true vector of frailties, number of players and number of observations per

each.

Table 3.2: Fixed parameters for the simulation study.

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

The first scenario (see the first block in Table 3.3) is designed according to the results obtained from the application data analysis. Within it, we consider three different settings based on the estimated coefficients obtained from these data. In the first setting, the vector of coefficients $\boldsymbol{\beta}$, that generates the underlying true data, is fixed to be the vector of coefficients estimated by the frailty model based on BeSS-selected variables. The second and third settings follow the same approach. The second setting uses the vector of coefficients $\boldsymbol{\beta}$ obtained by the frailty model based on Lasso, Elastic Net and Ridge regression-selected variables. Meanwhile, the third setting uses the vector obtained from the frailty model based on Cox Boosting-selected variables. Furthermore, we assume that the data consists of 66 players, equivalent to three average-sized teams, denoted as $L = 66$. Each player has three repeated observations, represented as $n_l = 3$ for all $l = 1, \ldots, L$, and there are $p = 28$ variables. To expand the design matrix of the application data, we incorporate resampled rows by drawing bootstrap samples with replacement and repeating each sampled value three times with added random noise. The grouping vector used for the Group Lasso remains consistent with the application data, reflecting the categorization based on the type of functional screening test.

Conversely, the second scenario and third scenario's design matrices (see the second and third blocks in Table 3.3) are generated from equally distributed normal variables. In both scenarios, we set the vector of coefficients, $\boldsymbol{\beta}$, to be the same, and we assume the frailty term to follow a normal distribution centred at zero with a standard deviation of 0.3. The key difference between these scenarios lies in the correlation structure among the variables. Scenario 2 assumes independent variables, whereas Scenario 3 considers a pairwise correlation between each pair of variables $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, $\rho_{i,j}$, to be $0.65^{|i-j|}$.

In both scenarios, we consider four sample sizes, determined by varying the number of players, denoted as L where $L \in \{22, 66, 132, 220\}$ players –equivalent to 1, 3, 6, and 10 football teams with an average of 22 players each. Each player has a different random

Table 3.3: Parameter settings for each scenario of the simulation study.

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

number of observations with $p = 50$ variables. The number of observations per player is generated using a truncated Poisson distribution with a mean of 3, resulting in a total of 60, 191, 391, and 670 observations. Importantly, the number of observations per player and the frailty vector remain consistent in the second and third scenarios. The grouping vector for Group Lasso consists of ten groups, each containing five variables, i.e. $G = 10$ and $n_g = 5$.

Software issues

All computations are performed in **R** version 3.6.2 (R Core Team, 2023), on a 64-

bit Linux platform with an Intel Core 2.2 GHz CPU of 2 cores and 243.4 GByte RAM. The code for the simulations is available in the following GitHub repository: `https://github.com/lzumeta/TimeToEvent-InjurySim`. The six regularized methods are implemented through **BeSS** 1.0.6 (Wen et al., 2020b), **glmnet** 4.1-7 (Friedman et al., 2010), **grpreg** 3.4.0 (Breheny and Huang, 2015) and **CoxBoost** 1.3 (Binder, 2013) **R** packages. Shared frailty Cox models are fitted using the `coxph()` function from the **survival** 3.5-5 (Therneau, 2020) package, and the BS and IBS are computed using the **pec** 2023.04.12 (Mogensen et al., 2012) package. The **furrr** 0.3.1 (Vaughan and Dancho, 2022) package is employed for running the simulation study.

3.4.2 Results

Table 3.4 summarizes the results for the three settings in Scenario 1. With regards to the first setting, the best model is BeSS (on which the setting is based), followed by Ridge regression. BeSS outperforms with an average of 4.17 wrongly selected variables, an MSE of 3.78, and IBS medians of 0.045 (between $[0, 1000]$) and 0.086 (between $[0, 3500]$). The frailty model based on Cox Boosting-selected variables also performs well.

The second setting shows no significant differences in model performances compared to the methods from which it's generated, i.e., the frailty model with six variables selected by Lasso, Elastic Net, and Ridge regression. Ridge regression performs the best in terms of MSE. BeSS is the second-best model with an average of 1.35 wrongly selected variables, an MSE of 5.05, and good Brier Scores. Cox Boosting also shows robust performance with respect to the IBS medians.

Regarding the third setting from Scenario 1, results indicate that not only Cox Boosting, the method on which the setting is based, performs well, but BeSS and Ridge regression are also suitable. BeSS stands out in certain metrics, particularly in the average number of wrongly selected variables (0.62 versus 6.89 for Cox Boosting) and MSE (5.24 versus 6.17 for Cox Boosting). Both methods exhibit similar IBS median values.

Table 3.5 presents results for Scenario 2 and Scenario 3 for sample sizes of $N_{\text{obs}} \in \{60, 191, 391\}$. Results for the setting with $N_{\text{obs}} = 670$ observations are in Table B2 in Appendix B. In general, differences between frailty models decrease with larger sample sizes, whether considering MSE or IBS. Figure 3.6 shows that prediction errors become smaller with more observations. For example, in Scenario 2, frailty models based on BeSS or Group Lasso-selected variables reduce their prediction error range by 67.9% (from 0.131 to 0.042) and 72.1% (from 0.176 to 0.049), respectively, when increasing the number of teams (and thus, the sample size) from 1 to 10. In Scenario 3, the decrease in the range of

Table 3.4: Simulation results for the three different settings within Scenario 1, involving 66 players with 3 observations each, resulting in a total sample size of 198. The measures analyzed include the AMS, ANFS, ANFNS, MSE and the median of IBS calculated over the $[0, 1000]$ and $[0, 3500]$ time intervals, for all models.

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

Table 3.5: Simulation results for Scenarios 2 and 3, which consider different correlation structures of covariates ($\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$) for a varying number of players, $L \in \{22, 66, 132\}$, resulting in $N_{\text{obs}} \in \{60, 191, 391\}$ observations, respectively. The measures analyzed include the AMS, ANFS, ANFNS, MSE and the median of IBS calculated over the $[0, 1000]$ and $[0, 3500]$ time intervals, for all models.

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

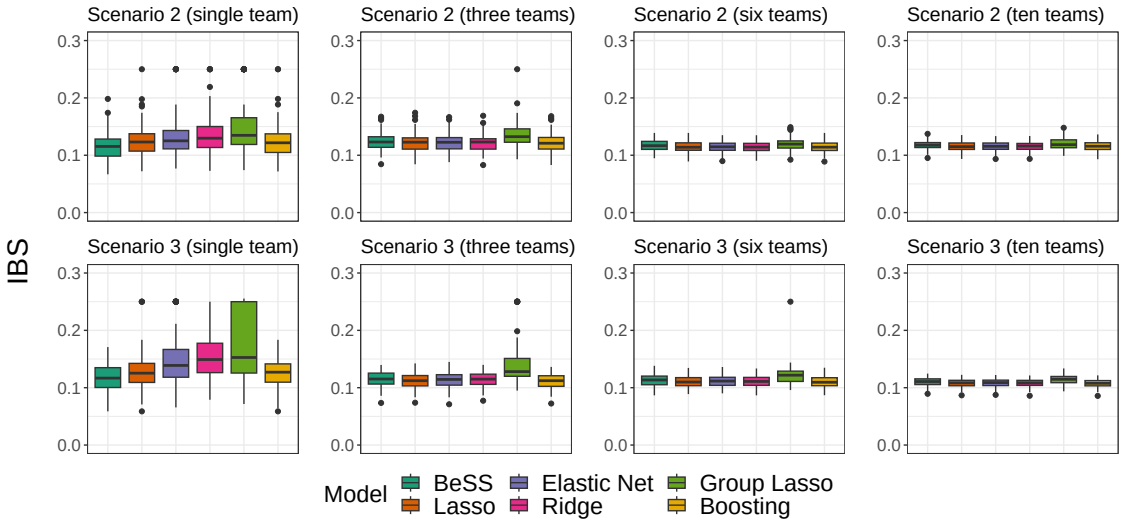


Figure 3.6: Distribution of the IBS over the $[0, 3500]$ time interval for each of the six frailty models considered across Scenarios 2 and 3, that consider different correlation structure of covariates, $\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$, for varying team sizes consisting of $L \in \{22, 66, 132, 220\}$ players.

prediction errors for these models is 67.8% (from 0.112 to 0.036) and 78.1% (from 0.183 to 0.04), respectively.

It's important to note that prediction errors depend on the time interval considered. In the [0, 1000] time interval, the IBS values are comparable across different models due to similar early-time predictions. At longer time intervals, the disparities in IBS become more apparent between models. Specifically, within the [0, 1000] range, Scenario 2 exhibits marginally lower median IBS values than Scenario 3. Conversely, over the longer $[0, 3500]$ interval, the median IBS values for Scenario 3 are slightly lower than those for Scenario 2.

When covariates in the data set exhibit low dependence, the resulting models tend to be sparser –that is, they include fewer variables. Consequently, the average model sizes in Scenario 2 (where $\rho_{ij} = 0$ for all $i \neq j$) are closer to the average size of the true model, which is five. In both scenarios, smaller sample sizes lead to more pronounced differences in prediction error curves and in the variable selection methods' ability to identify true effects. For the smallest sample size, Group Lasso, and to some extent, Ridge regression and Elastic Net, exhibit notably higher prediction error ranges and median prediction errors compared to other methods. This trend is even more pronounced with correlated covariates in Scenario 3 (see Figure B4 in Appendix B).

BeSS selects the fewest variables, followed by Ridge regression. In contrast, Group Lasso selects the largest number of variables, leading to more complex models (see Fig-

ure B3 in Appendix B). The number of falsely selected variables and incorrectly estimated coefficients align with each method's selection tendencies. BeSS, selecting fewer variables, has fewer falsely selected variables and a lower number of coefficients incorrectly estimated as 0, aligning with the true effect. Group Lasso, on the other hand, results in a high average number of falsely selected variables but a low number of coefficients incorrectly estimated as 0. In general, methods leading to sparse models, such as BeSS and Cox Boosting, perform better, especially with small sample sizes. As the sample size increases, differences between models decrease, except for Group Lasso.

3.5 Discussion

Recurrent events models have been widely used in numerous biomedical studies, but, to our knowledge, their application in the field of sports injury prevention has been limited. Our work aimed to provide an appropriate statistical modelling strategy for football injury data, which presents some challenges, rather to provide evidence about risk factors for lower-limb injuries. Research on sports injuries is undergoing a significant shift, with an increasing emphasis on more powerful analytical methods. We believe that, despite the limitations of our application, recurrent time-to-event methods hold great potential to advance sports injury research. Further investigations in larger cohorts, spanning multiple seasons and involving various sports teams, are necessary to apply the proposed methodological approach and increase knowledge of sports injury risk factors. In this context, it should be noted that such expansions may introduce other levels of complexity in the data; for instance, an additional random effect could account for team-specific variations resulting from different training styles.

Contributions and practical application

The analyses and simulation studies performed suggest that the methodology presented is useful for identifying screening tests associated with the risk of injury (variable selection), addressing the recurrent time-varying nature of sports injury data (frailty), for the sports medicine practice in a professional football team. However, as statisticians, it is important to convey to medical services in professional sports teams that despite conducting numerous functional screening tests, the small sample size of individuals and lower-limb injury events limit the usefulness for predicting the risk of injury. Our simulation study results confirm the assumptions about the reliability and robustness of estimated effects in such small data sets. In a real-world scenario involving 22 players, the model's predictive performance heavily relies on the choice of variable selection technique. BeSS and

likelihood-based Cox Boosting perform well with small sample sizes common in sports injury data. Conversely, Group Lasso, and to some extent, Ridge regression and the Elastic Net, show higher prediction errors. For larger cohorts, the choice of variable selection technique becomes less critical.

In conclusion, our work highlights the potential of shared frailty Cox models in sports injury prevention. Though conclusions from models based on very small sample sizes (e.g. a single football/sports team of about 20 players) should be drawn with caution due to high variability, using data from three or six teams already leads to strong improvements. Regardless of the chosen modelling strategy, the key to enhancing predictive performance and accuracy in injury predictions lies in the size of available data. To gain valuable insights into sports injury prediction and monitoring, we recommend sports clubs invest in collecting more data, such as conducting regular functional screening tests for several of their teams.

Limitations

Our work puts attention on the Cox model as one of the most classical approaches for modelling time-to-event data, and its extension to recurrent events data, the shared frailty Cox model. We first applied regularized Cox models, motivated by the large number of covariates present in the data. In the second step, we fitted shared frailty Cox models using a reduced set of selected variables. However, we acknowledge that the techniques used in the first step do not account for the correlation between groups of observations, which may result in flawed variable selection. A better approach would involve jointly performing both steps: selecting the important variables and fitting the model.

We now discuss the choice of the methods employed and also, the ongoing research on the (simultaneous) regularization of frailty models.

Alternative approaches and further work

While our focus has been on statistical regularization techniques for variable selection, it's important to note that modelling approaches with variable selection are not limited to these methods. There are various methods available, many of which come from the field of machine learning, including tree-based survival techniques like recursive partitioning and random forests, survival principal component analysis, support vector machines, and more (LeBlanc and Crowley, 1992; Bair et al., 2006; Li and Luan, 2002). However, it's worth mentioning that most machine learning algorithms are based on the assumption of independent and identically distributed (i.i.d.) training data. They often require large training data sets and may not perform well with imbalanced cases, such as those

involving a very low number of injuries, which is the case in our context. Thus, without modifications, most machine learning algorithms are not directly applicable to non-i.i.d. data. Future research aims to incorporate machine learning survival techniques into the comparison of available methods for sports injury data. Benchmark studies involving machine learning survival approaches have been conducted in other research areas, such as modelling disease outcomes with genome data (Herrmann et al., 2020) or multivariate and random survival trees (Su and Fan, 2004; Ishwaran et al., 2008), but many of these methods rely on large sample sizes for training and parameter tuning.

On the other hand, there are alternative survival methods for recurrent events data that could be considered, including parametric survival models, variance-corrected Cox models, and spline-based survival models. Nevertheless, for our statistical analyses based on real data, the shared frailty Cox model was preferred over all these alternatives. Parametric survival models, like accelerated failure time models (Pan, 2001), require making distributional assumptions about the time-to-event outcome. Variance-corrected Cox models, such as Andersen-Gill 1982, Prentice-Williams-Peterson 1981, and Wei-Lin-Weissfeld models 1989, address correlation by using robust standard errors to model the marginal distribution of each event time with corrected variance. In contrast, the shared frailty Cox model corrects dependence among recurrent event times by considering a random effect. That is to say, it assumes that some players are intrinsically more or less prone to experience an injury. Spline-based survival approaches provide a compelling framework, including generalized survival models (Liu et al., 2017) and piece-wise exponential additive mixed models (Bender et al., 2018). These models can estimate the baseline hazard with smooth functions, incorporate random effects, and offer flexibility for various covariate effects. However, they typically require estimating more parameters and demand larger data sets compared to the shared frailty Cox approach.

Lastly, the literature provides some strategies to simultaneously perform variable selection and frailty model estimation. A first approach was proposed by Fan and Li (2002), who used a penalized likelihood estimator with smoothly clipped absolute deviation penalty (SCAD), for variable selection in gamma frailty models. Androulakis et al. (2012) extend this methodology for penalized gamma frailty models, but as of yet, no open-source software implementation is available. A recent penalization approach by Groll et al. (2017) focuses on variable selection in frailty models with time-varying coefficients such that single varying effects are either included, included in the form of constant effects or totally excluded. The method is implemented in the **PenCoxFrail R** package (Groll, 2016). This method was beyond the scope of our work since we do not consider time-varying covariates or time-varying effects. Newly, Hohberg and Groll (2020) proposed a

more general Lasso Cox frailty approach allowing to perform variable selection, even for non-time-varying covariates.

Chapter 4

Flexible time-to-event modelling approaches for recurrent football injury data

Contributing article

Zumeta-Olaskoaga, L., Bender, A. & Lee, D.-J. (2023). Flexible modelling of time-varying exposures and recurrent events to analyze training load effects in team sports injuries. *Manuscript submitted for publication*.

Code repository

https://github.com/lzumeta/flex-mod-training-loads-recu-injuries

4.1 Context

In this chapter, we study flexible modelling approaches for analyzing time-varying exposures, such as training load, and recurrent events in the context of team sports injuries. Today, we have access to a wealth of regularly collected data, primarily through Global Positioning System (GPS) devices. These devices play a crucial role in monitoring and quantifying various facets of training load, including the duration of training sessions and competitions, distance covered, as well as speed and power output metrics. The ever-increasing amount of data now being collected opens up new opportunities but also introduces novel challenges.

The analysis of an athlete's exposure status over time is widely recognized as crucial

in understanding the aetiology of sports injuries, especially concerning recurrent, subsequent, and exacerbated injuries (Nielsen et al., 2020). Athletes, referred to as players in this dissertation, are consistently exposed to high competition demands, that, in turn, increase the strain on their bodies and exposure to injury risk. Effective injury prevention depends on the athlete's capacity to tolerate repeated exposures to injury risk.

The physical load from training and competition is often termed as training load, which is defined as "the cumulative stress placed on an individual from multiple training sessions and games over a period of time" (Gabbett et al., 2014). Consequently, training load can be applied to the athlete over varying time periods and with varying magnitudes (Soligard et al., 2016). Indeed, the study of training load is key to developing effective training plan strategies that enhance athletes' performance while also reducing their risk of injury. The relationship between training load and injury, however, remains uncertain (Windt et al., 2018; Griffin et al., 2020).

Quantifying this relationship requires the development of an etiologically plausible time-varying exposure model, which estimates how previous training affects the injury risk (Impellizzeri et al., 2023). The effects of past exposures may cumulate over time and exhibit complex forms of association. Additionally, the model must account for potential associations between subsequent injuries within players. To address these concerns, especially the dependencies resulting from subsequent injuries and the varying intensity and duration of past exposures, we propose a Piece-wise exponential Additive Mixed Model (PAMM, Bender et al. 2018) with weighted cumulative exposure-type (WCE) cumulative effects (Sylvestre and Abrahamowicz, 2009).

Related work

PAMMs are a semi-parametric extension of the Piece-wise Exponential Model (PEM, Holford 1980; Laird and Olivier 1981; Friedman et al. 1982) that allow for penalized estimation of flexible survival models with a wide range of covariate effects, such as non-linear, time-varying effects, cumulative effects, and/or random effects (refer to Bender et al. 2018 and Argyropoulos and Unruh 2015 for a thorough overview; see also Chapter 2). This framework has also been shown to support the estimation of cumulative effects of time-varying exposure histories. The WCE-type cumulative effect suggested by Sylvestre and Abrahamowicz (2009), a weighted sum of all past exposures over a relevant time window, is a common way to address this. The weight function assigns weights to past exposures based on the time elapsed since the exposure occurred, which, ideally, is determined according to the true underlying biological mechanism. They proposed to estimate the weight function using B-spline regression.

The PAMM methodology has been employed in many recent publications. Bender et al. (2019) explore complex exposure-lag-response associations and provide a general formulation of PAMMs that includes previous approaches for cumulative effects like the WCE model and the distributed lag non-linear model (DLNM, Gasparrini et al. 2017), as special cases. Ramjith et al. (2022) study the PAMM framework for recurrent events analysis and show that under the assumption of proportional hazards, PAMM and the shared frailty Cox model (McGilchrist and Aisbett, 1991) are equivalent. Danieli and Abrahamowicz (2019) and Li et al. (2022) introduce approaches to model cumulative effects of time-varying exposures with competing risks, via cause-specific hazards model and subdistribution hazards model for each competing event, respectively, by incorporating separate Cox WCE models with an event-specific weight function. Recently, in the field of sports medicine, Bache-Mathiesen et al. (2022) evaluated different methods to assess the cumulative effect of training load on the risk of injury in team sports and suggested the use of DLNM.

Aim of the work

In this work, we aim to extend the PAMM by incorporating WCE-type cumulative effects in the recurrent events setting combined with a method to identify a relevant time window in which past exposures have an effect. We further demonstrate the practical application of this model in the field of sports medicine.

Outline

The proposed modelling framework is detailed throughout section 4.2, which first introduces the PAMM framework for recurrent events with time-constant covariates and then focuses on how we adapt it to flexibly model time-varying exposures and recurrent events, in addition to how we penalize the model for identifying a relevant window. Section 4.3 illustrates a real-world application of this method to assess the cumulative effects of past training exposures on the hazard of subsequent injuries in a football team, while section 4.4 describes the simulation study carried out to evaluate the model performance. The final section 4.5 concludes the work with a discussion.

4.2 Methods

As already mentioned in Chapter 2, PAMMs transform a survival task into a Poisson regression task by partitioning the follow-up period into a finite number of intervals and assuming that hazards are piece-wise constant in each of these intervals.

Data in "standard" time-to-event format

player (I)	t (I)	status (I)	enum	position (I)
1	20	1	1	midfielder
1	120	0	2	midfielder
2	5	0	1	attacker
3	2	1	1	left-winger
3	22	0	2	left-winger

Data in Piece-wise Exponential Data (PED) format

player (I)	interval (I,j)	offset (I,j)	status (I,j)	enum (I, j)	position (I)
1	(0, 5]	log(5)	0	1	midfielder
1	(5, 20]	log(15)	1	1	midfielder
1	(0, 5]	log(5)	0	2	midfielder
1	(5, 20]	log(15)	0	2	midfielder
1	(20, 100]	log(80)	1	2	midfielder
2	(0, 5]	log(5)	0	1	attacker
3	(0, 5]	log(2)	1	1	left-winger
3	(0, 5]	log(5)	0	2	left-winger
3	(5, 20]	log(15)	0	2	left-winger

Figure 4.1: An illustration of data transformation into PED representation using the following interval cut-off points: 0, 5, 20, and 100.

When we have recurrent events, the data transformation is carried out as follows:

Let's consider that we have data in the "standard" time-to-event format, as shown on the right-hand side of Figure 4.1. Now, if we take, for instance, these 0, 5, 20, and 100 values as cut-points; we generate as many rows as there are intervals in which each player is at risk of injury. In Figure 4.1, player 3 has suffered an injury at time 2. For this first injury of this player (first event), we generate only one row, one interval, between $(0, 5]$. After recovering from this injury, we observe the player again. The player is at risk of his/her second event (`enum = 2`). Then, we generate two rows, two intervals, both with `status` = 0, since the last time we observed the player (`t = 20`, in gap-time approach), she/he was not injured. The offset column indicates how long the player has been at risk in that interval. Generally, the offset is the logarithm of the length of that interval. However, in cases like the first injury of player 3, where the player has suffered an injury, the offset is the logarithm of the total time they have been at risk in that interval, i.e., $\log(2)$.

The challenging aspect of this transformation lies in effectively integrating or combining the PED-transformed data which include event information and time-constant covariates, with the data that contain time-dependent covariate information. We have worked out this technical issue and implemented it in code into the **pammtools** (Bender and Scheipl, 2018) **R** package. Now, the `pammtools::as_ped()` function provides support for this scenario, i.e. for the context of recurrent events with time-dependent covariates. Find more details in this pull request on GitHub: https://github.com/adibender/pammtools/pull/224.

Formally, let partition the follow-up period $(0, t_{\max}]$ into J intervals with $J + 1$ cut points, i.e. $0 = \kappa_0 < \kappa_1 < \ldots < \kappa_J = t_{\max}$, and let assume the hazard to be constant in each interval. The general expression of the hazard rate of the i-th injury (event) of the

l-th player is given by,

$$\lambda_{i_l}(t|\boldsymbol{z}_l(t), l) := \exp\left(f(\boldsymbol{z}_l(t), t_j, l)\right) = \lambda_{i_l,j} \tag{4.1}$$

for all $t \in (\kappa_{j-1}, \kappa_j]$, $j = 1, \ldots, J$, $t > 0$ and $l = 1, \ldots, L$; $i = 1, \ldots, n_l$, where t is the time of interest, t_j a fixed time point in the j-th interval (e.g. $t_j := \kappa_j$), $\boldsymbol{z}(t) \in \mathbb{R}^p$ potentially time-dependent covariates and $f(\cdot)$ the effect of (time-dependent) covariates on the hazard, that can be potentially (non-linearly) time-varying and injury-specific. This general representation allows us to study different dependence structures arising in football injury data.

We generally use the event times as cut-points to partition the follow-up time, following the empirical discussion by Bender (2018).

Next, we introduce the PAMM framework for recurrent events. Initially, we present the simpler PAMM models for time-constant covariates and recurrent events –namely, the *stratified PAMM* and the *shared frailty PAMM*. Subsequently, we introduce the broader and more flexible PAMM for modelling time-dependent covariates in the context of recurrent events. Within this section, our emphasis is on analyzing WCE-type cumulative effects.

4.2.1 PAMM for recurrent events with time-constant covariates

Relaxing some of the terms in (4.1), we investigate the following models:

Stratified PAMM models

They assume different baseline hazards for each of the injury events and thus, consider the dependence induced by the previous injuries.

$$\begin{aligned}\lambda_{i_l}(t|\boldsymbol{x}_l) &:= \lambda(t|\boldsymbol{x}_l, l, i) = \lambda_{0,i}(t)\exp(\boldsymbol{x}_l'\boldsymbol{\beta}) = \\ &= \exp\left(\beta_{0,i} + f_{0,i}(t_j) + \boldsymbol{x}_l'\boldsymbol{\beta}\right), \quad \forall t \in (\kappa_{j-1}, \kappa_j].\end{aligned}$$

Shared frailty PAMM models

They assume a common baseline hazard for all events ($\lambda_{0,i} = \lambda_0$), but account for within-player correlation by introducing a frailty term.

$$\begin{aligned}\lambda_{i_l}(t|\boldsymbol{x}_l) &:= \lambda(t|\boldsymbol{x}_l, b_l, i) = \lambda_0(t)\exp(\boldsymbol{x}_l'\boldsymbol{\beta} + b_l) = \\ &= \exp\left(\beta_0 + f_0(t_j) + \boldsymbol{x}_l'\boldsymbol{\beta} + b_l\right), \quad \forall t \in (\kappa_{j-1}, \kappa_j],\end{aligned}$$

where $b \sim N(\mathbf{0}, \boldsymbol{D})$ is a Gaussian random effect.

4.2.2 PAMM for recurrent events and time-dependent covariates

As we proceed with the modelling of injuries, the hazard rate of the i-th injury (event) of the l-th player, given the player's training exposure history $\boldsymbol{z}_l(t) = \{z_l(t_z) : t_z \leq t\}$, is expressed as:

$$\begin{aligned}\lambda_{i_l}(t|\boldsymbol{z}_l(t), b_l) &= \lambda_0(t)\exp(g(\boldsymbol{z}_l(t), t_j) + b_l) = \\ &= \exp(\beta_0 + f_0(t_j) + g(\boldsymbol{z}_l(t), t_j) + b_l),\end{aligned} \tag{4.2}$$

for all $t \in (\kappa_{j-1}, \kappa_j]$, $t > 0$ and $t_j := \kappa_j$, where $\kappa_0 = 0$, $\kappa_J = t_{\max}$ and κ_j, $j = 1, \ldots, J-1$, are the $J+1$ cut points defining J intervals that partition the study follow-up time $(0, t_{\max}]$. In Eq. (4.2), the expression $\beta_0 + f_0(t_j)$ denotes the log-baseline hazard, where $f_0(t_j)$ is expressed as a smooth term of the form $\sum_{m=1}^{M'} \gamma_{0m} B_m(t_j)$. The term $g(\boldsymbol{z}_l(t), t) = \int_{\tau(t)} h(t, t_z, z_l(t_z))dt_z$ denotes the cumulative effect of $\boldsymbol{z}_l(t)$ at time t, i.e. the past exposure effects of $\boldsymbol{z}_l(t)$ cumulating over a relevant time-window $\tau(t)$, resulting in a sum of weighted effects. The dependence induced by subsequent injuries is accounted for by b_l, a Gaussian random effect (i.e. a shared frailty term) associated with player l, which acts as a random intercept term for the l-th player, i.e. $b_l \sim N(\mathbf{0}, \sigma_b^2)$.

For a WCE-type effect, in Eq. (4.2), we consider time-varying exposure effects weighted by latency $t - t_z$ and linear in $z(t_z)$. That is, the contribution of covariate $\boldsymbol{z}(t)$ observed at time t_z with value $z(t_z)$, is defined by $h(t, t_z, z(t_z)) := h(t - t_z)z(t_z)$, and called partial effect. Thus, the cumulative effect $g(\boldsymbol{z}(t), t)$ at follow-up time t is the integral of these partial effects over exposure times t_z contained within the so-called lag-lead window, $\tau(t)$, which controls how many observations of z contribute to the cumulative effect at time t (with the minimal requirement being that $t_z \leq t$).

Let $\boldsymbol{z}(t) = \{z(t_z) : t_z \leq t\} = \{z(t_{z,1}), \ldots, z(t_{z,Q})\}$ be the set of all registered exposure variables up to time t. Then, $g(\boldsymbol{z}(t), t)$ is estimated with penalized splines (e.g., by using P-splines (Eilers and Marx, 1996, 2021) that penalize the differences of neighbouring basis coefficients) and with quadrature weights $\Delta_q = t_{z,q} - t_{z,q+1}$ (and $t_{z,0} = t$), the time difference between two consecutive exposure measurements, for numerical integration, as follows:

$$\int_{\tau(t)} h(t - t_z)z(t_z)dt_z \approx \sum_{q=1}^{Q} \tilde{\Delta}_q \tilde{h}(t - t_{z,q}) = \sum_{q=1}^{Q} \tilde{\Delta}_q \sum_{m=1}^{M} \gamma_m B_m(t - t_{z,q}) \tag{4.3}$$

with $\tilde{\Delta}_q = z(t_z)(t_{z,q} - t_{z,q+1})$ if $t_{z,q} \in \tau(t)$ and 0 otherwise; $B_m(\cdot)$ B-spline basis functions and γ_m the associated spline coefficients.

Penalization of the weight function

One of the challenging issues is to determine a relevant time window $\tau(t)$. Without solid prior knowledge, it can be defined as $\tau(t) = \{t_z : t \geq t_z\}$, so all past exposures, collected before actual time t, contribute to the cumulative effect $g(\boldsymbol{z}(t), t)$. Yet, it is plausible that the effects of exposure variables may not be everlasting. As time passes, the effect of exposures recorded long ago may smoothly decrease to zero and eventually disappear. The exact length of the window, however, is usually unknown.

Subsequently, adapting the method by Obermeier et al. (2015) to the PAMM framework, we present two approaches to penalize the weight function, allowing it to transition smoothly to zero at the right end of the support interval: (i) *a constrained-effect approach* and (ii) *a ridge-penalty approach*.

Constrained-effect approach

In this approach, we force the weight function, and its first derivative, to reach the zero value at time $t - t_{z,Q}$, by imposing the last two coefficients in Eq. (4.3) to be equal to zero (Sylvestre and Abrahamowicz, 2009). Working in matrix notation, the right part of Eq. (4.3) is given by:

$$g(\boldsymbol{z}(t), t) \approx \tilde{\boldsymbol{\Delta}}' \, \boldsymbol{B} \, \boldsymbol{\gamma} = \tilde{\boldsymbol{\Delta}}' \, \tilde{\boldsymbol{\gamma}} \tag{4.4}$$

with

$$\boldsymbol{B} = \begin{pmatrix} B_1(t - t_{z,1}) & \dots & B_M(t - t_{z,1}) \\ \vdots & & \vdots \\ B_1(t - t_{z,Q}) & \dots & B_M(t - t_{z,Q}) \end{pmatrix},$$

denoting a $Q \times M$-dimensional basis matrix of B-splines of degree d, $\boldsymbol{\gamma}$ denoting a $M \times 1$ column-vector of associated spline coefficients and $\tilde{\boldsymbol{\Delta}}$ denoting a $Q \times 1$ column-vector having $\tilde{\Delta}_q$ as elements for $q = 1, \dots, Q$.

Therefore, the value $z(t_{z,Q})$ is assumed to have no impact on the current risk at t, by constraining the two last spline coefficients to zero, i.e. $\tilde{\gamma}_{Q-1} = \tilde{\gamma}_Q = 0$.

Ridge-penalty approach

This approach consists of adding a shrinkage L2-penalty to penalize the last B-spline basis coefficients of $\boldsymbol{\gamma}$ in Eq. (4.4), see Obermeier et al. (2015).

Hence, we seek that the last coefficient $\tilde{\gamma}_Q$ to be close to zero:

$$\tilde{\gamma}_Q = \sum_{m=1}^{M} B_m(t - t_{z,Q}) \, \gamma_m \approx 0. \tag{4.5}$$

At any chosen time point $t_{z,q} \in \tau(t)$, exactly $d+1$ d-degree B-spline basis functions are non-zero. Thus, Eq. (4.5) can be expressed as,

$$\tilde{\gamma}_Q = \sum_{m=M-d}^{M} B_m(t - t_{z,Q})\, \gamma_m \approx 0.$$

Notice also that we implicitly assume that $\tilde{\gamma}_{Q+1}$ is zero. Following, the last $\tilde{\gamma}_Q$ coefficient can be forced to shrink towards zero by penalizing the last $d+1$ γ-coefficients by an $M \times M$ shrinkage matrix $\boldsymbol{K}_r$, a diagonal matrix with all elements equal to zero except the last $d+1$, which have value one, i.e. $\boldsymbol{K}_r = \text{diag}(\underset{M-(d+1)\times 1}{\mathbf{0}}, \underset{(d+1)\times 1}{\mathbf{1}})$.

Then, an extra regularization parameter controls the shrinking of the last basis coefficients. Large values of this parameter imply strong shrinkage of the last $\tilde{\gamma}_Q$ coefficient and decrease its estimated values.

Estimation and inference

The estimation of the model coefficients $\boldsymbol{\gamma}$ can be carried out by maximizing the penalized likelihood (Wood, 2011). In this case, for the Poisson GAMM, the model deviance $l(\boldsymbol{\gamma}) = \sum_{i_l=1}^{N_l} \{\delta_{i_l} \log(\lambda_{i_l}(t|\boldsymbol{z}_l(t), b_l)) - \lambda_{i_l}(t|\boldsymbol{z}_l(t), b_l)\}$, with $\delta_{i_l} \in \{0, 1\}$ the i-th event indicator of subject l, is penalized with a smoothing and a shrinkage matrix, $\boldsymbol{K}_d$ and $\boldsymbol{K}_r$, giving rise to:

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}' \left(\lambda_d \boldsymbol{K}_d + \lambda_r \boldsymbol{K}_r\right) \boldsymbol{\gamma}.$$

For the smoothing penalty matrix, we use second-order differences, i.e., $\boldsymbol{K}_d = \boldsymbol{D}_2' \boldsymbol{D}_2$, where $\boldsymbol{D}_2$ is the matrix representation of applying the Δ operator to α twice: $\Delta^2\alpha = \Delta(\alpha_j - \alpha_{j-1}) = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$. The smoothing parameter λ_d penalizes large differences in adjacent basis coefficients, while the regularization parameter λ_r shrinks the last basis coefficient.

Therefore, the coefficients can be estimated via penalized iteratively reweighted least squares P-IRLS (Marx and Eilers, 1998; Wood, 2017). The P-IRLS consists of iteratively updating the coefficient estimates until convergence is reached using numerical optimization methods of the restricted maximum likelihood. This is implemented in the **gam** function from **mgcv** (Wood, 2017) **R** package.

Software specification

The analyses of the simulation study and the application are coded in **R** version 4.2.2, on a 64-bit Unix platform (x86_64 linux-gnu) computer, as well as on a high-performance cluster system. The code to reproduce these analyses is available at: https://github.com/lzumeta/flex-mod-training-loads-recu-injuries. The package **msm**

1.6.9 (Jackson, 2011) is used to draw piece-wise exponential survival times, **pammtools** 0.5.8 (Bender and Scheipl, 2018) and **mgcv** 1.8-41 (Wood, 2023) to fit the models, **batchtools** 0.9.16 (Lang et al., 2023) to structure, write down and submit the simulation experiment in a convenient and reproducible fashion and the package **injurytools** 1.0.1 (Zumeta-Olaskoaga and Lee, 2023) to structure and explore the external training load data set.

4.3 Application to external training load data

Data

We apply the proposed model to observational injury data from an elite male football team that competed in LaLiga during the 2017-2018 and 2018-2019 seasons. A total of $L = 36$ players were followed up. To monitor players' performance and health status, external training load variables (Soligard et al., 2016) were registered through tracking devices, on each match and training session. These variables measure the physical exertion that the player has been exposed to. A total of 72 non-contact time-loss injuries occurred among 23 players (64%, 23/36) and 15 players (65%, 15/23) were reinjured during the follow-up, see Figure 4.2.

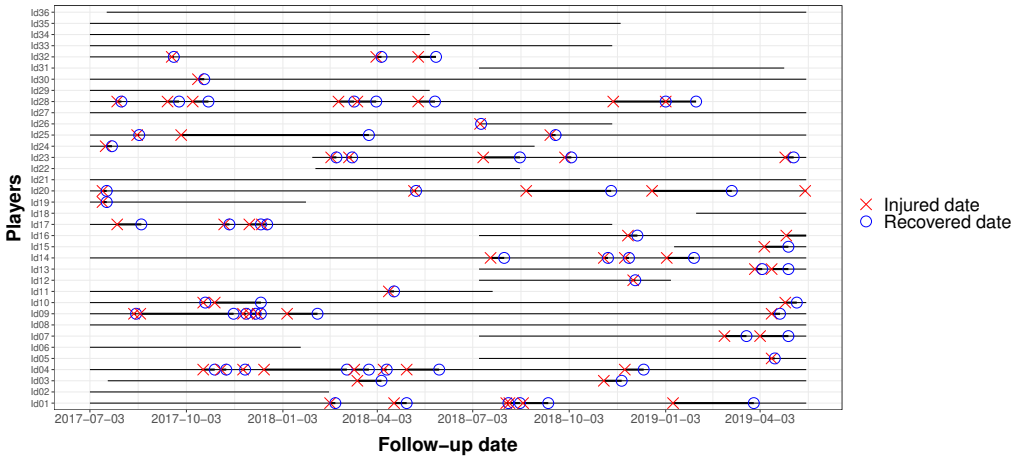


Figure 4.2: Timeline of the football players' follow-up period together with the injuries they sustained. The red cross indicates the exact injury date, the blue circle the recovery date and the bold black line the duration of the non-contact time-loss injury.

Our focus lies in the association between external training load and time-loss injuries (Fuller et al., 2006), namely, how the cumulative stress placed on a player from multiple training sessions and matches, over a period of time, affects his risk of a (subsequent) football injury.

Table 4.1: Descriptive characteristics of external training load data: summary statistics related to injury and exposure variables overall and by session type. Injury incidence and injury burden are reported per 1000 player-hours.

		Session type	
Injury-related variables	**Overall**	**Training**	**Match**
Injuries, n (%)	72	26 (36.1)	46 (63.9)
Days lost, n (%)	1595	591 (37.1)	1004 (62.9)
Total follow-up sessions, median (IQR)	220.5 (163 - 345)	198.5 (146 - 277)	37.5 (19 - 67)
Injury incidence, (95% CI)	4.07 (3.1 - 5)	1.47 (0.9 - 2)	2.6 (1.85 - 3.35)
Injury burden, (95% CI)	90.1 (85.6 - 94.5)	33.4 (30.4 - 36.1)	59.7 (53.2 - 60.2)
Exposure variables			
Average Speed (m/s), median (IQR)	3.8 (3.24 - 4.72)	3.71 (3.17 - 4.25)	6.46 (5.96 - 6.88)
Total Distance (m), median (IQR)	4689 (3586 - 6122)	4458 (3517 - 5525)	8552 (5138 - 10022)

n: number; IQR: interquartile range; 95% CI: 95% confidence interval

Table 4.1 shows the data's descriptive characteristics, overall and by session type. Injury incidence and injury burden are calculated as the number of injuries (I) per player exposure (ΔT) and the number of days lost due to injury ($\mathrm{n_d}$) per player exposure (Bahr et al., 2020), $I_\mathrm{r} = I/\Delta T$ and $I_\mathrm{br} = \mathrm{n_d}/\Delta T$, respectively. The first calculates the rate at which new injury occurs (likelihood), whereas the second how severe an injury is (consequences). We assume that the number of injuries I (and the number of days lost $\mathrm{n_d}$) throughout the total time under risk, ΔT, follows a Poisson distribution and compute the approximated confidence interval under large sample conditions. We specifically focus on the variables average speed per session (*Speed*) and total distance covered per session (*Dist*) as external training load variables that represent the intensity of each session. These metrics are chosen for their clear link to physical exertion and injury risk. Average speed reflects the sustained intensity throughout the session, while total distance indicates the overall workload. Both are crucial in evaluating the cumulative stress imparted on players, potentially leading to injuries when consistently high.

Modelling approach

We consider that the unit of the follow-up time t, as well as of the exposure time t_z, to be the n-th number of session (i.e., match and training sessions). The analysis time zero is defined as the first session (match or training session), from July 7, 2017, in which the player has taken part in the team. Players are followed until an injury occurs, or until they are transferred to another team, the end of the contract or the end of the study (May 18,

2019), whichever occurs first. Players fully recovered from an injury are followed again until one of the previously described events occurs, see Figure 4.2.

First, to explore the data, we fit a stratified hazards model. We stratify the baseline hazard by 'season' (X_1) and by 'injury number' (X_2) categorical variables, i.e. we allow the baseline hazards to be different for these groups.

- X_1 = 'season' (2-levels: 0 = '17-18' and 1 = '18-19').
- X_2 = 'injury number' (2-levels: 0 = 'at risk for a 1st-time injury', 1 = 'at risk for a subsequent injury').

Then, for player l in i-th stratum, we estimate this stratified baseline hazards model, with no covariates, in log-scale:

$$\log(\lambda_{i_l}(t|\boldsymbol{X}_l(t))) := \log(\lambda_{0,i}) = \beta_0 + X_1\beta_{0,1} + X_2\beta_{0,2} + \\ + f_0^{\text{17-18}}(t_j) + f_0^{\text{18-19}}(t_j) + f_0^{\text{first}}(t_j) + f_0^{\text{recurrent}}(t_j),\ \forall t \in (\kappa_{j-1}, \kappa_j] \tag{4.6}$$

Then, as our focus is on the association between external training load and time-loss injuries, we fit a PAMM with WCE cumulative effects model and a ridge penalty. Therefore, we consider that external training load is applied to the player over varying time periods and with varying magnitude by considering cumulative effects and we adjust for the type of session, whether training or match session, since it has been suggested as one of the primary risk factors Bahr and Holme (2003); Bahr et al. (2020); and account for subsequent injuries adding a random effect (Gaussian frailty). Thus, the log-hazard rate of player l of the fitted model is expressed as:

$$\log(\lambda(t|\boldsymbol{z}_l(t), b_l, i)) = \beta_0 + f_0(t_j) + z_l^{\text{type session}}(t_j)\beta_1 + g_1(\boldsymbol{z}_l^{\text{Speed}}(t), t) + g_2(\boldsymbol{z}_l^{\text{Dist}}(t), t) + b_l \\ \forall t \in (\kappa_{j-1}, \kappa_j],\ t_j := \kappa_j \text{ and } b_l \sim N(\boldsymbol{0}, \sigma_b), \tag{4.7}$$

where $\beta_0 + f_0(t_j)$ indicates the log-baseline hazard rate, $z_l^{\text{type session}}(t_j)$ the type of session undertaken by player l at t_j, g_1 and g_2 are non-linear time-varying effects of the training load variables and b_l a Gaussian random intercept term associated to player l. The cumulative effects, g_1 and g_2, are defined as $\int_{\tau_{\text{Speed}}(t)} h(t - t_z) z_l^{\text{Speed}}(t_z) dt_z$ and $\int_{\tau_{\text{Dist}}(t)} h(t - t_z) z_l^{\text{Dist}}(t_z) dt_z$, and each lag-lead window, $\tau_{\text{HRt}}(t)$ and $\tau_{\text{HSR}}(t)$, is chosen to be large enough to identify relevant past exposure effects by fitting a PAMM with a ridge penalization. All smooth terms are estimated using P-Splines with second-order difference penalties.

Importantly, we add a minimum lag time of one session to minimize confounding by indication bias (Signorello et al., 2002), i.e., we exclude the current session to have

an effect on the hazard of injury, ensuring $t > t_z$. The rationale behind this choice is that each session depends on the player's physical condition and, presumably, sessions in which a player was injured had lower intensity compared to sessions in which he had no complaints.

Results

The results of the estimated baseline hazards from the stratified PAMM model (4.6), i.e., a GAMM with two additive factor-smooth interaction terms (season and injury number), are shown in Figure 4.3. The fact of stratifying hazards captures the unobserved player-specific variability. It reveals that the risk of a subsequent injury is higher than the risk of experiencing a first-time injury in both seasons. Additionally, the differences between the risks of subsequent and first-time injuries are more pronounced at the beginning of the follow-up times and just after recovering from the first-time injury (i.e., the initial times of the risk sets). Moreover, the hazard rates for sustaining any injury (regardless of recurrence) are significantly higher in the 2017-2018 season compared to the 2018-2019 season.

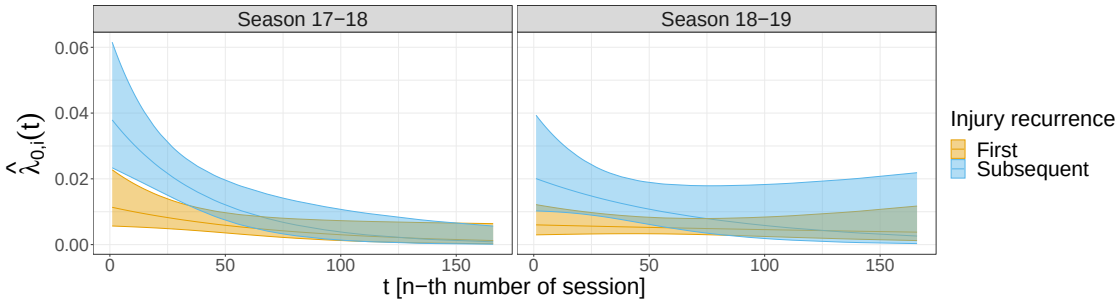


Figure 4.3: Estimated baseline hazards by a stratified PAM model with two-factor variables: injury recurrence, whether a first-time injury (orange) or a subsequent injury (blue); and season, 17-18 (left panel) or 18-19 (right panel).

On the other hand, the estimated cumulative effects in model (4.7) are computed considering that all recorded *Speed* and *Dist* values in the last 10 sessions prior to t (i.e., before three weeks approximately) could have an effect on the hazard of injury at time t, represented by the lag-lead windows $\tau_1(t) = \tau_2(t) = \{t_z : t > t_z \wedge t < t_z + 11\}$. We assume that these windows are large enough for the model to identify relevant past exposure effects.

The estimated partial effects corresponding to *Speed* and *Dist* training load variables, $\hat{h}_1(t - t_z)z_1(t_z)$ and $\hat{h}_2(t - t_z)z_2(t_z)$, are shown in Figure 4.4 (see also Figure C11 in Appendix C). The results suggest that no more than seven sessions in the past are of interest with regards to *Speed* and *Dist* variables cumulative effects. Both cumulative effects are

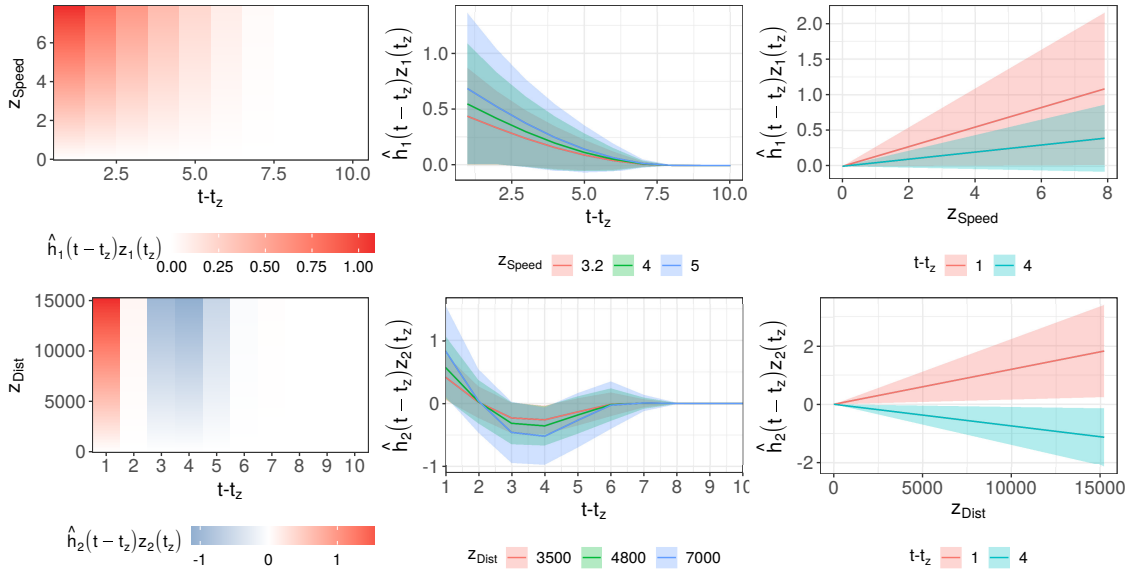


Figure 4.4: Top: Estimated partial effects surface (left-hand panel) and one-dimensional slices through the previous surface with respect to the covariate *Speed*, $z_1(t_z) = z_{\text{Speed}}(t_z) \in \{3.2, 4, 5\}$ (middle panel) and the latency $t - t_z \in \{1, 4\}$ (right panel) on the log-hazard scale. Bottom: the analogue for the covariate *Dist*, where the estimated one-dimensional partial effects are conditioned on the values $z_2(t_z) = z_{\text{Dist}}(t_z) \in \{3500, 4800, 7000\}$ and $t - t_z \in \{1, 4\}$.

estimated to have a non-linear decaying effect on the covariate $\boldsymbol{z}(t)$ with respect to latency and a linear effect on latency with respect to the covariate $\boldsymbol{z}(t)$. Regarding the estimated partial effects of both variables, the values contributing the most to the hazard are those most recently recorded, while the contribution of values recorded longer ago diminishes. Although there is not much difference in the trend, the greater the average speed and the total distance covered in recent sessions, the greater the impact on the resulting cumulative effect is. The Gaussian frailty term (random intercepts), which accounts for the correlation between subsequent injuries from the same player, is statistically significant (p-value < 0.01), with $\hat{\sigma}_b = 0.22$ as the estimated variance. Players who suffered more injuries (e.g., Id04, Id28) have a higher baseline hazard of injury, as observed in Figure 4.5, which shows the estimated smooth log-baseline hazard, $\hat{f}_0(t)$, together with the estimated player-specific smooth log-baseline hazard. Concerning the session type effect, match sessions have a higher risk of injury compared to training sessions ($\hat{\beta}_1 = 2.45$ and p-value $<$ 0.01). See also Table C5 and Figure C13 in Appendix C where the estimated linear and

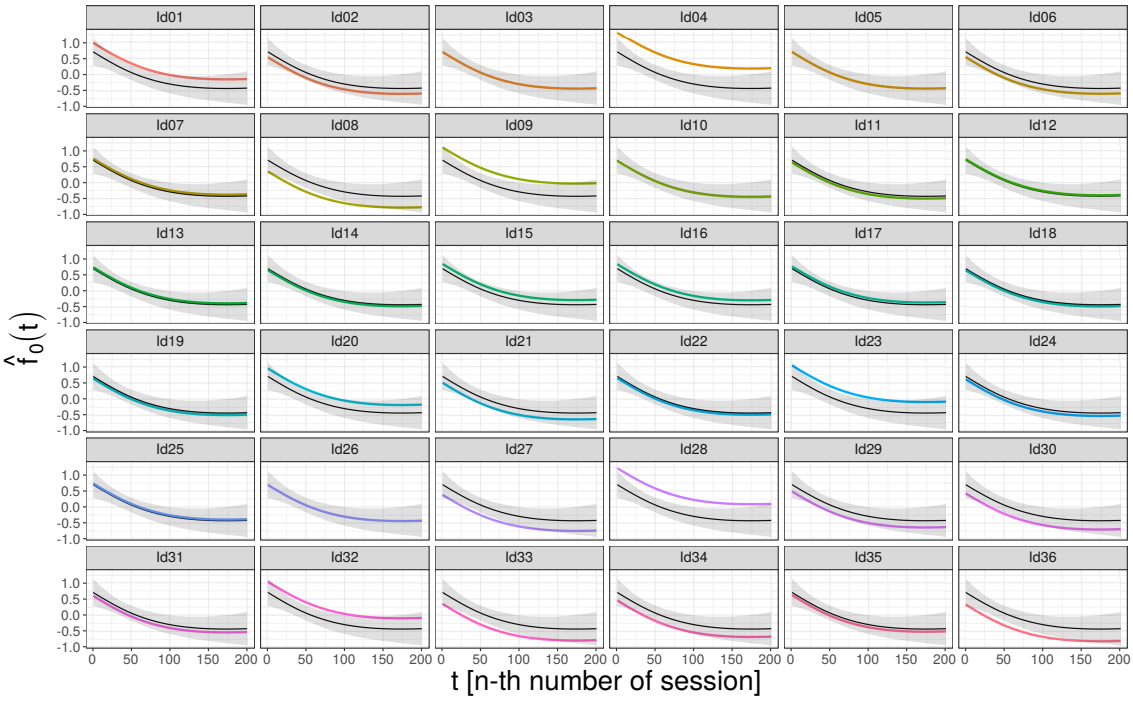


Figure 4.5: Estimated team smooth baseline hazard (in black) and player-specific smooth baseline hazard (coloured line in each panel), together with confidence intervals (grey shadow) of the team's smooth baseline hazard estimate, in log scale.

non-linear effects are presented.

4.4 Simulation study

We conduct extensive simulation studies to evaluate the proposed models and to investigate their properties. In particular, we aim to (i) assess the ability of the model to simultaneously estimate both, flexible WCE-type effects and heterogeneity resulting from recurrent events; and (ii) study the implementation of penalties on the basis coefficients to select the maximum length of the time window in which past exposures are cumulatively associated with the hazard.

4.4.1 Data generation

We draw survival times from the piece-wise exponential distribution. Let n_l be the maximum event number that individual l has been at risk for and $l = 1, \ldots, L$. Then, it suffices to specify a vector of piece-wise constant hazards $\boldsymbol{\lambda} = (\lambda_{1_1}, \lambda_{2_1}, \ldots, \lambda_{n_1}, \lambda_{1_2}, \lambda_{2_2}, \ldots, \lambda_{n_2}, \ldots, \lambda_{1_L}, \lambda_{2_L}, \ldots, \lambda_{n_L})$, in intervals defined by $J + 1$

cut-points, i.e., by the vector of interval borders $\boldsymbol{\kappa} = (0 = \kappa_0, \ldots, \kappa_J = t_{\max})$. That is, λ_{i_l} is composed of $(\lambda_{i_l 1}, \lambda_{i_l 2}, \ldots, \lambda_{i_l J})$, where each element $\lambda_{i_l j}$ is the hazard rate of i-th event for individual l in the interval j, $i_l = 1, \ldots, n_l$ and $j = 1, \ldots, J$; and can be defined through a function of time t, current and past exposure covariates $\boldsymbol{z}(t)$ and a random effect b_l, i.e. $\lambda_{i_l,j}(t|\boldsymbol{z}(t), b) = f(t, \boldsymbol{z}(t), b) = \exp\Big(\textit{const.} + f_0(t) + \int_{\{t_z : t \geq t_z\}} h(t - t_z) z(t_z) dt + b_l\Big)$, evaluated at time $t = \kappa_j$.

Then, we draw recurrent survival times from the piece-wise exponential distribution (PEXP), $t \sim \text{PEXP}(\boldsymbol{\lambda}, \boldsymbol{\kappa})$, for which the algorithm is outlined in Table C1 in Appendix C. The hazard rate vector $\boldsymbol{\lambda}$ is defined based on the simulation settings described in the following section. All further details on data generation are provided in section C.1 in Appendix C.

4.4.2 Scenarios and parameter settings

We simulate $N_{\text{sim}} = 500$ times a cohort of $L = 500$ individuals with exposures recorded at $t_{z,1} = 1, t_{z,2} = 2, \ldots, t_{z,Q=40} = 40$ days before the time at which we model the hazard, $\boldsymbol{z}_l(t) = (z_l(t_{z,1}), z_l(t_{z,2}), \ \ldots, z_l(t_{z,Q}))$, and draw survival times from the piece-wise exponential distribution under four different true weight functions, $h(t - t_z)$, (a) *exponential decay*, (b) *bi-linear* (c) *early peak* and (d) *inverted U* shapes, each defined over a $[0, t_{z,Q}]$ interval (see the black curves in Figure 4.6); and under three different levels of heterogeneity between recurrent events, $\sigma_b \in \{0.05, 0.5, 1\}$, indicating very low heterogeneity, low heterogeneity and high heterogeneity, respectively (see also Figure C1 in Appendix C).

We then fit three different PAMMs with WCE-type cumulative effects: a model with no constraint (*Uncons.*), adding a constraint (*Constr.*) and adding a ridge penalty (*Ridge*). The performance of the models is evaluated by graphical inspection of the estimated $\hat{h}(t - t_z)$ function in comparison to the true simulated $h(t - t_z)$ function; the accuracy of these WCE-type cumulative effects estimates are also evaluated via the mean RMSE, i.e., $\overline{\text{RMSE}}$, over all simulation runs, as:

$$\overline{\text{RMSE}} = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \sqrt{\frac{1}{N_{t_z}} \sum_{t - t_z = 0}^{40} \Big(h(t - t_z) - \hat{h}(t - t_z)^{(n)}\Big)^2},$$

where $N_{t_z} = 41$, since $t - t_z = \{0, 1, 2, \ldots, 40\}$ takes 40 +1 number of different values.

We assess the accuracy of the standard deviation of the random effects, $\hat{\sigma}_b$, through:

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \Big(\sigma_b - \hat{\sigma}_b^{(n)}\Big)^2}.$$

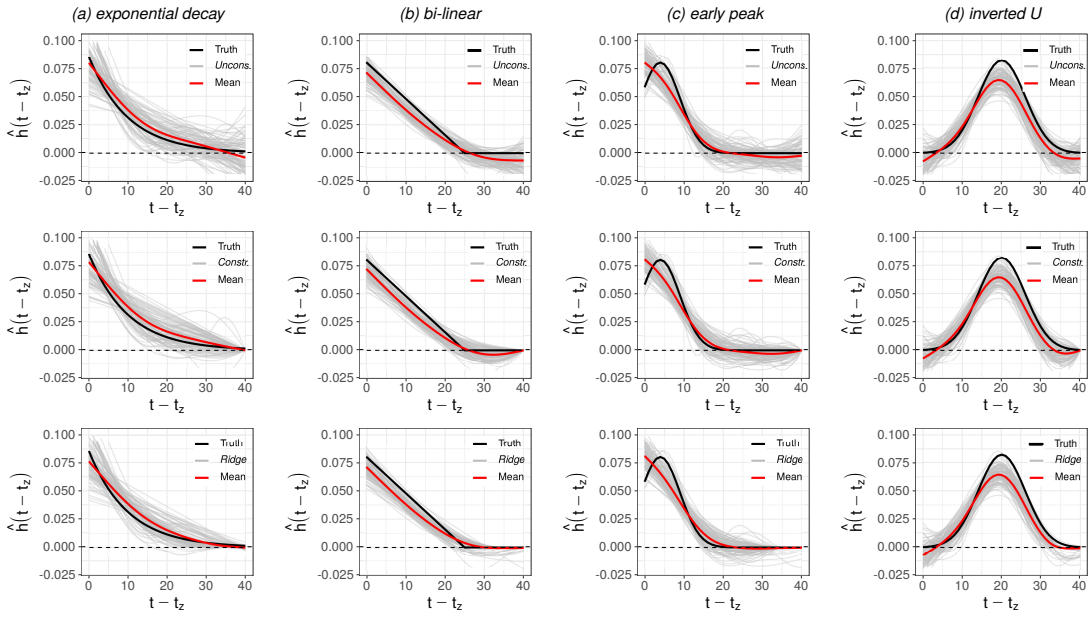


Figure 4.6: True vs. fitted partial effect weight function $h(t-t_z)$ for scenario $\sigma_b = 1$. Rows: *Uncons.* model (top row), *Constr.* model (middle row) and *Ridge* model (bottom row). Columns: (a) *exponential decay*, (b) *bi-linear*, (c) *early peak* and (d) *inverted U*. True shapes used for simulation are depicted as solid black lines and the mean (point-wise averages) of all simulation runs are depicted in solid red lines. A random sample of 100 individual estimated weight functions are shown as grey curves.

We also evaluate the rate at which the estimated confidence interval of WCE-type cumulative effects estimates contains the true estimand $h(t-t_z)$, computing the mean coverage at the $1-\alpha$ confidence level, i.e. $\overline{\text{Coverage}}_\alpha$, as:

$$\overline{\text{Coverage}}_\alpha = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \left[\frac{1}{N_{t_z}} \sum_{t-t_z=0}^{40} \mathbb{I}\left(h(t-t_z) \in \left[\hat{h}(t-t_z)^{(n)} \mp \zeta_{1-\alpha/2} \hat{\sigma}_{\hat{h}}^{(n)} \right] \right) \right],$$

where ζ_q is the q-quantile of the standard normal distribution and $\hat{\sigma}_{\hat{h}}$ the standard error of the estimated $\hat{h}(t-t_z)$.

4.4.3 Results

The results regarding the estimation of the weight function $\hat{h}(t-t_z)$ for scenarios with $\sigma_b = 1$ and true weight functions (a)-(d) are shown in Figure 4.6 (the rest of the scenarios are shown in Figures C2-C7 in Appendix C). In general, the model estimates effectively capture the underlying weight function. Models incorporating an additional penalty (middle and bottom panels), referred to as *Constr.* and *Ridge* models, perform

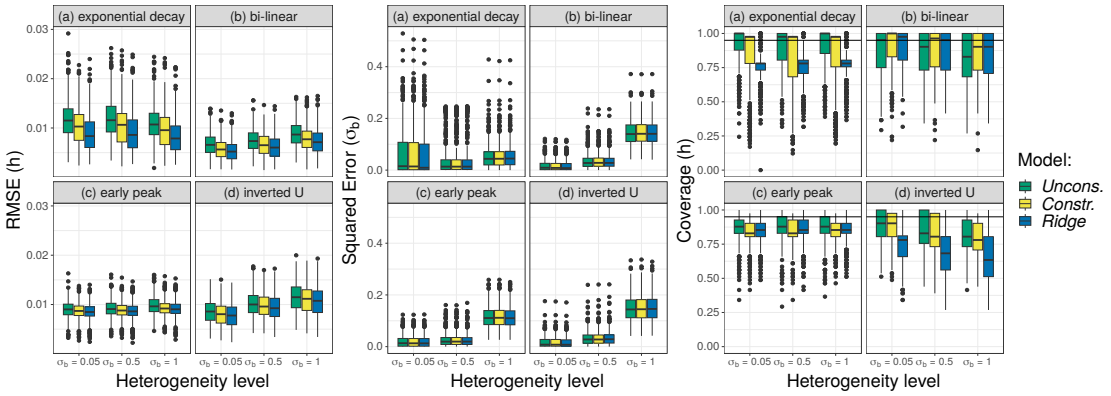


Figure 4.7: Distribution of the RMSE (left), the squared error of σ_b (middle) and the 95% point-wise coverage (right) across all simulation settings ($N_{\text{sim}} = 500$).

better for scenarios in which the exposures that occurred relatively long ago (e.g. from the 20th lag on) have little impact on the risk, as observed in shapes (b) and (c). Among the settings considered, the most accurate estimation of $h(t - t_z)$ is obtained for shape (b) *bi-linear* and a ridge penalty, according to RMSE and 95% coverage. This model has a mean RMSE of 0.007 and a mean coverage of 84.9% (refer to Table C2 in Appendix C).

Figure 4.7 shows boxplots of the distribution of the RMSE, the 95% point-wise coverage (across all time points), and the squared error of σ_b across all simulation settings. The shape of the true weight function is independent of the estimation of the standard deviation of the random effect, σ_b. Besides, the estimates are less accurate for higher values of σ_b. Note that the 95% coverage of $h(t-t_z)$ for shapes (a), (c), and (d) shows underfitting, specifically in models where the weight function is penalized, due to the form of the true weight function considered and the way the point-wise coverage is calculated.

4.5 Discussion

Contributions and practical application

We extended and assessed the PAMM model class to the context of recurrent events with time-dependent covariates, modelled as WCE-type cumulative effects. By introducing a ridge penalty to diminish the influence of past exposures registered long ago, we presented a method to determine a relevant time window based on data. Lastly, motivated by the research question regarding the association between external training load and (subsequent) time-loss injuries, we have applied the proposed methodology to the sports medicine context.

Simulations indicate that PAMM with ridge penalization is the method that yields the most accurate estimates for the partial effects, $\hat{h}(t, t_z, z(t_z)) = \hat{h}(t-t_z)z(t_z)$. The additional ridge penalization of the weight function enables us to identify the relevant window $\tau(t)$ at which past exposures cumulatively affect the hazard at time t, as also demonstrated in our application on football injury data. We propose using wide time windows to properly determine the exposure time at which, from that time on, the estimated effects are close to zero.

Limitations

Simulation studies indicate that the model can recover a number of clinically plausible shapes for the true weight function under various levels of heterogeneity. Without prior knowledge about the form of association for time-varying exposures, the model proved to capture well a variety of shapes, estimating them from the data via P-splines. However, for non-smooth effect shapes, such as piece-wise constant or bi-linear, alternative methods like adaptive splines (Friedman, 1991) or treed distributed lag non-linear models (TDLNM, Mork and Wilson 2022), might be of interest.

In addition, future research should consider evaluating the impact of the number of events per subject –kept fixed in our simulation study– on the model performance, as well as to explore distributions other than Gaussian, for example, Gamma distributed random effects, which are popular in the context of survival analysis (Balan and Putter, 2020).

Alternative approaches and further work

From a practical point of view, the presented modelling framework provides a suitable approach to flexibly model training load exposures and analyze their effects on subsequent football injuries, with respect to alternative measures of training load exposures commonly used in the literature (refer to Table C6 in Appendix C). For example, the widely known and used acute chronic workload ratio (ACWR, Hulin et al. 2014), and its variants (Lolli et al., 2019; Wang et al., 2020), limit to summarize past observations into a predefined unweighted metric, through a ratio of two rolling averages –last 7 days (acute load) over the last 28 days (chronic load). The same applies to the exponentially weighted moving averages (EWMA, Williams et al. 2017) metric, suggested as an alternative measure of rolling averages. While EWMA more accurately accounts for the decaying nature of fitness and fatigue effects over time compared to rolling averages, both may fall short in accurately reflecting various changes in past training exposures, as well as considering prespecified time windows that could either be superfluous or insufficient. Our method, in contrast, estimates cumulative effects and relevant time windows based on the data rather than predetermined metrics. Future research should assess the application

of negative binomial and zero-inflated models within the PAMM framework to address overdispersion and the excess of zeros issues (i.e., the low number of injuries).

By highlighting the potential value of PAMMs with WCE effects in assessing recurrent events in sports medicine, our work contributes to enriching the existing literature. We believe that this methodology would help in designing and comparing personalized training plans with insights into the risk of injury.

Chapter 5

Software development for sports injury data

Contributing article

Zumeta-Olaskoaga L., Lee D.-J. (2023). injurytools: A Toolkit for Sports Injury Data Analysis. https://cran.r-project.org/package=injurytools

Code repository

`https://github.com/lzumeta/injurytools`

In the previous chapters, we have highlighted specific code repositories that were purposefully created for each of the research objectives. The statistical open-source software **R** (R Core Team, 2023) has been employed as a comprehensive tool for implementing and evaluating statistical models, processing and organizing the data, and presenting results in both visual and tabular formats. All these computational developments have been made publicly available to ensure full reproducibility.

In the following, we outline the current landscape of existing **R** packages in the field of sports science and emphasize the necessity for a dedicated **R** tool for sports injury data. Subsequently, we present the **R** package we have developed, called **injurytools** (Zumeta-Olaskoaga and Lee, 2023). The package offers general and standardised routines that simplify the workflow of sports injury data analysis and is intended to be used in practice.

5.1 Statement of need

A systematic review by Casals et al. (2023), focusing on sports-related packages in the Comprehensive R Archive Network (CRAN) repository (https://cran.r-project.org/), has emphasized a growing trend in the development of **R** packages, along with books and tutorials within the **R** ecosystem, specifically designed for various sports environments. Of the eighty-one packages that met their eligibility criteria, as of 18 February 2021, fifty (61.7%) were categorized under the "*sports performance analysis*" category. The predominant functionality was web scraping ($n = 43$, 53.1%), with basketball ($n = 14$, 17.3%) being the most represented sport, closely followed by soccer (referred to as football in this dissertation, $n = 12$, 14.8%). Furthermore, the Sports Analytics CRAN Task View (https://CRAN.R-project.org/view=SportsAnalytics) provides a comprehensive list of packages in sports analytics.

There is a notable scarcity of **R** packages tailored for sports medicine. According to Casals et al. (2023), only a small fraction (four out of eighty-one, or 4.9%) of the packages are related to the "*athlete heatlh*" category. These existing packages are primarily focused on injury categorization or calculating Injury Severity Scores (ISS) based on the International Classification of Diseases (ICD) codes, potentially serving as alternatives to manual injury severity scoring. However, to our knowledge, there is currently no **R** package that specifically addresses the comprehensive needs of sports medicine, particularly in sports injury data management. To fill this gap, we have developed the **injurytools R** package following the guidelines of the most extensive resource on how to generate an **R** package (Wickham and Bryan, 2023).

As an additional remark, the development of the **injurytools R** package originated from our efforts to organize and standardize the code that we found ourselves using repeatedly. By consolidating this code into a comprehensive package, accompanied by detailed documentation, our goal is to enhance its broader dissemination and ease of use within the sports medicine community.

5.2 The injurytools R package

injurytools is a user-friendly **R** package developed for the field of sports medicine to facilitate the data analysis workflow and automate common tasks typically encountered in handling sports injury data.

The package is structured in four main blocks that include convenience functions de-

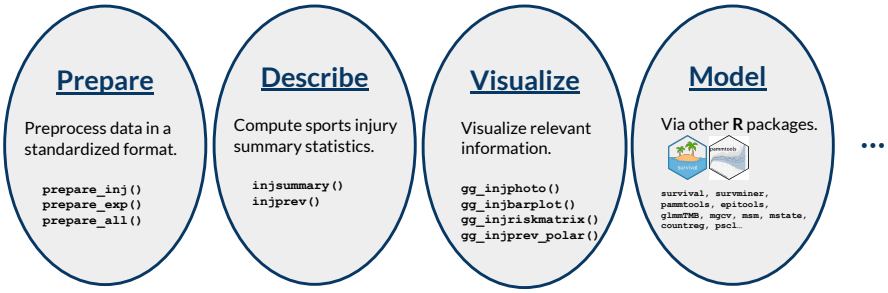


Figure 5.1: The sports injury data analysis workflow in the **injurytools R** package.

voted to (a) data preparation, (b) estimation of epidemiological measures, (c) data visualization and (d) data modelling (see Figure 5.1). Above all, all these functionalities are meant to offer standardized procedures for the specific field of sports injuries. On the **R** CRAN, there are packages related to some of the **injurytools** package features, such as the **Epi** (Carstensen et al., 2023), **epiR** (Stevenson et al., 2023), **epitools** (Omidpanah et al., 2020) packages, which contain functions for epidemiological data analysis. However, they are more broad in nature and do not cover the particular needs of the sports injury data analysis.

Furthermore, **injurytools** is a fully documented package that includes a companion website, available at https://lzumeta.github.io/injurytools/, created using the **R** package **pkgdown** (see Wickham et al., 2022). This website provides detailed function references and their corresponding help files, along with vignettes that demonstrate and guide users through the package's application.

5.2.1 Summary

The **injurytools** package's most recent version on CRAN is v.1.0.3, which was released on 14 November 2023. Table 5.1 contains a list of all the functions available in the package, as well as the data sets that come with it to exemplify those functions. For more details, check out each function's help files (`help(function)` or `?function` in **R**) and see the reference page on https://lzumeta.github.io/injurytools/reference/index.html.

Following, we showcase the main functionalities of the package on the included data set. These are the injury data for Liverpool Football Club's male first-team players over two consecutive seasons, 2017/2018 and 2018/2019, scrapped from https://www.transfermarkt.com/ website and comprise information on these player's

Table 5.1: Function reference in the **injurytools** package.

Function	Description
Prepare sports injury data	
`prepare_inj()`	Prepare injury data in a standardized format
`prepare_exp()`	Prepare exposure data in a standardized format
`prepare_all()`	Create the final and required data frame (`injd` object)
Compute injury summary statistics	
`injsummary()`	Estimate injury summary statistics
`injprev()`	Calculate injury prevalence
Visualize sports injury data	
`gg_injphoto()`	Plot injuries over the follow-up period
`gg_injbarplot()`	Plot player's injury incidence/burden ranking
`gg_injprev_polar()`	Plot polar area diagrams showing players' prevalence
`gg_injriskmatrix()`	Plot risk matrices
Data sets	
`raw_df_exposures`	Minimal example of exposure data
`raw_df_injuries`	Minimal example of injury data
`injd`	Example of an `injd` object

match exposures and injuries they sustained during the matches[4].

5.2.2 Usage

When using **injurytools** package, the very first step every user has to follow is to prepare data, to create a standardized data frame. Let us illustrate the functions intended to facilitate this data preprocessing step and what the final data set is like.

(*a*) *Data preparation*

Data can be collected in several ways and by several means. A conventional approach is to collect and store data as events occur. In the context of sports medicine, it is common to store injury records on one hand, and in a separate table, data related to training and competitions/matches (*exposure time* among others). Following this, we consider that the

[4]These data sets are provided for illustrative purposes. We warn that they might not be accurate and could potentially include discrepancies or incomplete information compared to what actually occurred.

user has the raw data in two separate data sets that we call **injury** and **exposure** data, respectively[5]. See Figure 5.2.

Exposure data

player	date	time_expo	...
Olivia	2022-07-03	70	
Olivia	2022-07-04	84	
Olivia	2022-07-06	75	
...	...	...	
Mia	2022-07-03	72	
Mia	2022-07-04	80	
...	...	...	

Injury data

player	date_injured	date_recovered	...
Mia	2022-09-09	2022-09-19	
Olivia	2022-11-01	2022-11-28	
Mia	2022-11-12	2022-11-16	
...	...	...	

Figure 5.2: Illustration of minimal data required.

Thus the early task is to tidy up these two sources of data. To this end, the functions provided by **injurytools** involve:

1. setting **exposure** and **injury** data in a standardized format and
2. integrating both sources of data into an adequate data structure.

We consider the `raw_df_injuries` and `raw_df_exposures` data sets available from the **injurytools** package and we standardize the **key column names** such as the player (subject) identifier, the dates of injury and recovery (if any), the training/match/season date and the amount of time of exposure; and set them proper names and formats by means of `prepare_inj()` and `prepare_exp()`:

```
df_injuries <- prepare_inj(df_injuries0   = raw_df_injuries,
                           player         = "player_name",
                           date_injured   = "from",
                           date_recovered = "until")
```

```
df_exposures <- prepare_exp(df_exposures0 = raw_df_exposures,
                            player        = "player_name",
                            date          = "year",
                            time_expo     = "minutes_played")
```

[5]If the data are not recorded this way, we suggest splitting both information into separate tables and then following the same functions provided by the package.

Then, we apply `prepare_all()` to the data sets tidied up above. It is important to specify the unit of exposure, i.e. the `exp_unit` argument (see `?prepare_all`):

```
injd <-  prepare_all(data_exposures = df_exposures,
                     data_injuries  = df_injuries,
                     exp_unit = "matches_minutes")
head(injd)
```

```
#> # A tibble: 6 × 19
#>   player t0         tf         date_injured date_recovered tstart     tstop
#>   <fct>  <date>     <date>     <date>       <date>         <date>     <date>
#> 1 adam-... 2017-07-01 2019-06-30 2017-07-31   2017-11-25     2017-07-01 2017-07-31
#> 2 adam-... 2017-07-01 2019-06-30 2018-03-31   2018-05-13     2017-11-25 2018-03-31
#> 3 adam-... 2017-07-01 2019-06-30 2018-09-04   2018-10-19     2018-05-13 2018-09-04
#> 4 adam-... 2017-07-01 2019-06-30 2018-11-09   2018-12-04     2018-10-19 2018-11-09
#> 5 adam-... 2017-07-01 2019-06-30 2019-01-06   2019-01-18     2018-12-04 2019-01-06
#> 6 adam-... 2017-07-01 2019-06-30 2019-04-01   2019-05-31     2019-01-18 2019-04-01
#> #  12 more variables: tstart_minPlay <dbl>, tstop_minPlay <dbl>, status <dbl>,
#> #   enum <dbl>, days_lost <dbl>, player_id <fct>, season <fct>,
#> #   games_lost <dbl>, injury <chr>, injury_acl <fct>, injury_type <fct>,
#> #   injury_severity <fct>
```

This last step integrates both the standardized injury and exposure data sets, converting them into an `injd` **S3** object with a structure suitable for further statistical analyses. The resulting data set will always include the columns listed below (standardized columns or those created by the function), as well as additional (optional) sports-related variables:

- `player`: the player identifier.
- `t0` and `tf`: the follow-up period of the corresponding player, i.e. the player's first and last dates observed (same value for each `player`).
- `date_injured` and `date_recovered`: the dates of injury and recovery of the corresponding observation (if any). Otherwise `NA`.
- `tstart` and `tstop`: the beginning and ending dates of the corresponding interval in which the observation has been at risk of injury.
- `tstart_x`and `tstop_x`: the beginning and ending times of the corresponding interval in which the observation has been at risk of injury (it depends on the unit of exposure time specified).
- `status`: the injury (event) indicator.

- `enum`: an integer indicating the recurrence number, i.e. the k-th injury (event), at which the observation is at risk.
- `days_lost`: the number of days lost due to injury occurred at `tstop`/`date_injured` (if any; otherwise 0). Namely, `date_recovered` - `date_injured` in days.

For example, the first row of `injd` corresponds to the player Adam Lallana, to the risk set that starts on 2017-07-01 and ends on 2017-07-31, after having played 236 minutes, when he got firstly (`enum` = 1) injured (`status` = 1). The second row corresponds to the risk set of being injured by a second injury (`enum` = 2), the set starts when he fully recovered in 2017-11-23 and finishes when he suffered another hamstring injury. These final data set is an **R** object of class `injd`,

```
class(injd)
```

```
#> [1] "injd"       "tbl_df"     "tbl"        "data.frame"
```

and have the following attributes:

```
str(injd, 1)
```

```
#> injd [108 × 19] (S3: injd/tbl_df/tbl/data.frame)
#>  - attr(*, "unit_exposure")= chr "matches_minutes"
#>  - attr(*, "follow_up")= tibble [28 × 3] (S3: tbl_df/tbl/data.frame)
#>  - attr(*, "data_exposures")='data.frame':   42 obs. of  19 variables:
#>  - attr(*, "data_injuries")= tibble [82 × 11] (S3: tbl_df/tbl/data.frame)
```

- `unit_exposure`: a character indicating the unit of exposure time used in this object.
- `follow_up`: a data frame consisting of one row per player with their first and last dates observed (`t0` and `tf` columns).
- `data_exposures`: the preprocessed exposure data frame.
- `data_injuries`: the preprocessed injury data frame.

To extract one of the attributes, for example, `unit_exposure`, type:

```
attr(injd, "unit_exposure")
```

```
#> [1] "matches_minutes"
```

(*b*) *Estimation of epidemiological measures*

Now, the preprocessed data are passed to `injsummary()` to calculate injury summary statistics:

```
injds <- injsummary(injd)
```

What `injsummary()` returns as its output is a list of two elements:

```
str(injds, 1)
```

```
#> [1] "matches_minutes"#> List of 2
#>  $ playerwise: tibble [28 × 9] (S3: tbl_df/tbl/data.frame)
#>  $ overall   : tibble [1 × 14] (S3: tbl_df/tbl/data.frame)
#>  - attr(*, "class")= chr [1:2] "injds" "list"
#>  - attr(*, "unit_exposure")= chr "matches_minutes"
#>  - attr(*, "unit_timerisk")= chr "100 player-match"
#>  - attr(*, "conf_level")= num 0.95
```

that is, the `injds` object consists of two data frames (two tables), which can be accessed by typing:

```
# the 'playerwise' data frame
injds[[1]]   ## or injds[["playerwise"]]
# the 'overall' data frame
injds[[2]] ## or injd[["overall"]]
```

The user can easily transform these objects and make them publication-ready. For instance, the data frame resulting from `injds[[2]]` is split into two tables and shown in Tables 5.2 and 5.3 in a LaTeX-styled form.

Table 5.2: The formatted output from `injds[[2]]` showing injury summary statistics.

	N injuries	**N days lost**	**Mean days lost**	**Median days lost**	**IQR days lost**	**Total exposure**	**Incidence**	**Burden**
TOTAL	82	2049	18.97	7.5	1-20.25	74690	9.88	246.9

All in all, `injsummary()` can be used to compute injury summary statistics either on a player-wise or team-wise basis. Additionally, the measures can be estimated for each type of injury by specifying the argument `var_type_injury`, which should indicate the name of the column, based on which injury summary statistics are computed. In this case,

Table 5.3: The formatted output from `injds[[2]]` showing injury summary statistics (*continuation*).

	Incidence	**95% CI for I_r**	**Burden**	**95% CI for I_{br}**
TOTAL	9.88	[7.7, 12]	246.9	[236.2, 257.6]

Tables 5.2 and 5.3 show the Liverpool FC's male first-team's number of injuries, number of days lost, the median number of days lost, total exposure time (minutes in matches), incidence and burden, along with their 95% confidence intervals, during 2017-2019.

Note that to provide numbers that are easy to interpret and to avoid small decimals, injury incidence and injury burden are reported "per 100 player-match exposure". As in this example exposure time is **minutes played in matches**, we multiply the rates by 90*100 (i.e. 90 minutes lasts a football match). The reported incidence rate is estimated as $\hat{I}_r = \frac{82}{74690} \times 90 \times 100$.

To calculate the injury prevalence and the proportions of injury-free players on a season basis, we use `injprev()` function:

```
prev_table1 <- injprev(injd, by = "season") ## by = "monthly"
prev_table1
```

```
#> # A tibble: 4 × 5
#>   season           type_injury     n n_player  prop
#>   <fct>            <fct>       <int>    <int> <dbl>
#> 1 season 2017/2018 Available       7       23  30.4
#> 2 season 2017/2018 Injured        16       23  69.6
#> 3 season 2018/2019 Available       2       19  10.5
#> 4 season 2018/2019 Injured        17       19  89.5
```

Overall, there were more injured players in the 2018/2019 season than in the previous season.

(*c*) *Data visualization*

We now keep on exploring the data graphically. **injurytools** offers modern visualization techniques. For example, to obtain a comprehensive picture of injury data, we just type `gg_injphoto()`:

```
gg_injphoto(injd,
            title   = "Overview of injuries:\nLiverpool FC 1st male team during
            2017-2018 and 2018-2019 seasons",
```

```
            by_date = "2 month",
            fix     = TRUE)
```

The outcome is shown in Figure 5.3, which gives us an overview of the injuries sustained by each player during the follow-up. Each player's timeline is depicted horizontally: the red cross indicates the exact injury date, the blue circle the recovery date and the bold black line indicates the duration of the injury (time-loss).

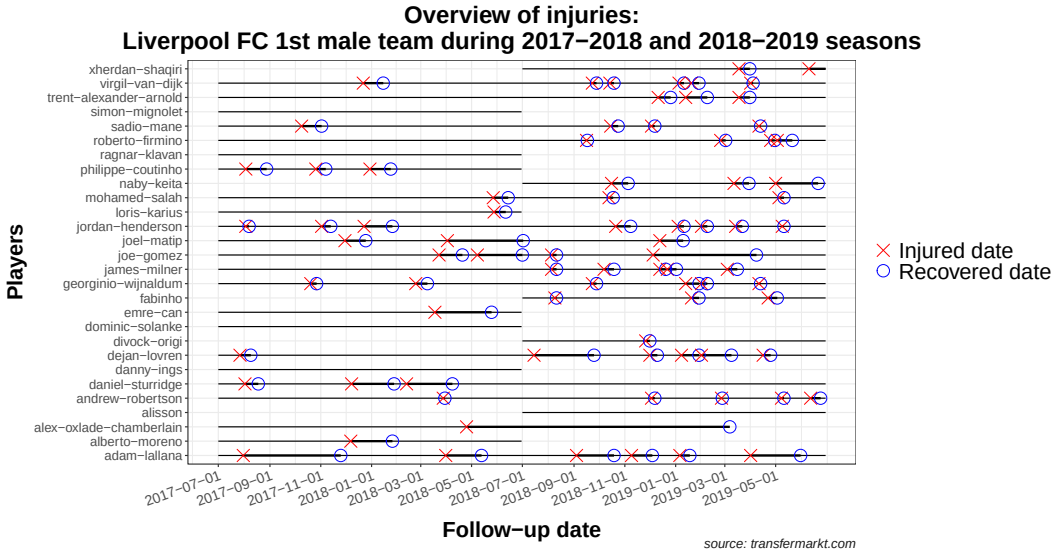


Figure 5.3: The output from the `gg_injphoto()` function applied to the included Liverpool FC players' injury data.

We now count how many injuries (red crosses in Figure 5.3) occurred and how severe they were (length of the thick black line), by type of injury. First, we use the `injsummary()` function and then, we plot the injury incidence vs. the mean time-loss graph, i.e. the so-called injury risk matrix (Fuller, 2018), through the `gg_injriskmatrix()` function:

```
# warnings set to FALSE
injds        <- injsummary(injd)
injds_perinj <- injsummary(injd, var_type_injury = "injury_type")
# injds
```

```
# warnings set to FALSE
gg_injriskmatrix(injds_perinj,
                 var_type_injury = "injury_type",
                 title = "Risk matrix")
```

Table 5.4 shows the information stored in the `injds_perinj` object (i.e. the `injds_perinjd[[2]]` data frame formatted), which among others include the injury incidence

Table 5.4: The formatted `injds_perinj` object (of `injds` class). Injury incidence and injury burden are reported as 100 player-matches.

Type of injury	N injuries	N days lost	Total expo	Incidence (95% CI)	Burden (95% CI)
Bone	11	173	74690	1.33 (0.54, 2.11)	20.85 (17.74, 23.95)
Concussion	16	213	74690	1.93 (0.98, 2.87)	25.67 (22.22, 29.11)
Ligament	9	596	74690	1.08 (0.38, 1.79)	71.82 (66.05, 77.58)
Muscle	25	735	74690	3.01 (1.83, 4.19)	88.57 (82.16, 94.97)
Unknown	21	332	74690	2.53 (1.45, 3.61)	40.01 (35.7, 44.31)

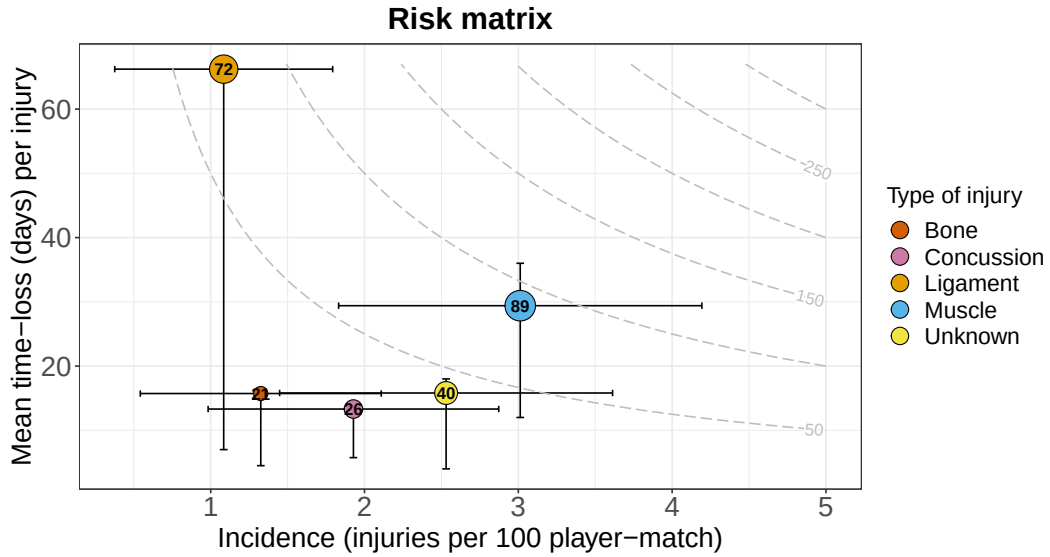


Figure 5.4: The output from the `gg_injriskmatrix()` function applied to the included Liverpool FC players' injury data.

and injury burden of the team during the follow-up by injury type.

This `injds_perinj` object is used to plot the injury risk matrix, displayed in Figure 5.4. The point estimate of injury incidence together with its confidence interval is plotted against the mean time-loss per injury together with ± IQR (days). The number shown inside the point and the point size itself, report the injury burden (days lost per player-exposure time), the bigger the size the greater the burden. Contour lines join the values for which the product between the x- and y-axes is the same, which results to be the injury burden[6].

As emphasized in Chapter 2, it is essential to report and evaluate injury incidence

[6]Injury incidence multiplied by the mean time-loss per injury results in injury burden.

(likelihood) and injury burden (severity) together rather than separately. To that end, the injury risk matrix is very useful.

After doing some exploratory data analysis and gaining a general sense of the data, we can dig a little deeper into the data and answer some of the questions that naturally arise. Thus, let's briefly compare injuries that occurred in the 2017/2018 season vs. the 2018/2019 season.

We prepare two `injd` objects:

```
# warnings set to FALSE
injd1 <- cut_injd(injd, datef = 2017)
injd2 <- cut_injd(injd, date0 = 2018)
```

```
## Plot just for checking whether cut_injd() worked well
p1 <- gg_injphoto(injd1, fix = TRUE, by_date = "3 months")
p2 <- gg_injphoto(injd2, fix = TRUE, by_date = "3 months")
grid.arrange(p1, p2, ncol = 2)
```

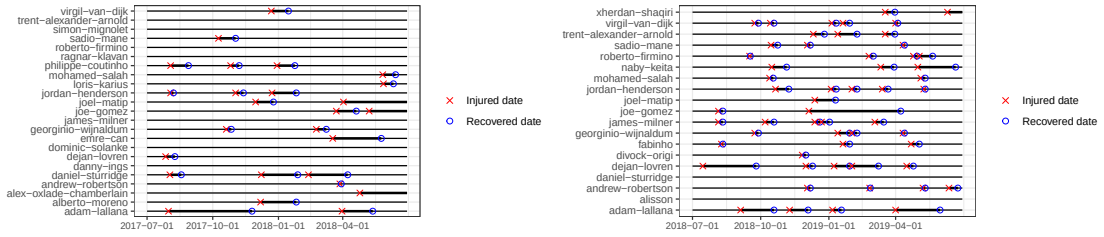


Figure 5.5: Overview of injuries that occurred during the 2017/2018 season (left) and 2018/2019 season (right). Output from the `gg_injphoto()` function.

Table 5.5: Injury summary statistics for each season. Injury incidence and injury burden are reported per 100 player-matches. Output from `injsummary()`.

Season	N injuries	N days lost	Total expo	Incidence (95% CI)	Burden (95% CI)
2017-2018	26	1141	37364	6.26 (3.86, 8.67)	274.84 (258.89, 290.78)
2018-2019	56	908	37326	13.5 (9.97, 17.04)	218.94 (204.7, 233.18)

Table 5.5, numerically, and Figure 5.5, visually, outline the comparison between both seasons.

Who were the most injured players? And the most severely affected?

```
injds1 <- injsummary(injd1)
injds2 <- injsummary(injd2)

p11 <- gg_injbarplot(injds1)  ## type = "incidence" by default
p12 <- gg_injbarplot(injds1, type = "burden")
p21 <- gg_injbarplot(injds2)
p22 <- gg_injbarplot(injds2, type = "burden")

# grid.arrange(p11, p21, p12, p22, nrow = 2)
```

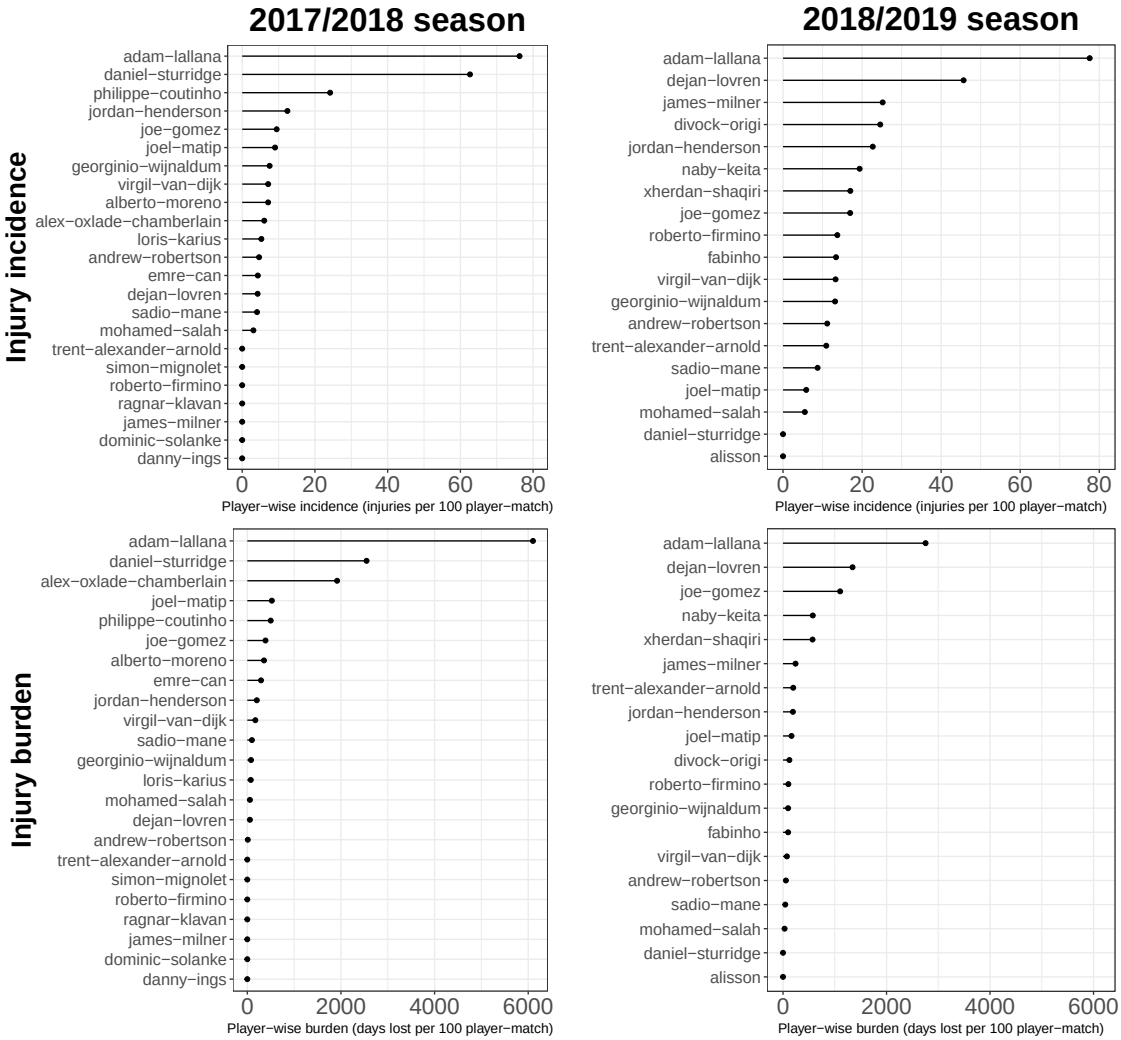


Figure 5.6: Incidence and injury burden among players for each season. Output from the `gg_injbarplot()` function.

As Figure 5.6 shows, in the 2017/2018 season, players with the highest injury incidence

rate and injury burden (all types of injuries) were Adam Lallana and Daniel Sturridge, with 76.3 and 62.6 injuries, per 100 player-matches, and 6102 and 2548 days lost, per 100 player-matches, respectively. In the next 2018/2019 season, Adam Lallana is again the player with the highest injury incidence rate and injury burden, with 77.6 injuries and 2754.3 days lost per 100 player-matches. Dejan Lovren follows him, with 45.7 injuries and 1343 days lost per 100 player-matches.

Which injuries were more frequent? And more burdensome?

We compute injury summary statistics per type of injury and plot the risk matrices.

```
# warnings set to FALSE
## Calculate summary statistics
injds1_perinj <- injsummary(injd1, var_type_injury = "injury_type")
injds2_perinj <- injsummary(injd2, var_type_injury = "injury_type")

## Plot
p1 <- gg_injriskmatrix(injds1_perinj, var_type_injury = "injury_type",
                       title = "Season 2017/2018", add_contour = TRUE)
p2 <- gg_injriskmatrix(injds2_perinj, var_type_injury = "injury_type",
                       title = "Season 2018/2019", add_contour = TRUE)

# Print both plots side by side
# grid.arrange(p1, p2, nrow = 1)
```

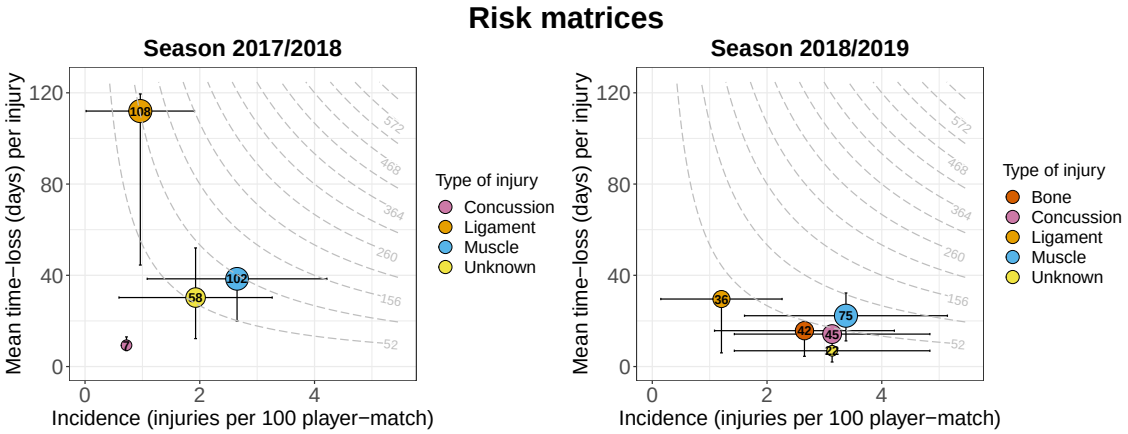


Figure 5.7: Injury risk matrices for each season. Output from the `gg_injriskmatrix()` function.

According to Figure 5.7, injuries that occurred in the 2017/2018 season were more burdensome, especially those related to ligaments, that resulted in 108 days lost per 100 player-matches (36 in the 2018/2019 season). Muscle-related injuries were the next most

burdensome ones in the 2017/2018 season, with 102 days lost per 100 player-matches (75 in the 2018/2019 season).

How many players were injury-free in each month?

We plot polar area diagrams:

```
gg_injprev_polar(injd, by = "monthly")
```

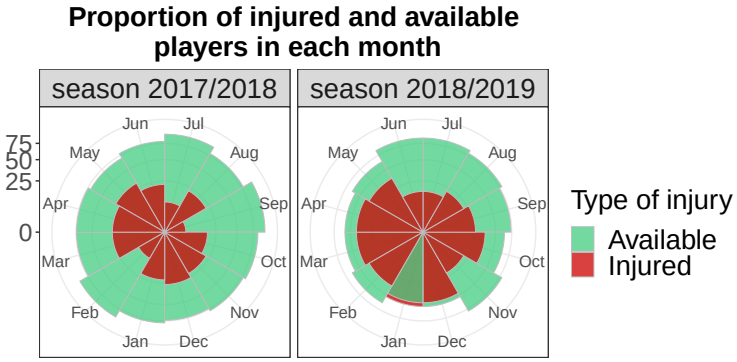


Figure 5.8: Injury prevalence on a monthly basis for each season. Output from the `gg_injprev_polar()` function.

Overall, there were more injured players in the 2018/2019 season than in the previous season. Looking at a monthly basis, see Figure 5.8, there were more differences with regards to player availability, especially during the winter January/February months.

(*d*) *Data modelling*

We now provide a practical demonstration of modelling injuries in the **R** software. Note that we make use of other packages to model the relationships between injuries and variables of interest. We also skip some necessary code for the sake of brevity and refer the reader to the corresponding **injurytools** Vignettes to find out all the steps.

When injuries are **viewed as count data**, we first explore the distribution of the rate variables by plotting histograms as in Figure 5.9. This is useful for deciding which modelling strategy to use. To this end, the following packages might be of interest: **stats**, **lme4**, **glmmTMB** and **pscl**.

We illustrate how to fit four different regression models, namely, the Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial models (see Chapter 2), by modelling the injury burden according to the player's position in the 2017/2018

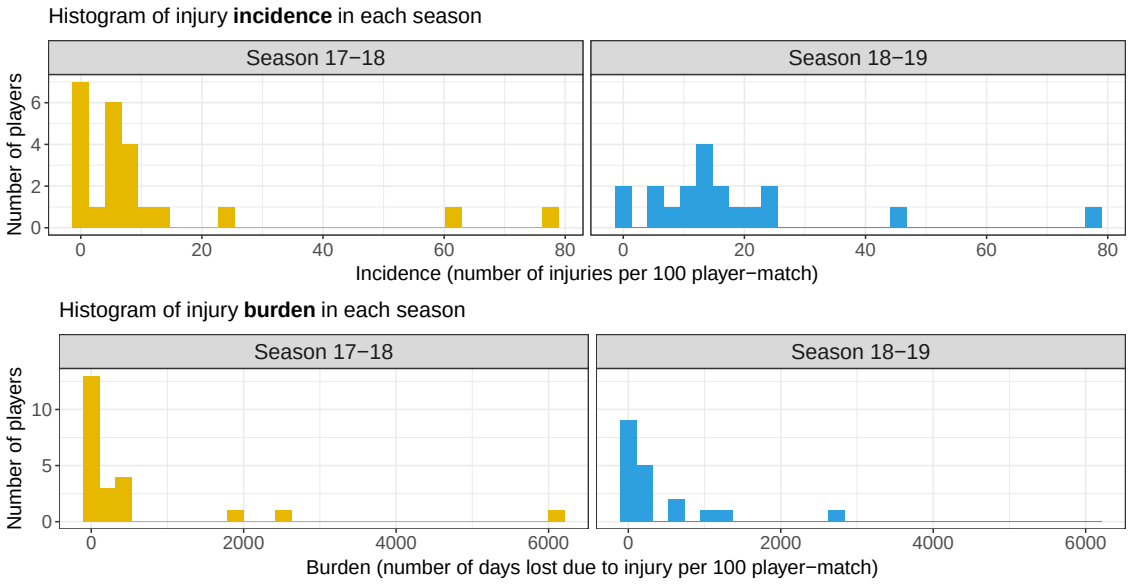


Figure 5.9: Histograms of injury incidence and injury burden for each season of the included data set.

season[7]:

```
# poisson
burden_glm_pois <- glm(ndayslost ~ positionb, offset = log(totalexpo),
                       data = injds1718p,
                       family = poisson)
# poisson random effects model
# burden_glmm_pois <- glmer(formula = ndayslost ~ positionb + (1 | player),
#                           offset = log(totalexpo),
#                           data = injdsp,
#                           family = poisson)

# negative binomial
burden_glm_nb <- glm.nb(ndayslost ~ positionb + offset(log(totalexpo)),
                        data = injds1718p)

# zero-inflated poisson
burden_zinfpois <- zeroinfl(ndayslost ~ positionb | positionb,
                            offset = log(totalexpo),
                            data = injds1718p,
                            link = "logit",
                            dist = "poisson",
```

[7]We use the previously prepared `injds1718p` data frame.

```
                              trace = FALSE, EM = FALSE)

# zero-inflated negative binomial
burden_zinfnb <- zeroinfl(ndayslost ~ positionb | positionb,
                              offset = log(totalexpo),
                              data = injds1718p,
                              link = "logit",
                              dist = "negbin",
                              trace = FALSE, EM = FALSE)
```

We can do the analogue to model the injury incidence as the response variable, e.g.:

```
# example incidence (poisson reg)
incidence_glm_pois <- glm(ninjuries ~ positionb, # + offset(log(totalexpo))
                          offset = log(totalexpo),
                          data = injds1718p,
                          family = poisson)
```

As of now, let us interpret the output of the `burden_glm_pois` model:

```
summary(burden_glm_pois)
cbind(estimate = exp(coef(burden_glm_pois)) * c(100, 1, 1),
      exp(confint(burden_glm_pois)) * c(100, 1, 1)) # for 100 player-matches
#> Waiting for profiling to be done...
```

Table 5.6: Estimated coefficients of the `burden_glm_pois` model. IR stands for injury burden rate and IRR for injury burden rate ratio.

	Estimate	**95 % CI**
$\widehat{\text{IR}}$ (attacker)	197.48	[172.84, 224.35]
$\widehat{\text{IRR}}$ (Defender vs. attacker)	0.97	[0.81, 1.15]
$\widehat{\text{IRR}}$ (Midfielder vs. attacker)	2.76	[2.37, 3.22]

As Table 5.6 shows, the estimated injury burden of attackers is 197.5 days lost per 100 player-matches. Besides, the injury burden of midfielders is significantly higher than that of attackers (Adam Lallana plays as a midfielder). The corresponding estimated injury burden rate ratio is 2.76. However, the fit is not good enough, since:

```
> injds1718p |>
    group_by(positionb) |>
    summarize(mean = mean(injburden),
              median = median(injburden))
# A tibble: 3 x 3
```

```
  positionb  mean median
  <fct>     <dbl>  <dbl>
1 Attack     457.   55.5
2 Defender   188.  113.
3 Midfield  1432.  248.

> coefs_p <- coef(burden_glm_pois)
> coefs_p <- exp(c(coefs_p[[1]], coefs_p[[1]] + coefs_p[[2]],
                   coefs_p[[1]] + coefs_p[[3]])) * 90 * 100
> coefs_nb <- coef(burden_glm_nb)
> coefs_nb <- exp(c(coefs_nb[[1]], coefs_nb[[1]] + coefs_nb[[2]],
                    coefs_nb[[1]] + coefs_nb[[3]])) * 90 * 100
> data.frame(positionb = levels(injds1718p$positionb),
             estimate_pois = coefs_p,
             estimate_nb = coefs_nb)
  positionb estimate_pois estimate_nb
1    Attack      197.4757    455.7904
2  Defender      191.2059    187.9884
3  Midfield      545.2393   1427.4855
```

Finally, we compare the four models. We compute the conditional predicted mean probabilities of each model and display them over the histogram of the data to examine the fits, see Figure 5.10.

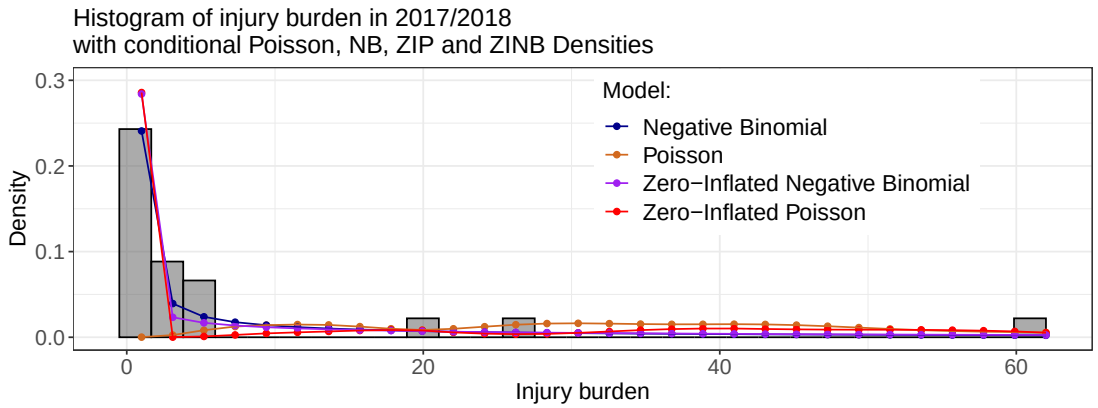


Figure 5.10: Histogram of injury burden in the 2017/2018 season, along with the conditional predicted mean probabilities from each model.

Besides, we compute goodness of fit measures such as AIC, BIC and deviance explained and present them in Table 5.7. According to these measures, the Negative Binomial model (`burden_glm_nb`) fits these data best.

Table 5.7: Goodness of fit measures of the fitted models ordered according to the BIC.

Model	AIC	BIC	Deviance Explained
Negative binomial model	200.83	205.00	13.29
Zero-inflated Negative Binomial model	204.48	211.79	2.77
Zero-inflated Poisson model	1871.10	1877.36	11.03
Poisson model	2417.26	2420.40	11.25

On the other hand, when injuries are **viewed as time-to-event data**, we can estimate the probability of being injury-free over time or estimate which factors and to what extent they affect this probability using packages such as **survival**, **survminer**, **coxme** or **pammtools**.

We first show the application of the well-known Kaplan-Meier (KM) method and Cox Proportional Hazards (Cox PH) model on sports injury data and, after that, we describe two possible survival modelling strategies that take into account the recurrence of injuries, and data include repeated observations per player.

Methods for time to first injury

We prepare the data so that for each separate season we have an `injd` object with each observation (row) corresponding to **time to first injury** (or end of the season, or a transfer to another team, i.e. censored observation). The final data frames are called `injd1718_sub` and `injd1819_sub`. Then,

```
## we join both data sets by row
injd_sub <- bind_rows("17-18" = injd1718_sub,
                      "18-19" = injd1819_sub,
                      .id = "season")
```

We estimate the survival probabilities, $\hat{S}_{\text{KM}}(t)$, in each season, as follows:

```
fit <- survfit(Surv(tstart_day, tstop_day, status) ~ seasonb,
               data = injd_sub)
fit
```

```
#> Call: survfit(formula = Surv(tstart_day, tstop_day, status) ~ seasonb,
#>     data = injd_sub)
#>
#>                     n events median 0.95LCL 0.95UCL
#> seasonb=2017/2018 23     16    265     152      NA
```

```
#> seasonb=2018/2019 19     17    106      84     165
```

The number of first-time injuries in both seasons is similar (16 vs. 17), but the median survival probability is lower in the 2018/2019 season, i.e. in 2018/2019 the estimated probability of being injury-free on or after the 106th day is less than or equal to 0.5 (equivalently, the estimated probability of surviving 106 days (three months and a half) is 0.5), whereas in 2017/2019 the probability of surviving the same time is 0.696 (see Figure 5.11).

Next, we plot the Kaplan-Meier curves for each season based on the above results via the `survminer::ggsurvplot()` function. The graphic is shown in Figure 5.11. Additionally, we have added information on the risk sets over time, the estimated median survival probabilities for each curve and the p-value obtained from the log-rank test. There are statistical differences regarding the survival probabilities of first-time injuries between the 2017/2018 and 2018/2019 seasons.

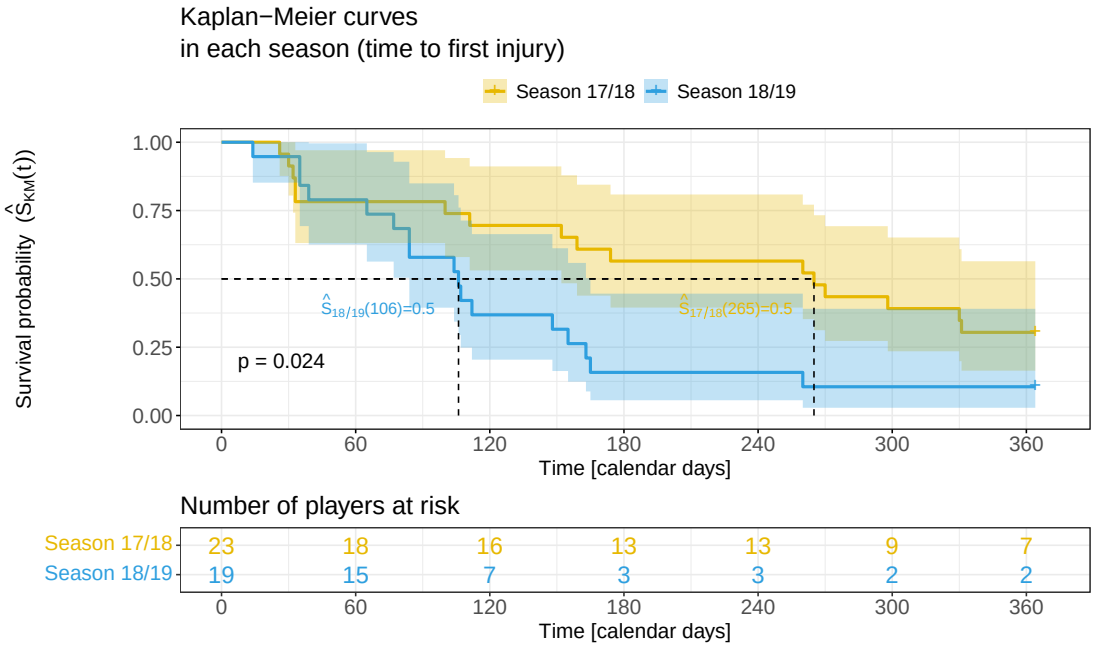


Figure 5.11: Estimated Kaplan-Meier curves for each season from the `fit` model.

We fit a Cox PH model, that relates some player-related covariates (e.g. `positionb`, `age` and `yellow`) to the injury outcome, through the hazard function, to the `injd1819_sub` named data frame, as:

```
## create positionb column
## (so that the categories are: Attack, Defender, Goalkeeper and Midfield)
injd1819_sub <- mutate(injd1819_sub,
```

```
                        positionb = factor(str_split_i(position, "_", 1)))

cfit <- coxph(Surv(tstop_day, status) ~ positionb + age + yellows,
              data = injd1819_sub |>
                     filter(positionb != "Goalkeeper") |> droplevels())
```

The estimated effects of the `cfit` model are displayed in Figure 5.12 using the `survminer::ggforest()` function. It shows the hazard ratios and 95% confidence intervals, together with the p-values of each covariate, and further information about the goodness of fit of the `cfit` model. The results, however, are not meaningful.

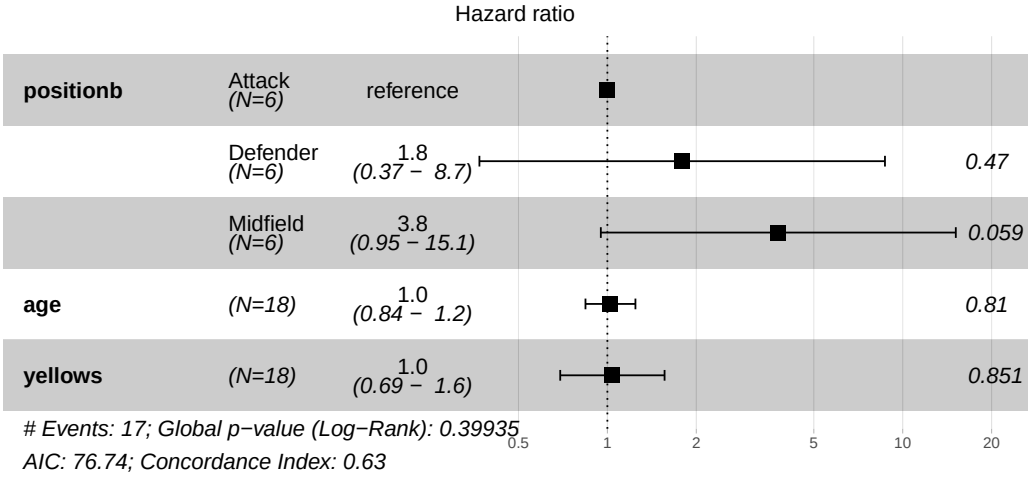


Figure 5.12: Estimated hazard ratios, 95% confidence intervals and further details from the `cfit` model.

Then, we check if the proportional hazards assumption for the Cox PH model holds by computing the Schoenfeld residuals. See Figure 5.13. The PH assumption is violated as the Global Schoenfeld Test p-value reveals.

Models for time to (subsequent) injuries

We use the (previously prepared) `injd_sub` data to fit a stratified Cox PH model. With this model, we fit a different baseline hazard function for each level (stratum) of the `seasonb` covariate (strata), i.e. $\lambda(t|\boldsymbol{x}) = \lambda_{0,k}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})$ for $k = 1, 2$.

```
sfit <- coxph(Surv(tstart_day, tstop_day, status) ~ age + strata(seasonb),
              data = injd_sub)
summary(sfit)
```

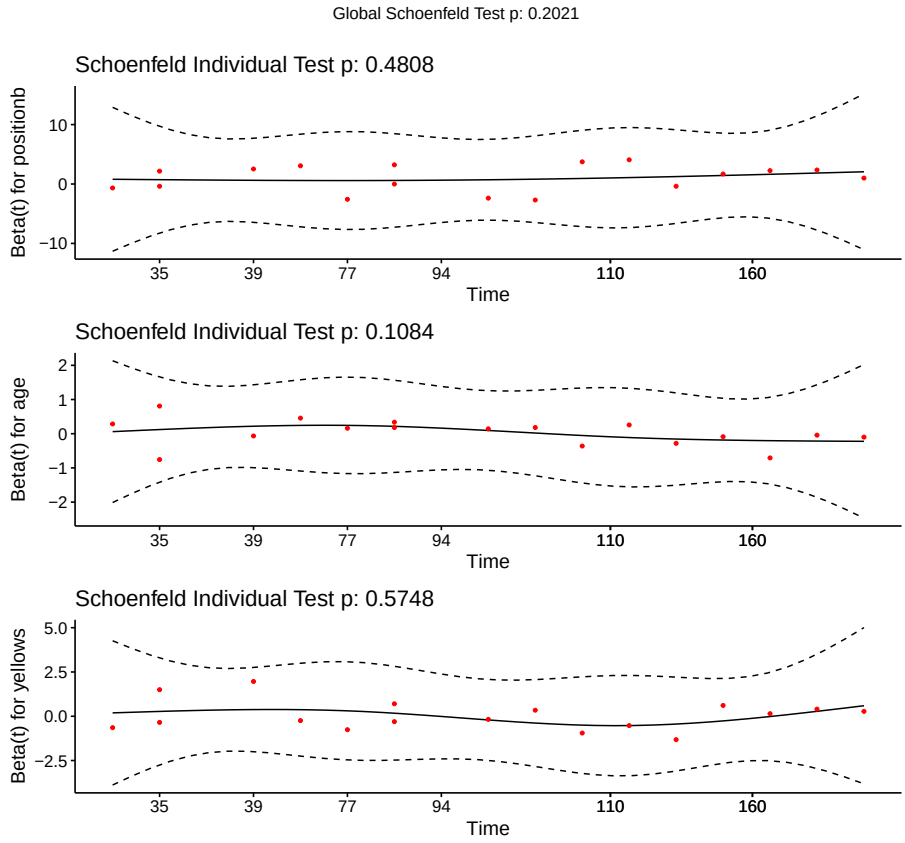


Figure 5.13: Schoenfeld residuals for each covariate in the `cfit` model.

```
#> Call:
#> coxph(formula = Surv(tstart_day, tstop_day, status) ~ age + strata(seasonb),
#>     data = injd_sub)
#>
#>   n= 42, number of events= 33
#>
#>        coef exp(coef) se(coef)     z Pr(>|z|)
#> age 0.01749   1.01764  0.05541 0.316    0.752
#>
#>     exp(coef) exp(-coef) lower .95 upper .95
#> age     1.018     0.9827    0.9129     1.134
#>
#> Concordance= 0.6  (se = 0.069 )
#> Likelihood ratio test= 0.1  on 1 df,   p=0.8
#> Wald test            = 0.1  on 1 df,   p=0.8
#> Score (logrank) test = 0.1  on 1 df,   p=0.8
```

The effect of age, $\widehat{\text{HR}}_{\text{age}} = \exp\left(\hat{\beta}_{\text{age}}\right) = 1.02$, is not significant. However, we will keep on and illustrate how to plot the estimates of two players of different ages, 18 years old vs.

36 years old in both seasons, based on the fitted stratified model.

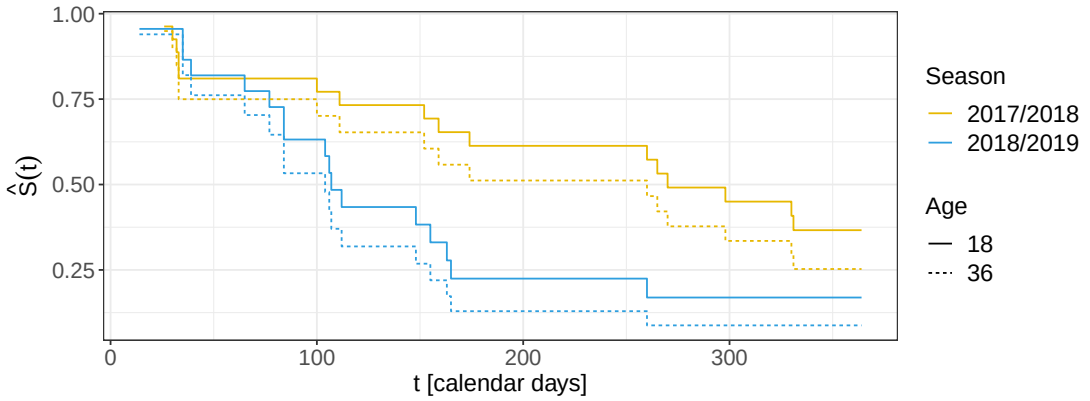


Figure 5.14: Estimated survival probabilities of two players, a 18 year-old vs. 36 years old, in both seasons, from the `sfit` model.

Figure 5.14 shows that the estimated risk of injury for the older player is higher in both seasons, as the estimated survival probability decreases more rapidly. Moreover, there are differences in the estimated baseline hazards. The risk of injury in the 2018/2019 season is higher.

As a final model example, we fit a shared frailty model in which the frailty term follows a Gamma distribution using the `frailty(player)` syntax inside `survival::coxph()` function's formula:

```
sffit <- coxph(Surv(tstart_minPlay, tstop_minPlay, status) ~
                 age + days_lost +
                 frailty(player, distribution = "gamma"), data = injd)
```

Alternatively, we can use the **coxme** package (there are also more packages) and fit a model with a log-normal frailty using the `(1 | player)` syntax:

```
sffit2 <- coxme(Surv(tstart_minPlay, tstop_minPlay, status) ~
                  age + days_lost + (1 | player), data = injd)
```

By fitting this model, we are able to model the dependence between several survival times through a frailty term that is shared by all the survival times pertaining to a player. That is, the survival times of a player who sustains multiple injuries have the same level of frailty attached to them.

```
summary(sffit)
```

```
#> Call:
#> coxph(formula = Surv(tstart_minPlay, tstop_minPlay, status) ~
#>     age + days_lost + frailty(player, distribution = "gamma"),
#>     data = injd)
#>
#>   n= 104, number of events= 81
#>
#>                           coef      se(coef) se2      Chisq  DF    p
#> age                        0.489503 0.278965            3.08  1.00 7.9e-02
#> days_lost                 -0.006381 0.009761 0.006744   0.43  1.00 5.1e-01
#> frailty(player, distribut                               230.57 15.72 3.3e-40
#>
#>           exp(coef) exp(-coef) lower .95 upper .95
#> age          1.6315     0.6129    0.9443     2.819
#> days_lost    0.9936     1.0064    0.9748     1.013
#>
#> Iterations: 10 outer, 184 Newton-Raphson
#>      Variance of random effect= 2.673384   I-likelihood = -171.4
#> Degrees of freedom for terms= -2.3  0.5 15.7
#> Concordance= 0.882  (se = 0.026 )
#> Likelihood ratio test= 122.5  on 13.88 df,   p=<2e-16
```

The estimated variance of the frailty term (random effect) is $\hat{\sigma}^2 = 2.69$ and the p-value of the frailty term is significant. See also the model output in Figure 5.15.

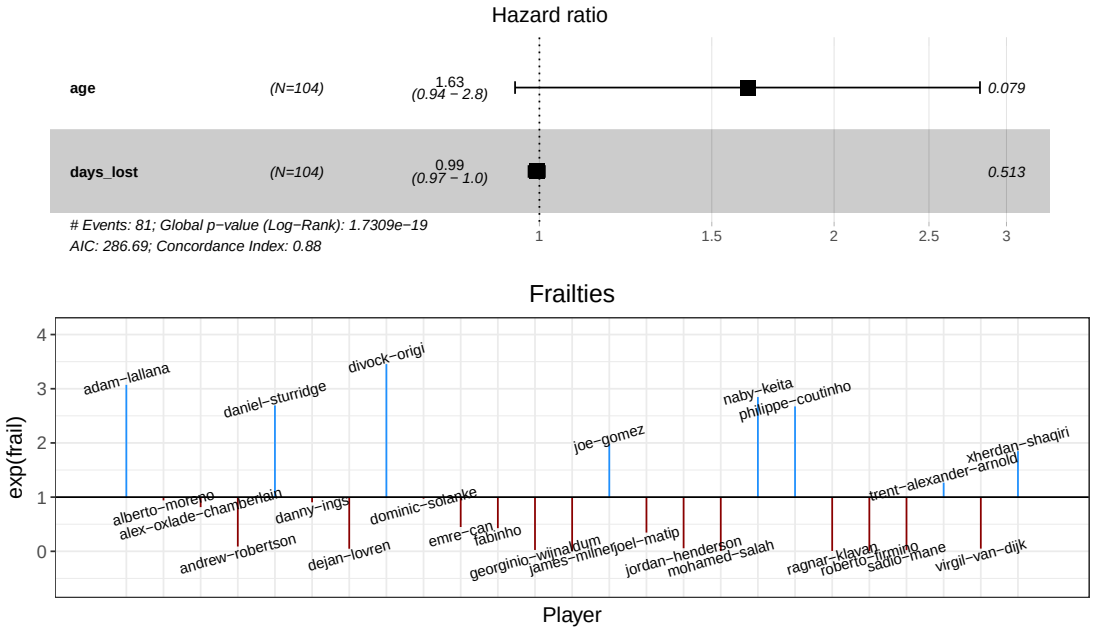


Figure 5.15: Estimated hazard ratios and frailty terms from the `sffit` model.

Part III.
Conclusions and Further Research

Chapter 6

Conclusions and further research

This dissertation investigated the suitability of various statistical modelling approaches for tackling specific research questions and contexts within sports injury prevention research. In what follows, we present the main conclusions related to each specific objective:

(*a*) *To assess different variable selection methods together with shared frailty Cox models for identifying biomechanical risk factors for subsequent sports injuries.*

We addressed the issue of the number of covariates being too large for a statistical model to be comprehensible and interpretable in the context of functional screening tests. These tests evaluate the strength, power, joint stability (e.g. knees, ankle, neck, etc.), movement patterns and asymmetries. Their primary goal is to evaluate the absence of potential physical risks to athletes.

We compared several regularized Cox methods, including Best Subset Selection (BeSS), Least Absolute Shrinkage and Selection Operator (Lasso), Elastic Net, Ridge regression, and Group Lasso; and Boosting in Cox regression. We also discussed the use of the most relevant variables for fitting shared frailty Cox models. Furthermore, we conducted a simulation study to better understand the robustness of these models across three possible scenarios.

We demonstrated that predictive performance significantly improves with the availability of more player observations. Methods that lead to sparse models and favour interpretability, such as BeSS and Boosting in Cox regression, are preferred when the sample size is small. As the sample size increases, differences between models become less pronounced.

We emphasize that our primary objective was to analyze an appropriate statistical

modelling approach for functional screening tests and sports injury data, rather than to provide evidence about risk factors for lower-limb injuries. Predominantly, the selected variables included those from unilateral drop jump tests and the active straight leg raise test. Given the small sample size, the associations observed in our application study should be interpreted with caution. This work highlighted the existing potential of shared frailty Cox models.

(*b*) *To develop and assess a flexible recurrent time-to-event approach for modelling the effects of training load on subsequent sports injuries.*

We proposed a flexible modelling approach to analyze complex, non-linear associations between external training loads and subsequent time-loss injuries. External training load refers to the cumulative stress placed on athletes from multiple training sessions and competitions over a period of time. It is measured using various metrics, such as distance covered, high-speed running, time, and sprints, among others; metrics that are recorded regularly.

To achieve this, we extended and assessed the PAMM model class within the context of recurrent events with time-varying covariates, specifically modelling them as WCE-type cumulative effects. Additionally, we introduced a method to determine a relevant time window based on data, incorporating a ridge penalty to minimize the influence of exposures recorded long ago.

Simulation studies demonstrated the model's capability to accurately recover various shapes for the true weight function under different levels of heterogeneity. The model effectively estimated these shapes from the data using P-splines. Furthermore, the results indicated that PAMM with ridge penalization offers the highest accuracy in estimating partial effects. Consequently, this method is particularly effective in identifying the relevant time window during which past exposures cumulatively affect the hazard.

The modelling framework flexibly modelled the cumulative effect of the *Speed* and *Dist* exposure variables on subsequent injuries using the application data. The results suggested that no more than seven past sessions were relevant concerning the cumulative effects of the *Speed* and *Dist* variables. Moreover, these effects were estimated to have a non-linear decaying impact. The frailty term accounted for the correlation between subsequent injuries from the same player with an estimated variance of $\hat{\sigma}_b = 0.22$. Furthermore, we compared this model with alternative measures of training load exposures, namely different variants of the Acute Chronic Workload Ratio measure that utilize rolling averages, exponentially weighted moving averages, and others. Our model was proven to provide the best fit, in addition to being able to estimate cumulative effects and determine

a relevant time window based on data.

The work highlighted the potential value of PAMMs with WCE-type cumulative effects in evaluating recurrent events in sports medicine. It contributed to the existing literature and offered insights for designing and comparing personalized training plans related to injury risk.

(*c*) *To develop software that implements the statistical methods for analyzing sports injury data proposed in this dissertation.*

We implemented in the **R** software the statistical modelling approaches proposed in this dissertation. To this end, we used existing functionalities in **R** and also developed new ones. Working with various injury data sets across different types of tasks –preprocessing, exploring, modelling, reporting, etc.– and creating **R** functions that pursued similar purposes, allowed us to reflect on and identify some essential coding features. Eventually, we established a standardized workflow and integrated the code into an **R** package named **injurytools**, complete with accompanying documentation.

We are confident that our public code repositories and the **injurytools** package facilitate the use of the proposed approaches and make them more accessible to practitioners. This has several important implications: it supports the development of these research fields, facilitates the transfer of knowledge, and brings sports scientists closer to user-friendly statistical tools, which can presumably support their decision-making process. Finally, we hope that other researchers can verify our findings, reproduce the analysis, and build upon our work.

Further research

The research conducted for this dissertation has highlighted several topics and identified potential areas for further investigation in the statistical analysis of sports injuries, an area that is still emerging.

Firstly, we are interested in evaluating how the number of injuries per player –kept fixed in our simulation studies– affects the model performance either in the context of functional tests data or in the external training load data. Connected to this, the distribution of the frailty term is also a subject of investigation. In this dissertation, we considered the log-normal distribution (the Gaussian distribution in the linear predictor scale). Our choice of a log-normal distribution for the frailty was influenced by the application data sets. Log-normal frailty models fitted to these data showed better fits according to AIC

rather than models with a Gamma frailty. This choice was also based on the distribution's intuitive nature and its compatibility with more flexible predictor structures. However, we recognize that the Gamma frailty model has a closed-form solution, making it computationally less demanding. We find it interesting to assess the influence of the frailty term's distribution in both contexts.

Secondly, in the context of functional screening tests data, we chose the shared frailty Cox model due to the data's characteristics, which include few injuries and many variables. We consider an interesting topic for future study, comparing our approach with parametric survival models and flexible spline-based survival models. Although the parametric models require a specific distribution to be considered for the time-to-event outcome, and the flexible spline-based survival models require more parameters to be estimated, both might show some benefits in some other situations: predictions can be extrapolated farther than the maximum followed-up time observed in the case of parametric survival models, like the Weibull Accelerated Failure Time (AFT) model (Kalbfleisch and Prentice, 2011); or very flexible covariate effects can be modelled with spline-based survival models, like with the PAMM.

We are also interested in analyzing the simultaneous modelling of variable selection techniques and shared frailty Cox models. In this regard, the works by Groll et al. (2017) and Hohberg and Groll (2020) are particularly noteworthy. Alternatively, Multivariate Survival Tree (MST) models (Fan and Li, 2002; Su and Fan, 2004) might be of interest. MST is a decision tree method capable of capturing non-linear relationships and interactions between covariates and recurrent time-to-event survival outcomes. Therefore, it can identify how risk factors interact, rather than simply selecting isolated risk factors (Bittencourt et al., 2016).

Thirdly, we encouraged collecting data from multiple teams and across various seasons. But, we are aware that having more data introduces additional complexity. This complexity extends to various aspects of the study, including the study design itself, such as deciding which data to collect and when to collect it. Additionally, it encompasses team-level factors that can potentially influence injury risk, such as training regimens, playing styles, and the quality of each team's medical staff, among others. A random effects model with different (multiple) levels of nesting, which addresses the between-teams heterogeneity (or another type of hierarchy), would be of interest, such as the nested frailty models, frailty interaction models or joint models (Rondeau et al., 2012; Tsiatis et al., 1995).

Fourthly, we want to emphasize that we did not intend to analyze any causal relation-

ships; causality was beyond the scope of this dissertation. We believe that causal inference is an active and ongoing research field (Hernán and Robins, 2010; Shrier, 2007; Bittencourt et al., 2016; Kalkhoven et al., 2021) that will provide valuable insights in the near future to sports injury prevention research.

Finally, we plan to continue extending and updating the **injurytools** package, currently at version v.1.0.3, by adding new features and improving support.

Concluding remarks

Identifying the most suitable statistical model to estimate injury occurrence is of significant interest in practical applications. Nevertheless, we acknowledge that accurately predicting an injury is an inherently challenging task, if not impossible (Lee and Zumeta-Olaskoaga, 2022). Sports injuries occur and will continue to occur. "*We control what we can, and know to expect the unexpected*" says Lindsay Slater, a sports scientist at the University of Illinois (Fiscutean, 2021). There are a variety of factors, including lifestyle, biological makeup, genetic characteristics and contextual ones, that influence an athlete's susceptibility to sports injuries (Van Mechelen et al., 1992; Meeuwisse, 1994). Some have suggested a more ecological view that includes context at multiple levels, i.e. at the individual, socio-cultural and environmental levels (Finch, 2006; Bolling et al., 2018); qualitative research that recognizes multiple realities and seeks to understand and interpret relationships between these realities surrounding sports injury.

Although predicting injuries with absolute certainty is unfeasible, accurately assessing an individual's risk level in relation to physical activity and injuries is entirely achievable. In this regard, statistical modelling is essential for understanding and quantifying the risk of sports injuries. The primary focus here is on comprehending relevant concepts such as association, causality, uncertainty, and complexity rather than solely predicting an athlete's injury (Meeuwisse, 1994; Shrier, 2007; Bittencourt et al., 2016).

All these approaches are notably data-intensive. Unfortunately, in practical applications, it is often challenging to obtain sufficient data for advanced statistical analyses. On one hand, sports injuries, being unique and specific, occur infrequently. This is particularly true for specific types of sports injuries, such as hamstring injuries, anterior cruciate ligament injuries, spondylolysis, among others. On the other hand, sports injury data are often limited due to the reluctance of sports clubs to share their information with competitors. In this sense, when it comes to football, there are noteworthy research initiatives that collect injury data from various clubs. These include the Union of European Football

Association (UEFA) Champions League (UCL) Injury Study, the UEFA Elite Club Injury Study, the UEFA Women's Elite Club Injury Study, and the European Football Lab. The primary goal of these initiatives, and the annual reports they produce, is to understand the nature and frequency of injuries among elite football players. They aim to help UEFA and the clubs develop strategies for injury prevention and player welfare, and to enhance the overall quality of the sport.

In this context, it is crucial to highlight the importance of data collection, standardization and storage in sports medicine (Bahr et al., 2020). It's also worth mentioning that in this dissertation we invested a considerable amount of time in data preprocessing. Bearing this in mind, we deem it essential to establish standardized data collection protocols across various sports to make the task of data preprocessing as effortless and automatic as possible. Universally accepted standards for recording injury-related data would mitigate inconsistencies between data sets, discrepancies in categorical variables (such as different names referring to the same category), irregularities in data collection periodicity, and the inclusion of values that are not missing at random. Furthermore, such standards could help in identifying and minimizing potential sources of bias (https://catalogofbias.org/) and erroneous data, among other issues. With a well-defined research goal, researchers in this field would be able to outline a study design and determine the specific data necessary for collection (Nielsen et al., 2020), without being overly concerned with the challenges related to data recording. This approach would ultimately ensure both the reliability and validity of the research findings.

While our simulation studies provided valuable insights, we acknowledge the importance of integrating physiological, psychological, and contextual understanding into this complex and multifaceted problem. We strongly advocate for multi- and interdisciplinary research in this area. Collaborating with a diverse team, including sports scientists, coaches, physiologists, physical therapists, physicians, and the athletes themselves, fosters solutions that encompass all these aspects. Such collaboration ensures that the statistical approach aligns with the practical needs of athletes and coaches, and thereby, it is more likely to yield actionable insights. In this sense, effective communication among collaborators and colleagues, as conveying findings in a clear and common language, is key. "*Having huge volumes of data in one thing; making sense of it is another*" (Derek McHugh, a data scientist in Kitman Labs (Fiscutean, 2021)). As statisticians, our analyses and the collection of data are undoubtedly useless without properly communicating the extracted results. Similarly, communication plays a crucial role, in elite football clubs. It has been claimed that the communication quality between the medical team and the head coach, as well as the leadership style of the head coach, is correlated with injury rates, training

attendance, and match availability, in elite football clubs (Ekstrand et al., 2019, 2018).

We consider that this dissertation represents a significant step towards a more comprehensive understanding of sports injuries from a statistical perspective. Our vision is to see research translated into practical applications, grounded in a commitment to sound methodology and adherence to best practices. Ongoing research is needed to provide valuable insights into the field of sports medicine, and in particular, into the understanding of the physical demands of modern football, that will eventually assist sports clubs and medical teams in managing players' performance, reducing injury rates, and enhancing players' health and safety.

Summary of the contributions

In the following, we outline the scientific contributions made during the course of this doctoral thesis:

Scientific articles (derived from this dissertation)

- Zumeta-Olaskoaga, L., Weigert, M., Larruskain, J., Bikandi, E., Setuain, I., Lekue, J., Küchenhoff, H., and Lee, D.-J. (2023). "Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models". *AStA Advances in Statistical Analysis*, 107(1-2), 101-126. doi: 10.1007/s10182-021-00428-2.

- Zumeta-Olaskoaga L., Bender A., Lee D.-J. (2023). "Flexible modelling of time-varying exposures and recurrent events to analyze training loads effects in team sports injuries". *Submitted*.

- Zumeta-Olaskoaga L., Lee D.-J. (2023). "injurytools: A Toolkit for Sports Injury Data Analysis". *Under preparation*.

Other scientific articles (in collaboration)

- Monasterio X., Gil S.M., Bidaurrazaga-Letona I., Lekue, J.A., Santisteban, J., Diaz-Beitia, G., Lee, D.-J., Zumeta-Olaskoaga, L., Martin-Garetxana, I., Bikandi, E., Larruskain J. (2023). "The burden of injuries according to maturity status and timing: A two-decade study with 110 growth curves in an elite football academy". *European Journal of Sport Science*, 23:2, 267-277. doi: 10.1080/17461391.2021.2006316.

- Monasterio X., M. Gil S., Bidaurrazaga-Letona I., Lekue J.A., Diaz-Beitia G., Santisteban J.M., Lee D.-J., Zumeta-Olaskoaga L., Martin-Garetxana I., Larruskain J. (2023).

"Peak Height Velocity Affects Injury Burden in Circa-PHV Soccer Players". *International Journal of Sports Medicine*, 44(4):292-297. doi: 10.1055/a-1983-6762.

Dissemination articles

- Lee D.-J., Zumeta-Olaskoaga L. (2022). "Can we really predict injuries in team sports?". *Boletín de Estadística e Investigación Operativa* (*BEIO*). Vol 38, No 3.

Conference contributions

- Zumeta-Olaskoaga L. and Lee D.-J. *"Statistical Modeling in sports injury prevention". Oral contribution*. IV Jornadas Científicas de Estudiantes de la SEB (JJSEB 2019) 5-6 September 2019, Albacete (Spain).

- Zumeta-Olaskoaga L., Weigert M., Larruskain J., Bikandi E., Küchenhoff H. and Lee D.-J. *"Statistical modelling for time-to-event sports injury data: assessing biomechanical risk factors". Oral contribution*. I Congreso Virtual de la Sociedad Española de Epidemiología (SEE) y da Associaçao Portuguesa de Epidemiologia (APE), 21-23 y 29-30 October 2020, Online.

- Zumeta-Olaskoaga L., Larruskain J., Lekue J.A., Bikandi E., Setuain I. and Lee D.-J. *"Flexible time-to-event modelling approaches for recurrent football injury data". Poster contribution*. Royal Statistical Society 2021 International Conference (RSS 2021), 6-9 September 2021, Online.

- Zumeta-Olaskoaga L., Monasterio X., Larruskain J., Lekue J.A., Santisteban J.M., Diaz-Beitia G. and Lee D.-J. *"Estimation of injury patterns according to maturity status and timing in an elite football academy based on zero-inflated models". Oral contribution*. XVIII Congreso Español de Biometría (CEB 2022), 25-27 May 2022, Madrid (Spain).

- Zumeta-Olaskoaga L., Bender A., Küchenhoff H. and Lee D.-J. *"Modelling the recurrence of injuries in football players using piece-wise exponential additive mixed models". Poster contribution*. 36th International Workshop on Statistical Modelling (IWSM 2022), 18-22 July 2022, Trieste (Italy).

- Zumeta-Olaskoaga L. and Lee D.-J. *"injurytools: a toolkit for sports injury data analysis". Oral contribution*. VI Jornadas Científicas de Estudiantes de la SEB (JSEB 2022), 14-16 September 2022, Valencia (Spain).

- Zumeta-Olaskoaga L., Bender A., Küchenhoff H. and Lee D.-J. *"Estimating the risk of time-loss injuries in football players through recurrent time-to-event methods". Oral*

contribution. 15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022), 17-19 December 2022, Londres (United Kingdom).

- Zumeta-Olaskoaga L., Bender A. and Lee D.-J. *"How do past training exposures affect injury risk in football?". Oral contribution*. XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría (CEB-EIB 2023), 27-30 June 2023, Vigo (Spain).

- Zumeta-Olaskoaga L., Bender A. and Lee D.-J. *"Flexible modelling of time-varying training exposures on the risk of recurrent injuries in football". Oral contribution*. 37th International Workshop on Statistical Modelling (IWSM 2023), 17-21 July 2023, Dortmund (Germany).

Awards

- **Best Poster Award** at the 36th International Workshop on Statistical Modelling 2022 (IWSM 2022), Trieste.
  Awarded work: Zumeta-Olaskoaga L., Bender A., Küchenhoff H. and Lee D.-J. *Modelling the recurrence of injuries in football players using piece-wise exponential additive mixed models.*

Other conference contributions

- Lee D.-J., Zumeta-Olaskoaga L., Larruskain J., Bikandi E., Setuain I. and Lekue J.A. *"Modelling and prediction in time-to-event sports injury data: a penalized Cox regression approach". Oral contribution*. XXXIX Congreso Nacional de Estadística e Investigación Operativa y de las XIII Jornadas de Estadística Pública (SEIO 2022), 7-10 June 2022, Granada (Spain).

- Renteria J., Zumeta-Olaskoaga L., Bikandi E., Larruskain J. and Lee D.-J. *"Potential risk factors of injuries in professional football using Multivariate Survival Trees: a comparison of female vs. male football players". Poster contribution*. XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría (CEB-EIB 2023), 27-30 June 2023, Vigo (Spain).

- Álvarez O., Zumeta-Olaskoaga L., Martínez-Minaya J. and Lee D.-J. *"A zero-inflated Bayesian modeling of sports injury risk incidences?". Poster contribution*. XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría (CEB-EIB 2023), 27-30 June 2023, Vigo (Spain).

- Renteria J., Zumeta-Olaskoaga L., Bikandi E., Larruskain J. and Lee D.-J. *"Multivariate Survival Trees for prediction of lower limb injuries in professional male and female*

football players". Poster contribution. 37th International Workshop on Statistical Modelling (IWSM 2023), 17-21 July 2023, Dortmund (Germany).

Outreach activities

Public lectures

- *Emakumeak Zientzian Atzo eta Gaur ekitaldia, "Florence Nightingale eta Lore Zumeta Olaskoaga"*, 7. Emakumeak Zientzian proiektua. Bidebarrieta liburutegia, Bilbao (2023-02-17).
- *"Analizando lesiones deportivas mediante la ciencia de datos"*, Gymkana: La ciencia de los datos, Mary Eleanor Spear (Stem4Girls UC3M). Universidad Carlos III de Madrid. Getafe (2022-02-11).
- *"Estatistika, kirola eta lesionatzeko arriskua. Zer zerikusi dute?"*, VII. Zarautz Zientziaz Blai. ZarautzON. Cine Modelo, Zarautz (2021-11-16).
- *"Happy Stats Hour: Sports Analytics"*, Adrià Arbués, Guillermo Villacampa and Lore Zumeta, Societat Catalana d'Estadística (SoCE) (2021-01-19).

Video

- *"#STATPíldora: Estadística y Deporte"*, Proyecto Stat Wars (2023-06-21).
- *Zientzialari 157 "Biziraupenerako analisiari esker lesio bat sufritzeko arriskua estima daiteke"*, Zientzia Kaiera UPV/EHUko Kultura Zientifikoko Katedra (2021-07-13).

Radio

- *Matematika aplikatua, futbolarien lesioak aurreikusteko. Lore Zumetaren tesian ari da ereduak bilatzen*, Euskadi Irratia, Faktoria (2023-02-09).

Written

- *Lore Zumeta: "Futboleko lesioen mekanismoa hobeto ulertu nahi dugu"*, Zarauzko Hitza, (2021-11-16).
- *Lore Zumeta, matematikaria: "Ikerketan hasi berria naiz, eta dena dut deskubritzeko"*, Zientzia Kaiera hedabide digitala, Emakumeak Zientzian (2020-01-03).
- *"The scientists who inspired us (II): Florence Nightingale"* and *"Inspiratzen gaituzten emakumeak (II): Florence Nightingale"*, Lore Zumeta, BCAM News (2020-03-26).

Bibliography

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

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

Argyropoulos, C. and Unruh, M. L. (2015). Analysis of time to event outcomes in randomized controlled trials by generalized additive models. *PLoS One*, 10(4):e0123784.

Bache-Mathiesen, L. K., Andersen, T. E., Dalen-Lorentsen, T., Clarsen, B., and Fagerland, M. W. (2022). Assessing the cumulative effect of long-term training load on the risk of injury in team sports. *BMJ Open Sport & Exercise Medicine*, 8(2):e001342.

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

Bahr, R., Clarsen, B., Derman, W., Dvorak, J., Emery, C. A., Finch, C. F., Hägglund, M., Junge, A., Kemp, S., 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)). *Orthopaedic Journal of Sports Medicine*, 8(2).

Bahr, R., Clarsen, B., and Ekstrand, J. (2018). Why we should focus on the burden of injuries and illnesses, not just their incidence. *British Journal of Sports Medicine*, 52(16):1018–1021.

Bahr, R. and Holme, I. (2003). Risk factors for sports injuries—a methodological approach. *British Journal of Sports Medicine*, 37(5):384–392.

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

Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research*, 29(11):3424–3454.

Banister, E. W. and Calvert, T. W. (1980). Planning for future performance: implications for long term training. *Canadian Journal of Applied Sport sciences. Journal Canadien des Sciences Appliquees au Sport*, 5(3):170–176.

Bender, A. (2018). *Flexible modeling of time-to-event data and exposure-lag-response associations*. PhD thesis, Ludwig-Maximilians-Universität München, LMU.

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

Bender, A. and Scheipl, F. (2018). pammtools: Piece-wise exponential additive mixed modeling tools. *arXiv preprint arXiv:1806.01042*.

Bender, A., Scheipl, F., Hartl, W., Day, A. G., and Küchenhoff, H. (2019). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*, 20(2):315–331.

Bender, A., Scheipl, F., Kopper, P., and Burk, L. (2023). *pammtools: Piece-Wise Exponential Additive Mixed Modeling Tools for Survival Analysis*. R package version 0.5.8, `https://CRAN.R-project.org/package=pammtools`.

Binder, H. (2013). *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*. R package version 1.4, `https://CRAN.R-project.org/package=CoxBoost`.

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

Bishop, C., Turner, A., Gonzalo-Skok, O., and Read, P. (2020). Inter-limb asymmetry during rehabilitation understanding formulas and monitoring the "magnitude" and "direction". *Aspetar Sports Medicine Journal*, 9(1):18–22.

Bittencourt, N. F., Meeuwisse, W., Mendonça, L., Nettel-Aguirre, A., Ocarino, J., and Fonseca, S. (2016). Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept. *British Journal of Sports Medicine*, 50(21):1309–1314.

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

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

Breheny, P., Zeng, Y., and Kurth, R. (2015). *grpreg: Regularization Paths for Regression Models with Grouped Covariates*. R package version 3.4.0, `https://CRAN.R-project.org/package=grpreg`.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133.

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

Carstensen, B. (2005). Demography and epidemiology: Practical use of the Lexis diagram in the computer age. In *Annual meeting of Finnish Statistical Society*, volume 23, page 24.

Carstensen, B., Plummer, M., Laara, E., and Hills, M. (2023). *Epi: Statistical Analysis in Epidemiology*. R package version 2.0.65, `https://CRAN.R-project.org/package=Epi`.

Casals, M., Fernández, J., Martínez, V., Lopez, M., Langohr, K., and Cortés, J. (2023). A systematic review of sport-related packages within the R CRAN repository. *International Journal of Sports Science & Coaching*, 18(2):621–629.

Casals, M. and Finch, C. F. (2017). Sports Biostatistician: a critical member of all sports science and medicine teams for injury prevention. *Injury Prevention*, 23(6):423–427.

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

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

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

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

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

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

Crossley, K. M., Patterson, B. E., Culvenor, A. G., Bruder, A. M., Mosler, A. B., and Mentiplay, B. F. (2020). Making football safer for women: a systematic review and meta-analysis of injury prevention programmes in 11 773 female football (soccer) players. *British Journal of Sports Medicine*, 54(18):1089–1098.

Danieli, C. and Abrahamowicz, M. (2019). Competing risks modeling of cumulative effects of time-varying drug exposures. *Statistical Methods in Medical Research*, 28(1):248–262.

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

Dorney, K., Dodington, J. M., Rees, C. A., Farrell, C. A., Hanson, H. R., Lyons, T. W., Lee, L. K., and for Kids®, I. F. C. (2020). Preventing injuries must be a priority to prevent disease in the twenty-first century. *Pediatric Research*, 87(2):282–292.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

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

Eilers, P. H. and Marx, B. D. (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, pages 89–102.

Ekstrand, J. (2013). Keeping your top players on the pitch: the key to football medicine at a professional level. *British Journal of Sports Medicine*, 47(12):723–724.

Ekstrand, J., Lundqvist, D., Davison, M., D'Hooghe, M., and Pensgaard, A. M. (2019). Communication quality between the medical team and the head coach/manager is as-

sociated with injury burden and player availability in elite football clubs. *British Journal of Sports Medicine*, 53(5):304–308.

Ekstrand, J., Lundqvist, D., Lagerbäck, L., Vouillamoz, M., Papadimitiou, N., and Karlsson, J. (2018). Is there a correlation between coaches' leadership styles and injuries in elite football teams? A study of 36 elite teams in 17 countries. *British Journal of Sports Medicine*, 52(8):527–531.

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

Finch, C. (2006). A new framework for research leading to sports injury prevention. *Journal of Science and Medicine in Sport*, 9(1-2):3–9.

Fiscutean, A. (2021). Data scientists are predicting sports injuries with an algorithm. *Nature*, 592(7852):S10–S11.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Friedman, M. et al. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1):101–113.

Fuller, C. W. (2018). Injury risk (burden), risk matrices and risk contours in team sports: a review of principles, practices and problems. *Sports Medicine*, 48(7):1597–1606.

Fuller, C. W., Bahr, R., Dick, R. W., and Meeuwisse, W. H. (2007). A framework for recording recurrences, reinjuries, and exacerbations in injury surveillance. *Clinical Journal of Sport Medicine*, 17:197–200.

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

Gabbett, T. J., Hulin, B. T., Blanch, P., and Whiteley, R. (2016). High training workloads alone do not cause sports injuries: how you get there is the real issue.

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

Gabbett, T. J., Whyte, D. G., Hartwig, T. B., Wescombe, H., and Naughton, G. A. (2014). The relationship between workloads, physical performance, injury and illness in adolescent male football players. *Sports Medicine*, 44:989–1003.

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

Gasparrini, A., Scheipl, F., Armstrong, B., and Kenward, M. G. (2017). A penalized framework for distributed lag non-linear models. *Biometrics*, 73(3):938–948.

Gerds, T. A. (2020). *Prediction Error Curves for Risk Prediction Models in Survival Analysis*. R package version 2020.11.17, `https://CRAN.R-project.org/package=pec`.

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

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

Griffin, A., Kenny, I. C., Comyns, T. M., and Lyons, M. (2020). The association between the acute: chronic workload ratio and injury and its application in team sports: a systematic review. *Sports Medicine*, 50:561–580.

Groll, A. (2016). *PenCoxFrail: Regularization in Cox Frailty Models*. R package version 1.0.1. https://cran.r-project.org/package=PenCoxFrail.

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

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

Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., and Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12):738–742.

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

Hastie, T., Friedman, J., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J., and Yang, J. (2010). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 4.1.7, `https://CRAN.R-project.org/package=glmnet`.

Hernán, M. A. and Robins, J. M. (2010). Causal Inference: What If.

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

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

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

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

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, pages 299–305.

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

Hulin, B. T., Gabbett, T. J., Blanch, P., Chapman, P., Bailey, D., and Orchard, J. W. (2014). Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers. *British Journal of Sports Medicine*, 48(8):708–712.

Hyman, J. M. (1983). Accurate monotonicity preserving cubic interpolation. *SIAM Journal on Scientific and Statistical Computing*, 4(4):645–654.

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

Impellizzeri, F. M., Shrier, I., McLaren, S. J., Coutts, A. J., McCall, A., Slattery, K., Jeffries, A. C., and Kalkhoven, J. T. (2023). Understanding training load as exposure and dose. *Sports Medicine*, pages 1–13.

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

Jackson, C. (2022). *msm: Multi-State Markov and Hidden Markov Models in Continuous Time*. R package version 1.6.9, https://CRAN.R-project.org/package=msm.

Jackson, C. H. (2011). Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, 38(8):1–29.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Kalkhoven, J. T., Watsford, M. L., Coutts, A. J., Edwards, W. B., and Impellizzeri, F. M. (2021). Training load and injury: causal pathways and future directions. *Sports Medicine*, 51:1137–1150.

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

Killen, N. M., Gabbett, T. J., and Jenkins, D. G. (2010). Training loads and incidence of injury during the preseason in professional rugby league players. *The Journal of Strength & Conditioning Research*, 24(8):2079–2084.

Kim, Y. and Lee, J. L. (2019). Common mistakes in statistical and methodological practices of sport management research. *Measurement in Physical Education and Exercise Science*, 23(4):314–324.

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

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lang, M., Bischl, B., and Surmann, D. (2023). *batchtools: Tools for Computation on Batch Systems*. R package version 0.9.16, https://CRAN.R-project.org/package=batchtools.

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

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

Lee, D.-J. and Zumeta-Olaskoaga, L. (2022). Can we really predict injuries in team sports? *Boletın de Estadıstica e Investigación Operativa BEIO*, page 149.

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

Li, X., Chang, C.-C. H., Donohue, J. M., and Krafty, R. T. (2022). A competing risks regression model for the association between time-varying opioid exposure and risk of overdose. *Statistical Methods in Medical Research*, 31(6):1013–1030.

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

Lolli, L., Batterham, A. M., Hawkins, R., Kelly, D. M., Strudwick, A. J., Thorpe, R., Gregson, W., and Atkinson, G. (2019). Mathematical coupling causes spurious correlation within the conventional acute-to-chronic workload ratio calculations. *British Journal of Sports Medicine*, 53(15):921–922.

Lutter, C., Jacquet, C., Verhagen, E., Seil, R., and Tischer, T. (2022). Does prevention pay off? Economic aspects of sports injury prevention: a systematic review. *British Journal of Sports Medicine*, 56(8):470–476.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall / CRC.

McCulloch, C. E., Searle, S., and Neuhaus, J. (2003). *Generalized, Linear, and Mixed Models*. John Wiley & Sons Ltd, Hoboken.

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

Meeuwisse, W. H. (1994). Assessing causation in sport injury: a multifactorial model. *Clinical Journal of Sport Medicine*, 4(3):166–170.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19.

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

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

Monasterio, X., Gil, S., Bidaurrazaga-Letona, I., Lekue, J. A., Diaz-Beitia, G., Santisteban, J. M., Lee, D.-J., Zumeta-Olaskoaga, L., Martin-Garetxana, I., and Larruskain, J. (2023a). Peak height velocity affects injury burden in circa-PHV soccer players. *International Journal of Sports Medicine*, pages 292–297.

Monasterio, X., Gil, S., Bidaurrazaga-Letona, I., Lekue, J. A., Santisteban, J. M., Diaz-Beitia, G., Lee, D.-J., Zumeta-Olaskoaga, L., Martin-Garetxana, I., Bikandi, E., et al. (2023b). The burden of injuries according to maturity status and timing: A two-decade study with 110 growth curves in an elite football academy. *European Journal of Sport Science*, 23(2):267–277.

Mork, D. and Wilson, A. (2022). Treed distributed lag nonlinear models. *Biostatistics*, 23(3):754–771.

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

Nevill, A. M., Holder, R. L., and Cooper, S.-M. (2007). Statistics, truth, and error reduction in sport and exercise sciences. *European Journal of Sport Science*, 7(1):9–14.

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

Nielsen, R. O., Chapman, C. M., Louis, W. R., Stovitz, S. D., Mansournia, M. A., Windt, J., Møller, M., Parner, E. T., Hulme, A., Bertelsen, M. L., et al. (2018). Seven sins when interpreting statistics in sports injury science. *British Journal of Sports Medicine*, 52(22):1410–1412.

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

Nielsen, R. Ø., Shrier, I., Casals, M., Nettel-Aguirre, A., Møller, M., Bolling, C., Bittencourt, N. F., Clarsen, B., Wedderkopp, N., Soligard, T., et al. (2020). Statement on methods in sport injury research from the first methods matter meeting, Copenhagen, 2019. *Journal of Orthopaedic & Sports Physical Therapy*, 50(5):226–233.

Obermeier, V., Scheipl, F., Heumann, C., Wassermann, J., and Küchenhoff, H. (2015). Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(2):395–412.

Omidpanah, A., Aragon, T. J., Fay, M. P., and Wollschlaeger, D. (2020). *epitools: Epidemiology Tools*. R package version 0.5-10.1, `https://CRAN.R-project.org/package=epitools`.

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

Pedersen, T. L. (2023). *patchwork: The Composer of Plots*. R package version 1.1.3, `https://CRAN.R-project.org/package=patchwork`.

Phillips, L. H. (2000). Sports injury incidence. *British Journal of Sports Medicine*, 34(2):133–136.

Preisser, J. S., Stamm, J. W., Long, D. L., and Kincade, M. E. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries research*, 46(4):413–423.

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

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramjith, J., Bender, A., Roes, K. C., and Jonker, M. A. (2022). Recurrent events analysis with piece-wise exponential additive mixed models. *Statistical Modelling*, 0(0).

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

Rommers, N., Rössler, R., Goossens, L., Vaeyens, R., Lenoir, M., Witvrouw, E., and D'Hondt, E. (2020). Risk of acute and overuse injuries in youth elite soccer players: body size and growth matter. *Journal of Science and Medicine in Sport*, 23(3):246–251.

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

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

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

Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., et al. (2021). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*, 55(2):118–122.

Sartori, S. (2011). *Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets*. PhD thesis, Università degli Studi di Milano.

Shrier, I. (2007). Understanding causal inference: the future direction in sports injury prevention. *Clinical Journal of Sport Medicine*, 17(3):220–224.

Shrier, I., Steele, R. J., Hanley, J., and Rich, B. (2009). Analyses of injury count data: some do's and don'ts. *American Journal of Epidemiology*, 170(10):1307–1315.

Signorello, L. B., McLaughlin, J. K., Lipworth, L., Friis, S., Sørensen, H. T., and Blot, W. J. (2002). Confounding by indication in epidemiologic studies of commonly used analgesics. *American Journal of Therapeutics*, 9(3):199–205.

Soligard, T., Schwellnus, M., Alonso, J.-M., Bahr, R., Clarsen, B., Dijkstra, H. P., Gabbett, T., Gleeson, M., Hägglund, M., Hutchinson, M. R., et al. (2016). How much is too

much?(Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. *British Journal of Sports Medicine*, 50(17):1030–1041.

Stevenson, M., Sergeant, E., Heuer, C., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., Sebastiani, P., Solymos, P., Yoshida, K., Jones, G., Pirikahu, S., Firestone, S., Kyle, R., Popp, J., Jay, M., Cheung, A., Singanallur, N., Szabo, A., and Rabiee, A. (2023). *epiR: Tools for the Analysis of Epidemiological Data*. R package version 2.0.65, https://CRAN.R-project.org/package=epiR.

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

Stovitz, S. D. and Shrier, I. (2012). Injury rates in team sport events: tackling challenges in assessing exposure time. *British Journal of Sports Medicine*, 46(14):960–963.

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

Sylvestre, M.-P. and Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine*, 28(27):3437–3453.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.2-7, https://CRAN.R-project.org/package=survival.

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

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

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

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

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

Van Mechelen, W., Hlobil, H., and Kemper, H. C. (1992). Incidence, severity, aetiology and prevention of sports injuries: a review of concepts. *Sports Medicine*, 14:82–99.

Vaughan, D. and Dancho, M. (2022). *furrr: Apply Mapping Functions in Parallel using Futures*. R package version 0.3.1, https://CRAN.R-project.org/package=furrr.

Waldén, M., Mountjoy, M., McCall, A., Serner, A., Massey, A., Tol, J. L., Bahr, R., D'Hooghe, M., Bittencourt, N., Della Villa, F., et al. (2023). Football-specific extension of the IOC consensus statement: methods for recording and reporting of epidemiological data on injury and illness in sport 2020. *British Journal of Sports Medicine*.

Wang, C., Vargas, J. T., Stokes, T., Steele, R., and Shrier, I. (2020). Analyzing activity and injury: lessons learned from the acute: chronic workload ratio. *Sports Medicine*, 50(7):1243–1254.

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

Wen, C., Zhang, A., Quan, S., and Wang, X. (2020a). *BeSS: An R Package for Best Subset Selection in Linear, Logistic and CoxPH Models*. R package version 1.0.6, https://CRAN.R-project.org/package=BeSS.

Wen, C., Zhang, A., Quan, S., and Wang, X. (2020b). BeSS: An R Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models. *Journal of Statistical Software*, 94(4):1–24.

Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 29(3):268–275.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H. and Bryan, J. (2023). *R packages*. O'Reilly Media, Inc. https://r-pkgs.org/.

Wickham, H., Hesselberth, J., Salmon, M., and RStudio (2022). *pkgdown: Make Static HTML Documentation for a Package*. R package version 1.1.3, https://CRAN.R-project.org/package=pkgdown.

Williams, S., West, S., Cross, M. J., and Stokes, K. A. (2017). Better way to determine the acute: chronic workload ratio? *British Journal of Sports Medicine*, 51(3):209–210.

Windt, J., Ardern, C. L., Gabbett, T. J., Khan, K. M., Cook, C. E., Sporer, B. C., and Zumbo, B. D. (2018). Getting the most out of intensive longitudinal data: a methodological review of workload–injury studies. *BMJ open*, 8(10):e022626.

Wood, S. (2023). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.8-41, https://CRAN.R-project.org/package=mgcv.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Yau, K. K., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Zumeta-Olaskoaga, L. and Lee, D.-J. (2023). *injurytools: A Toolkit for Sports Injury Data Analysis*. R package version 1.0.3, https://CRAN.R-project.org/package=injurytools.

Zumeta-Olaskoaga, L., Weigert, M., Larruskain, J., Bikandi, E., Setuain, I., Lekue, J., Küchenhoff, H., and Lee, D.-J. (2023). Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models. *AStA Advances in Statistical Analysis*, 107(1-2):101–126.

Appendices

Appendix A

A.1 Overall effects in zero-inflated negative binomial (ZINB) models

For simplicity, let us consider a binary variable X, e.g., $x_l = 1$ for player l having a previous injury, and $x_l = 0$, otherwise. Then, given the model in Eq. (2.4), the probability of an excess zero is modelled by logistic regression and expressed as,

$$p_l(x_l) = \frac{\exp(\gamma_0 + x_l\gamma_1)}{1 + \exp(\gamma_0 + x_l\gamma_1)}, \tag{A.1}$$

and the mean injury count for the at-risk players is modelled via a negative binomial model (i.e., log-linear model) as follows,

$$\lambda_l(x_l) = \exp(\beta_0 + x_l\beta_1).$$

The regression coefficient γ_1 in Eq. (A.1) represents the log odds ratio of having an excess zero or being in the not-at-risk group for the effect of $x_l = 1$ relative to $x_l = 0$. The coefficient β_1 represents the log of the incidence rate ratio (IRR) for the effect of $x_l = 1$ relative to $x_l = 0$ in the at-risk group, i.e. $\log(\lambda_l(x_l = 1)/\lambda_l(x_l = 0))$. But γ_1 and β_1 are not of primary interest, we seek to derive their contributions to the overall population effects.

The expected mean of a ZINB distributed variable Y is $\mathbb{E}(Y|X) = \lambda(x)(1-p(x))$. Thus, the overall incidence rate is derived as,

$$\begin{aligned}
\text{IR} := \mathbb{E}(Y|x_l) &= \lambda(x_l)(1 - p(x_l)) = \\
&= \exp(\beta_0 + x_l\beta_1)\left(1 - \frac{\exp(\gamma_0 + x_l\gamma_1)}{1 + \exp(\gamma_0 + x_l\gamma_1)}\right) = \\
&= \frac{\exp(\beta_0 + x_l\beta_1)}{1 + \exp(\gamma_0 + x_l\gamma_1)}.
\end{aligned}$$

Then, the IRR for the overall effect of x_l on injuries is,

$$\begin{aligned}\text{IRR} := \frac{\mathbb{E}\left(Y|(x_l=1)\right)}{\mathbb{E}\left(Y|(x_l=0)\right)} = \\ = \frac{\exp\left(\beta_0+\beta_1\right)}{1+\exp\left(\gamma_0+\gamma_1\right)} \Big/ \left(\frac{\exp\left(\beta_0\right)}{1+\exp\left(\gamma_0\right)}\right) = \\ = \exp\left(\beta_1\right)\frac{1+\exp\left(\gamma_0\right)}{1+\exp\left(\gamma_0+\gamma_1\right)}.\end{aligned}$$

A.2 Equivalence between the Cox and Poisson model

Let's assume a partition of the follow-up time into a finite number of intervals J, with $J+1$ cut points at $\kappa_0 < \ldots < \kappa_J$, and let's assume that the baseline hazard is piece-wise constant, i.e., it remains constant within each interval j: $\lambda_0(t) = \lambda_j$ for all $t \in (\kappa_{j-1}, \kappa_j]$. Then, following the formulation of the Cox PH model, the l-th individual's hazard function is expressed as,

$$\lambda(t;\boldsymbol{x}_l) = \lambda_0(t)\exp\big(\boldsymbol{x}_l'\boldsymbol{\beta}\big) := \lambda_j \exp\big(\boldsymbol{x}_l'\boldsymbol{\beta}\big) = \lambda_{lj}, \quad \forall t \in (\kappa_{j-1}, \kappa_j]. \tag{A.2}$$

Now, let $j(l)$ be the index of the interval for which $t_l \in (\kappa_{j-1}, \kappa_j]$, and let $\delta_{lj} \in \{0,1\}$ be the event indicator for individual l in interval j with $\delta_{lj(l)} = \delta_l$. Assuming model (A.2), the contribution of this individual l to the log-likelihood function, $\boldsymbol{l}(\cdot)$, is,

$$\begin{aligned}\boldsymbol{l}_l(\boldsymbol{\beta}) = \log\Big(f(t_l;\boldsymbol{x}_l)^{\delta_l} S(t_l;\boldsymbol{x}_l)^{1-\delta_l}\Big) = \log\Big(\lambda(t_l;\boldsymbol{x}_l)^{\delta_l} S(t_l;\boldsymbol{x}_l)\Big) = \\ = \delta_l \log\big(\lambda_{lj(l)}\big) - \sum_{j=1}^{j(l)} \lambda_{lj} t_{lj} = \\ = \sum_{j=1}^{j(l)} (\delta_{lj}\log\lambda_{lj} - \lambda_{lj}t_{lj}),\end{aligned} \tag{A.3}$$

where $f(t_l;\boldsymbol{x}_l) = \lambda(t_l;\boldsymbol{x}_l)S(t_l;\boldsymbol{x}_l)$, $S(t_l;\boldsymbol{x}_l) = \exp(-\Lambda(t_l;\boldsymbol{x}_l)) = \exp\Big(-\sum_{j=1}^{j(l)}\lambda_{lj}t_{lj}\Big)$, and $\delta_l\lambda_{lj(l)} = \sum_{j=1}^{j(l)}\delta_{lj}\lambda_{lj}$, since $\delta_{lj} = 0, \forall j \neq j(l)$.

On the other hand, if $\delta_{lj} \sim \text{Po}(\mu_{lj})$ with mean $\mu_{lj} = \lambda_{lj}t_{lj}$ and probability density function $f(\delta_{lj}) = \mu_{lj}^{\delta_{lj}}\exp(-\mu_{lj})/\delta_{lj}!$, where we can ignore the factorial since $\delta_{lj} \in \{0,1\}$, and thus, $\delta_{lj}! = 1$. It follows that the contribution of individual l to the Poisson log-

likelihood is given by,

$$\begin{aligned}
\boldsymbol{l}_l(\boldsymbol{\beta}) &= \log\left(\prod_{j=1}^{j(l)} f(\delta_{lj})\right) = \\
&= \sum_{j=1}^{j(l)} \log\left(\mu_{lj}^{\delta_{lj}} \exp(-\mu_{lj})\right) = \\
&= \sum_{j=1}^{j(l)} (\delta_{lj} \log(\mu_{lj}) - \mu_{lj}) = \\
&= \sum_{j=1}^{j(l)} (\delta_{lj} \log(\lambda_{lj}) + \delta_{lj} \log(t_{lj}) - \lambda_{lj} t_{lj}).
\end{aligned} \tag{A.4}$$

Therefore, the Poisson log-likelihood in Eq. (A.4) is proportional to Eq. (A.3), since the term $\delta_{lj} \log(t_{lj})$ is independent of the parameters of interest. Consequently, parameter estimates $\boldsymbol{\beta}$ can be obtained by optimizing the Poisson likelihood in Eq. (A.4).

Appendix B

B.1 Complementary information on functional screening tests data

Table B1 shows the variables included in functional screening tests data.

In these data, the definition used for the Limb Symmetry Index (LSI), which quantifies the inter-limb asymmetry or the discrepancies in strength, function and mobility of the legs, is the **Bilateral Strength Asymmetry** formula (Bishop et al., 2020; Impellizzeri et al., 2007):

$$\text{LSI} = \frac{|\text{Left leg} - \text{Right leg}|}{\max(\text{Left leg}, \text{Right leg})}.$$

An LSI value of 0 indicates total symmetry between both legs, while an LSI > 0 indicates asymmetry towards the leg for which the test value was higher. This definition is unique for each player and functional screening test. The associated effect of the variable can be easily interpreted in a statistical model.

B.2 Bootstrap .632+ estimates of the Brier score

In this section, we briefly describe the *bootstrap .632+ approach* used in the calculation of the Brier Score to avoid overfitting.

The bootstrap .632+ method was proposed in Efron (1983) and discussed in Efron and Tibshirani (1997). In the latter, they discussed the cross-validation and the bootstrap estimates of prediction error and showed that the bootstrap .632+ method substantially outperforms the cross-validation in their simulation experiments. Besides, the bootstrap .632+ method for the Brier score estimate was specifically introduced in Binder and Schumacher (2008) and demonstrated to provide accurate estimates.

Table B1: Detailed information on the 28 variables from the functional screening tests data, namely, the type of screening test and the number of tests for each type.

Type of screening test	**Number of tests**
Anthropometrics & previous injury	5
Height	
Weight	
Tibia + Femur length	
Sum of 6 skinfolds	
Previous injury	
Active Straight Leg Raise (ASLR)	2
ASLR lumbar strength LSI	
ASLR ROM LSI	
Cross Over Hop	3
Horizontal jumping distance LSI	
Horizontal jumping forces LSI	
Horizontal jumping impact forces LSI	
Core Strength Side	1
Core side plank LSI	
Drop Jump Kinetics	4
Drop jump impact forces 1st landing LSI	
Drop jump impact forces 2nd landing LSI	
Drop jump mechanical power LSI	
Drop jump vertical propulsion LSI	
Hand-Held Dynamometry	2
Hamstring strength isometric AKE LSI	
Hamstring strength isometric knee flexion 15º LSI	
Isokinetics	4
Isokinetic concentric knee extension 60º LSI	
Isokinetic concentric knee flexion 60º LSI	
Isokinetic isometric RTD knee flexion 30º LSI	
Isokinetic isometric RTD knee extension 90º LSI	
KT1000	1
Anterior-posterior knee laxity LSI	
Range of Motion (ROM)	4
Internal-external hip rotation ROM LSI	
Knee flexion ROM LSI	
Hip extension ROM LSI	
Ankle dorsiflexion ROM LSI	
Star Excursion Balance Test (SEBT)	2
SEBT knee extended LSI	
SEBT knee flexed LSI	
TOTAL	28

Namely, the bootstrap .632+ estimate of the Brier score –the prediction error curve– is a weighted linear combination of the apparent estimate, the bootstrap cross-validation estimate and the no information estimate. It is implemented in the **pec R** package (Gerds, 2020), and the following definition is given by them:

$$\text{Boot632plusErr}(t,\hat{S}) = \left(1 - \frac{0.632}{1-0.368\cdot\omega}\right)\text{AppErr}(t,\hat{S}) + \left(\frac{0.632}{1-0.368\cdot\omega}\right)\text{BootCvErr}(t,\hat{S}),$$

where

$$\omega = \frac{\min(\text{BootCvErr}(t,\hat{S}),\text{NoInfErr}(t,\hat{S})) - \text{AppErr}(t,\hat{S})}{\text{NoInfErr}(t,\hat{S}) - \text{AppErr}(t,\hat{S})}.$$

The constant 0.632 is independent of the sample size and corresponds to the probability of drawing with replacement subject i into the bootstrap sample: $P(\{(Y_i, X_i)\} \in D_b) = 1 - (1 - 1/N)^N \approx (1 - e^{-1}) \approx 0.632$.

We refer the reader to Mogensen et al. (2012) and the **R** help page of `pec::pec()` function for the definitions of the apparent, the bootstrap cross-validation and the no information estimate, that is, $\text{AppErr}(t,\hat{S})$, $\text{BootCvErr}(t,\hat{S})$ and $\text{NoInfErr}(t,\hat{S})$, respectively.

B.3 Simulation study

Data-generating process

We base on the method proposed by Harden and Kropko (2019), referred to as the "random spline method", for simulating Cox data. A function in the **coxed R** package implements the method with several options for user control. In this work, we modify the method to allow the inclusion of a frailty term.

Below, we describe the procedure used to generate survival times, i.e. the data-generating process (DGP), where we know and have control of the correct specification of covariates and true values of the coefficients. We assume that the DGP is given by a stochastic process that aligns with a Cox shared frailty model. First, we describe the generation of the baseline hazard function, and then, we outline the second part of generating individual durations.

Step 1 Generating the baseline hazard function

This is the crucial part of the "random spline method" since it addresses the challenge of the shape of the baseline hazard to be unspecified (no parametric). The procedure is:

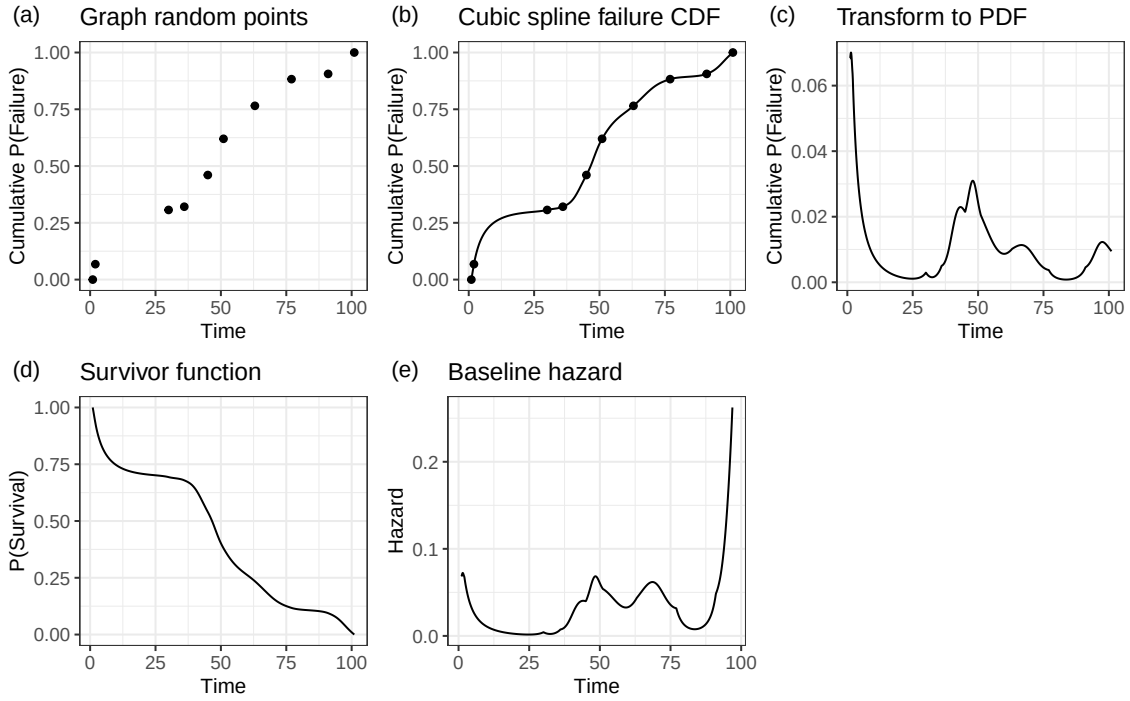


Figure B1: *Reproduced from Harden and Kropko (2019)*. Generating a baseline hazard function via the random spline method. (a) An example of the randomly drawn time points. (b) The cubic spline fit to those points to create the failure cumulative distribution function (CDF). (c) The transformation from the failure CDF to a failure probability density function (PDF). (d) Plots the survivor function and (e) graphs the baseline hazard function

- Create a time index of length $T_{\max}$ and draw k points, where $k << T_{\max}$. For illustration purposes, we choose integers from 1 to 100 ($T_{\max} = 100$), and draw $k = 10$ points, $(x_1, y_1), \ldots, (x_10, y_10)$. The x-coordinates for two of the k points are set as the minimum and maximum of the time index, e.g. $x_1 = 1$ and $x_{10} = 100$. Then, randomly draw the remaining $k-2$ points from the remaining time points with uniform probability. Set the y-coordinates at the minimum time to be 0 and at the maximum time to be 1. Randomly draw the other $k - 2$ y-coordinates from a $\mathcal{U}(0, 1)$. Finally, sort the coordinates in ascending order, as the cumulative distribution function (CDF) must be non-decreasing, resulting in the order $(x_1 = 0, y_1 = 0), (x_2, y_2), \ldots, (x_{10} = 100, y_{10} = 1)$. See Figure B1, panel (a).
- Construct the cumulative distribution function (CDF) for event occurrences by fitting the previously drawn k points with a cubic smoothing spline. The smoothing function presented in Hyman (1983) is used to preserve the mono-

tonicity (Figure B1, panel (b)).

- Transform the CDF function into the baseline hazard function. To do so, first, construct the probability density function (PDF) for failure times by computing the first differences of the CDF at each time point (see Figure B1, panel (c)). Generate the survivor function by subtracting the failure CDF from 1 (see Figure B1, panel (d)). Finally, compute the baseline hazard by dividing the failure PDF by the survivor function (see Figure B1, panel (e)).

Step 2 Generating individual durations

Once the baseline hazard function is generated, individual survival times are drawn in a way that depends on user-controlled covariates and coefficient values.

- p covariates are set, either randomly or as specified by the user, forming $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_p})_{\text{Nxp}}$. Also, true values for p coefficients are defined as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. We set frailty terms for each player –family of observations, or cluster of the data. These terms, denoted as $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$, follow a Gamma distribution with shape and scale parameters equal to one and a tenth, i.e. $\boldsymbol{\alpha}\, \Gamma(k = 1, \theta = \frac{1}{10})$. Then, we set the linear predictor vector as $\boldsymbol{\alpha} \exp(\boldsymbol{X}'\boldsymbol{\beta})$.
- If the hazard of the i-th injury event for the l-th player at time t is expressed as,

$$\lambda_{i_l}(t) = \alpha_l \lambda_0(t) \exp\big(\boldsymbol{X}'_{i_l}\boldsymbol{\beta}\big),$$

we have that, in terms of survival probability,

$$S_{i_l}(t) = \exp(-\Lambda_{i_l}(t)) = \exp\left(-\int_0^t \lambda_0(s)\alpha_k \exp\big(\boldsymbol{X}'_{i_l}\boldsymbol{\beta}\big)ds\right) =$$
$$= \exp\left(-\int_0^t \lambda_0(s)ds\right)^{\alpha_k \exp\left(\boldsymbol{X}'_{i_l}\boldsymbol{\beta}\right)} = S_0(t)^{\alpha_k \exp\left(\boldsymbol{X}'_{i_l}\boldsymbol{\beta}\right)}.$$

- Once we construct the true individual-specific survival function, we generate survival times by drawing $\mathcal{U}[0, 1]$. For all N observations, we determine the time point at which each individual observation's survival function becomes less than this randomly drawn uniform value, see Figure B2.
- Lastly, we censor some of the observations by randomly selecting, with a uniform or other distribution, observations to be censored (this conforms to the Cox model's assumption that, conditional on the covariates, the censoring mechanism is independent of the DGP that produces the durations).

  Note: if the survival time drawn is equal to $T_{\max}$ we directly censor this observation, keeping control of the censorship percentage.

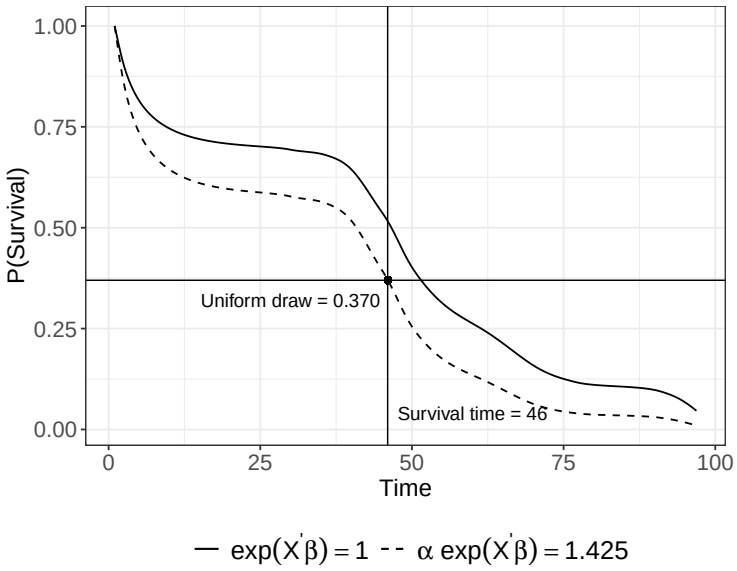


Figure B2: *Reproduced from Harden and Kropko (2019)*. Drawing a duration for an observation from the simulated survivor function. The solid line is the baseline survivor function, which represents the survival probability for an observation with a 0 for every covariate. The dashed line is the survivor function for an observation whose powered linear predictor is 1.425. This observation has a risk of failure at time t conditional on survival through time t that is 42.5 per cent higher than the baseline.

Further simulation results

The source code to reproduce the simulation study, as well as the following results, can be found at: `https://github.com/lzumeta/TimeToEvent-InjurySim`. In the following, we present additional results from the simulation study in tabular and graphical forms:

- In Table B2, we report the results of the last setting within Scenarios 2 and 3 ($N_{\text{obs}} = 670$).
- In Tables B3-B5, we report additional performance measures to quantify the uncertainty of the simulation estimates.
- In Figures B3-B4, we show complementary graphical results from Scenario 2 and Scenario 3 of the simulation study.

Table B2: Simulation results for Scenarios 2 and 3, which consider different correlation structures of covariates, $\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$, for $L = 220$ players resulting in $N_{\text{obs}} = 670$ observations. The measures analyzed include the AMS, ANFS, ANFNS, MSE and the median of IBS calculated over the $[0, 1000]$ and $[0, 3500]$ time intervals, for all models.

Sample size (N_{obs})	**Correlation structure** ($i \neq j$)	**Frailty model including vars. that selected**	**AMS** (5)	**ANFS** (0)	**ANFNS** (0)	**MSE** (0)	**IBS** (0)	
							[0,1000]	**[0, 3500]**
$N_{\text{obs}} = 670$	$\rho_{ij} = 0$	BeSS	**2.94**	**0.66**	2.72	**0.52**	**0.033**	0.109
		Lasso	12.41	8.66	1.25	0.77	**0.033**	**0.107**
		Elastic Net	14.18	10.34	1.16	0.79	**0.033**	0.108
		Ridge	7.53	4.24	1.71	0.65	**0.033**	0.108
		Group Lasso	21.1	16.1	**0**	0.80	0.034	0.110
		Cox Boosting	7.36	4.28	1.92	0.67	**0.033**	**0.107**
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	**3.03**	**1.52**	2.69	0.88	0.037	0.104
		Lasso	15.41	11.15	0.74	0.90	**0.036**	0.102
		Elastic Net	17.22	12.70	0.48	0.91	**0.036**	0.102
		Ridge	9.26	4.91	0.65	**0.79**	**0.036**	0.102
		Group Lasso	32.15	27.15	**0**	1.44	0.037	0.107
		Cox Boosting	10.53	6.89	1.36	0.90	**0.036**	**0.101**

Additional performance measures

In addition to the mean square error (MSE) reported in Chapter 3, we introduce two additional performance measures, Bias and empirical standard error, to provide a better interpretation of the MSE measure.

The MSE is calculated as,

$$\text{MSE} = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \sum_{j=1}^{p} \left(\hat{\beta}_j^{(n)} - \beta_j \right)^2,$$

the bias as,

$$\text{Bias} = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \sum_{j=1}^{p} \left(\hat{\beta}_j^{(n)} - \beta_j \right),$$

and the empirical standard error (EmpSE) as the long-run standard deviation of $\widehat{\boldsymbol{\beta}}$ over the N_{sim} repetitions. That is,

$$\text{EmpSE} = \sqrt{\frac{1}{N_{\text{sim}} - 1} \sum_{n=1}^{N_{\text{sim}}} \sum_{j=1}^{p} \left(\hat{\beta}_j^{(n)} - \overline{\hat{\beta}}_j \right)^2},$$

where,

$$\overline{\hat{\beta}}_j = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \hat{\beta}_j^{(n)}, \qquad \text{for each } \; j = 1, \ldots, p.$$

The EmpSE depends only on the estimates $\hat{\beta}_j$.

Table B3: Additional simulation results for the three different settings within Scenario 1, for 66 players with 3 observations each, resulting in a sample size of 198.

Model	**Bias**	**EmpSE**	**MSE**
True model: frailty (*BeSS*)			
BeSS	-0.07	1.48	3.85
Lasso	0.22	1.57	4.29
Elastic Net	0.24	1.72	4.84
Ridge	0.04	**1.42**	**3.78**
Group Lasso	**-0.03**	2.70	9.24
Boosting	0.23	1.57	4.29
True model: frailty (*Lasso*)			
BeSS	**-0.36**	1.76	5.05
Lasso	-0.42	1.96	5.67
Elastic Net	-0.52	2.10	6.26
Ridge	-0.42	**1.65**	**4.66**
Group Lasso	-1.31	3.58	14.57
Boosting	-0.41	1.96	5.65
True model: frailty (*Boosting*)			
BeSS	0.17	1.69	**5.24**
Lasso	0.23	1.85	6.06
Elastic Net	**0.16**	2.12	7.15
Ridge	0.23	**1.69**	5.50
Group Lasso	0.36	3.10	12.38
Boosting	0.22	1.88	6.17

Table B4: Additional simulation results for Scenarios 2 and 3, which consider different correlation structures of covariates ($\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$) for a varying number of players $L \in \{22, 66\}$ resulting in $N_{\text{obs}} \in \{60, 191\}$ observations, respectively.

Sample size (N_{obs})	**Correlation structure** ($i \neq j$)	**Frailty model including vars. that selected**	**Bias**	**EmpSE**	**MSE**
$N_{\text{obs}} = 60$	$\rho_{ij} = 0$	BeSS	-0.61	**1.63**	**3.69**
		Lasso	-1.08	6.32	56.37
		Elastic Net	1.61	14.73	271.34
		Ridge	**-0.22**	46.69	57.02
		Group Lasso	-28.34	113.52	$> 10^5$
		Cox Boosting	-0.64	5.47	42.77
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	**-0.14**	1.87	4.90
		Lasso	-0.37	**1.34**	**2.72**
		Elastic Net	-3.72	46.29	2758.8
		Ridge	1.42	23.54	776.0
		Group Lasso	-21.55	$> 10^5$	$> 10^5$
		Cox Boosting	-0.37	1.38	2.92
$N_{\text{obs}} = 191$	$\rho_{ij} = 0$	BeSS	-0.35	**0.72**	**0.96**
		Lasso	-0.37	0.81	1.10
		Elastic Net	-0.33	0.86	1.21
		Ridge	**-0.15**	0.85	1.21
		Group Lasso	0.36	1.64	3.39
		Cox Boosting	-0.25	0.84	1.18
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	**-0.12**	**0.89**	**1.28**
		Lasso	-0.48	1.03	1.54
		Elastic Net	-0.61	1.11	1.78
		Ridge	-0.40	1.01	1.50
		Group Lasso	-182.56	7068.27	$> 10^5$
		Cox Boosting	-0.39	1.01	1.48

Table B5: Additional simulation results for Scenarios 2 and 3, which consider different correlation structures of covariates ($\rho_{ij} = 0$ and $\rho_{ij} = 0.65^{|i-j|}$) for a varying number of players $L \in \{132, 220\}$ resulting in $N_{\text{obs}} \in \{391, 670\}$ observations, respectively.

Sample size (N_{obs})	**Correlation structure** ($i \neq j$)	**Frailty model including vars. that selected**	**Bias**	**EmpSE**	**MSE**
N_{obs} = 391	$\rho_{ij} = 0$	BeSS	**0.13**	**0.58**	**0.57**
		Lasso	0.29	0.74	0.82
		Elastic Net	0.46	0.78	0.89
		Ridge	0.39	0.69	0.74
		Group Lasso	0.74	0.89	1.09
		Cox Boosting	0.33	0.72	0.79
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	**-0.16**	**0.77**	1.02
		Lasso	-0.64	0.89	1.18
		Elastic Net	-0.71	0.94	1.27
		Ridge	-0.36	0.81	**0.99**
		Group Lasso	-0.35	1.40	2.40
		Cox Boosting	-0.51	0.86	1.12
N_{obs} = 670	$\rho_{ij} = 0$	BeSS	0.10	**0.53**	**0.54**
		Lasso	0.17	0.70	0.77
		Elastic Net	0.23	0.71	0.79
		Ridge	**0.09**	0.63	0.65
		Group Lasso	0.41	0.74	0.80
		Cox Boosting	0.15	0.64	0.67
	$\rho_{ij} = 0.65^{\|i-j\|}$	BeSS	**-0.08**	0.72	0.88
		Lasso	-0.76	0.77	0.90
		Elastic Net	-0.75	0.78	0.91
		Ridge	-0.34	**0.70**	**0.79**
		Group Lasso	-0.45	1.09	1.44
		Cox Boosting	-0.56	0.76	0.90

Figure B3: Heatmaps of variables correctly selected (in green) and wrongly selected (in red) across different simulation settings. Columns: Scenario 2 (left) and Scenario 3 (right). Rows: setting 1 (first row), setting 2 (second row), setting 3 (third row) and setting 4 (fourth row).

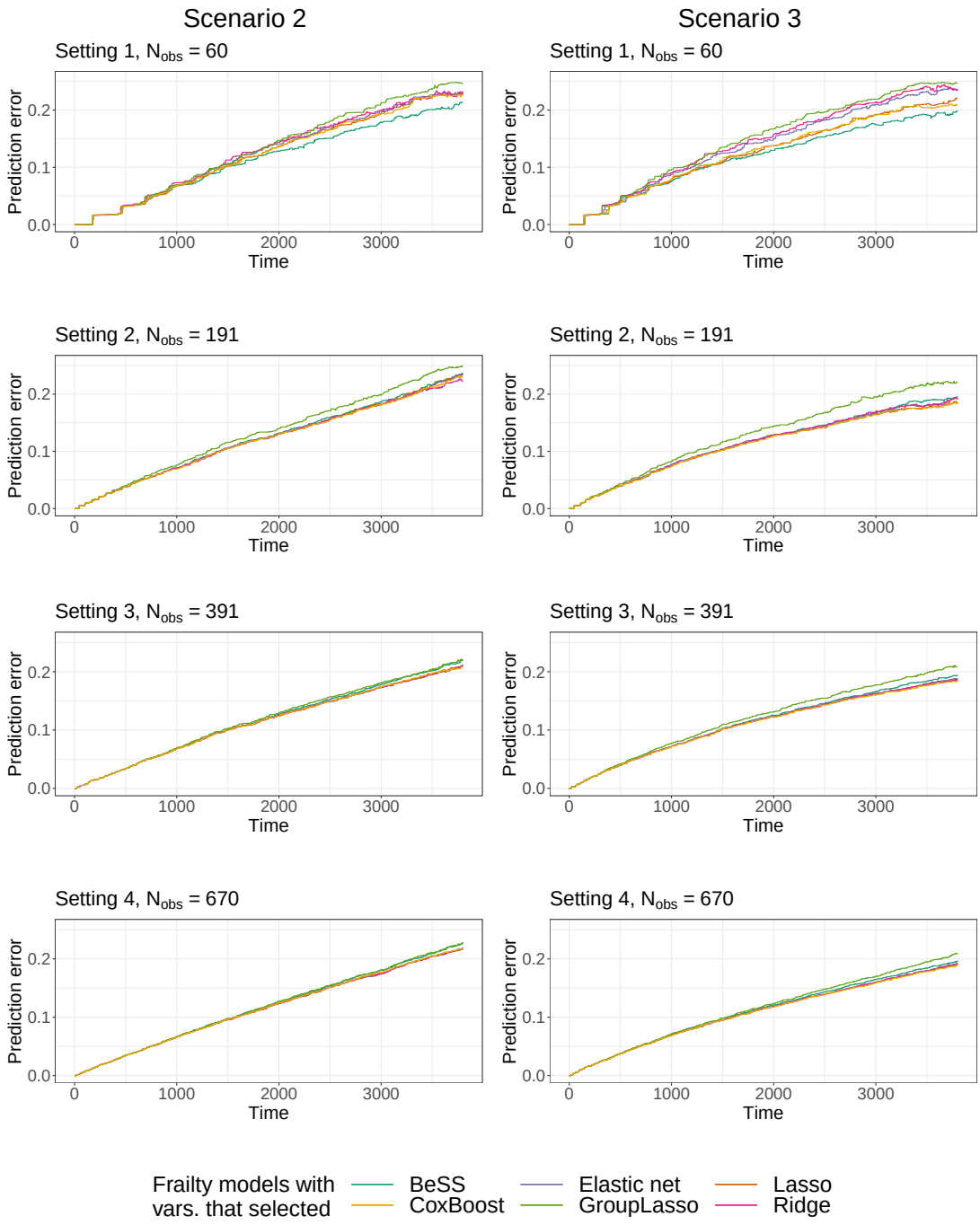


Figure B4: The point-wise averages of the Brier Score curve estimates using the bootstrap .632+ method (the point-wise average prediction error) for frailty models with variables that selected BeSS, Lasso, Elastic net, Ridge regression, Group Lasso and Cox Boosting methods, across different simulation settings. Columns: Scenario 2 (left) and Scenario 3 (right). Rows: setting 1 (first row), setting 2 (second row), setting 3 (third row) and setting 4 (fourth row).

Appendix C

In this appendix, we present additional information and results that, for brevity, were not included in Chapter 4. In Section C.1, we give a more detailed explanation of the simulation study, outlining the data generation process and all the scenarios considered; whereas in Section C.2, we provide supplementary analyses conducted on the football injury data.

C.1 Simulation study

To evaluate the performance of the three models, that is, PAMMs with WCE-type cumulative effects: with no constraint (*Uncons.*), adding a constraint (*Constr.*) and adding a ridge penalty (*Ridge*); we simulate $N_{\text{sim}} = 500$ times a cohort of $L = 500$ individuals with exposures recorded at $t_{z,1} = 1, t_{z,2} = 2, \ldots, t_{z,Q=40} = 40$ days before the time at which we model the hazard, $\boldsymbol{z}_l(t) = (z_l(t_{z,1}), z(t_{z,2}), \ \ldots, z(t_{z,Q}))$. Individuals' follow-up starts after 40 days of exposure, such that every individual has a complete exposure history of 40 exposures at the beginning of the follow-up.

In the following, we describe the scenario settings that were kept fixed across the simulation runs (section C.1.1), the data generation part that was random (section C.1.2) and the performance measures used to evaluate the results (section C.1.4). Finally, we present the simulation results we obtained (section C.1.5). The **R** code to reproduce these analyses is available at: https://github.com/lzumeta/flex-mod-training-loads-recu-injuries.

C.1.1 Simulation scenarios

We set six different true weight functions for the WCE-type –meaning, partial effects of $h(t, t_z, z(t_z)) = h(t - t_z)z(t_z)$ type– cumulative effects, **(a)-(f)**, each defined over a $[0, Q]$ interval, and under three different levels of heterogeneity between recurrent events, $\sigma_b \in \{0.05, 0.5, 1\}$, indicating very low heterogeneity, low heterogeneity and high hetero-

geneity, respectively.

For the true weight functions of the WCE-type cumulative effect, we stand on and adapt the simulation setting presented in Sylvestre and Abrahamowicz (2009) for cumulative effects of time-dependent exposures in Cox's PH model, and set the following six true weight functions (the last two true weight functions not shown in the main work):

(a) *Exponential decay:* $h(t-t_z) = \frac{4.5}{100} e^{-\frac{1}{10}(t-t_z)}$.

(b) *Bi-linear*: $h(t-t_z) = \left(1 - \frac{t-t_z}{25}\right) * 0.04$ for $t - t_z \leq 25$ and 0 otherwise.

(c) *Early peak*: probability density function of $N(0.04; 0.06)$ left-truncated at $t = 0$.

(d) *Inverted U*: probability density function of $N(0.2; 0.06)$ left-truncated at $t = 0$.

(e) *Constant*: $h(t-t_z) = 0.02$ for $0 \leq t - t_z \leq 20$ and 0 otherwise.

(f) *Hat*: $h(t-t_z) = \begin{cases} \text{increasing}, & \text{if } t - t_z \leq 19 \\ \text{plateau}, & \text{if } 19 < t - t_z \leq 22 \\ \text{decreasing}, & \text{if } 22 < t - t_z \leq 27 \\ 0, & \text{otherwise.} \end{cases}$

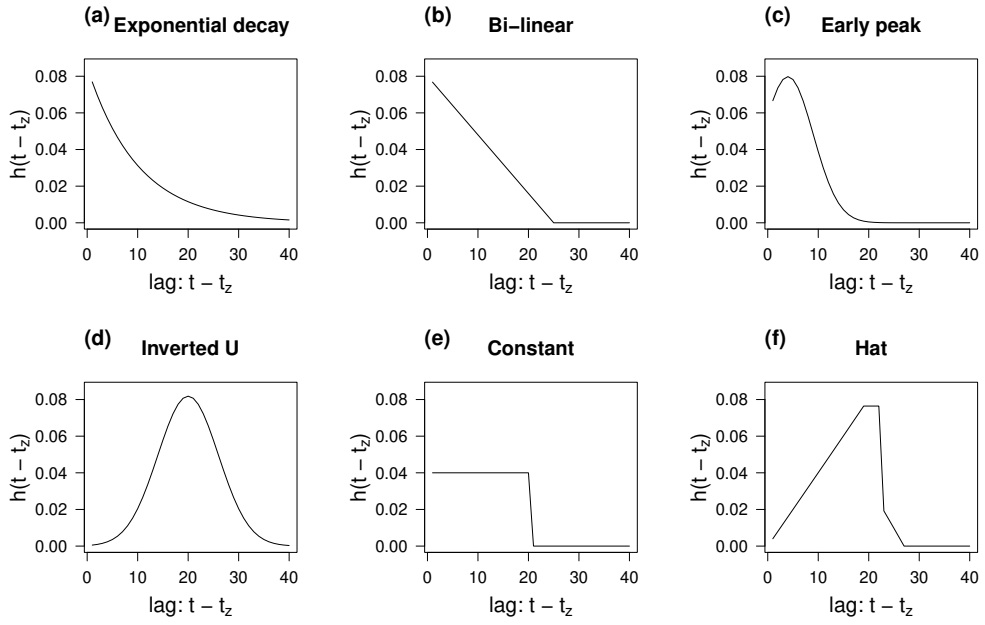


Figure C1: Each of the true weight functions, **(a)**-**(f)**, considered.

Each function assigns weights to past exposures based on the time elapsed since the exposure occurred (see Figure C1). The first two functions, **(a)** and **(b)**, assume that

weights decrease monotonically as the time elapsed since exposure was recorded increases. Functions **(c)** and **(d)** are non-monotonic where the weights first increase and then decrease. In contrast to the function **(d)** (*inverted U*), in function **(c)** (*early peak*), the maximum weight is assigned to more recent exposures and from the 20-th lag on, the weights are zero. In other words, exposures recorded over 20 or more time units ago have no impact at the current time. The function **(e)** assigns *constant* weights to past exposures and again, from the 20-th lag on the resulting exposure effect disappears. Its resulting WCE-type cumulative effect corresponds to a standard unweighted cumulative sum of the time-varying exposure variable calculated over the previous 20 units of time. The last function **(f)** is specifically designed as an extreme case to evaluate the ridge penalization on the basis coefficients.

On the other hand, the number of events per subject, n_l, is kept fixed across all simulation runs. It is drawn from a truncated Poisson distribution with a lower truncation point equal to zero, so that we condition the variable, e.g. Y, to be $Y > 0$. The vector b_l associated with each individual l is drawn from a Gaussian distribution with a mean of 0 and a standard deviation $\sigma_b \in \{0.05, 0.5, 1\}$, and it is also kept fixed across all simulation runs and all true weight functions.

In this manner, we construct the underlying true hazard value, $\lambda_{i_l j}$ ($l = 1, \ldots, L = 500,\ i_l = 1, \ldots, n_l$ and $j = 1, \ldots, J = 40$), for each event of each subject (i_l) at each time point (κ_j). This value defines the intervals of the piece-wise constant hazard vector and remains fixed in each scenario of the simulation study. Each subject's hazard value can be seen as the sum of: an intercept, a smooth baseline, the cumulative effect and the random effect.

C.1.2 Generation of recurrent survival times

Once we have the vector of piece-wise constant hazards $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{J=40})$ in intervals defined by time points $\boldsymbol{\kappa} = (\kappa_0, \ldots, \kappa_{J=40})$, we replicate the rows of the data depending on the number of events that each individual is at risk of (we kept this number, i.e. n_l, fixed), in order to generate recurrent survival times. Then, we draw survival times from the piece-wise exponential distribution, i.e. $t \sim \text{PEXP}(\boldsymbol{\lambda}, \boldsymbol{\kappa})$ for which the algorithm is summarized in Table C1. The main function to draw piece-wise constant rates is implemented in **R** as `msm::rpexp()`.

Table C1: Pseudo-algorithm for drawing survival times from the piece-wise exponential distribution (PEXP), adapted from Bender et al. (2019).

Let κ_{j-1} be the left border of interval $(\kappa_{j-1}, \kappa_j],\ j = 1, \ldots, J$:

1. For $l = 1, \ldots L$:
   - 1.1. For $i_l = 1, \ldots, n_l$:
     - (a) Set $j = 1$
     - (b) For $j = 1, \ldots, J$
       - i. Draw survival time $t'_{i_l j}$ from Exp($\lambda_{i_l j}$), set $t_{i_l} = \kappa_{j-1} + t'_{i_l j}$
       - ii. If $\kappa_{j-1} < t_{i_l} \leq \kappa_j$ or $j = J$: accept t_{i_l}
       - iii. Else: $j = j + 1$
   - 1.2. Return vector of survival times $(t_1, \ldots, t_{n_l})$ for subject l.
   - 1.3. Order the above survival times vector and the subject's i_l-th event number indicator(`enum`).
2. Return vector of survival times:
   $(t_{1_1}, t_{2_1} \ldots, t_{n_1}, t_{1_2}, t_{2_2}, \ldots, t_{n_2}, \ldots, t_{1_L}, t_{2_L}, \ldots, t_{n_L})$.

C.1.3 Model fitting

For model fitting, we use P-splines with second-order difference penalties and, 10 knots for the smooth baseline hazard term and 15 knots for the WCE-type smooth term. We use the restricted maximum likelihood (REML) optimization routine within the **mgcv R** package.

C.1.4 Performance measures

The performance measures we use to assess the performance of the models are the mean RMSE, mean 95% pointwise coverage, squared error, BIC and the deviance explained.

- The mean RMSE integrates the bias and the variance of $\hat{h}$ into one summary measure. The root of the sum of squared differences between the estimated $\hat{h}$ value and the true h value, computed across all covariates z and $t - t_z$ lag points, and then averaged over all the simulation runs.

$$\overline{\text{RMSE}}(h) = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \sqrt{\frac{1}{N_{t_z}} \sum_{t-t_z=0}^{40} \left(h(t - t_z) - \hat{h}(t - t_z)^{(n)} \right)^2},$$

where $N_{t_z} = 41$, since $t - t_z = \{0, 1, 2, \ldots, 40\}$ takes 40 +1 number of different values.

$$\text{RMSE}(\sigma_b) = \sqrt{\frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \left(\sigma_b - \hat{\sigma}_b^{(n)}\right)^2},$$

- The mean 95% coverage measures the proportion that the estimated pointwise 95% confidence interval contains h. It is calculated as:

$$\overline{\text{Coverage}}_\alpha = \frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \left[\frac{1}{N_{t_z}} \sum_{t-t_z=0}^{40} \mathbb{I}\left(h(t-t_z) \in \left[\hat{h}(t-t_z)^{(n)} \mp \zeta_{1-\alpha/2} \hat{\sigma}_{\hat{h}}^{(n)}\right]\right)\right],$$

  where ζ_q is the q-quantile of the standard normal distribution and $\hat{\sigma}_{\hat{h}}$ the standard error of the estimated $\hat{h}(t - t_z)$. The closer to 0.95 the better.

- The squared error of σ_b is the summand of RMSE(σ_b), i.e. $(\sigma_b - \hat{\sigma}_b)^2$.

- The BIC and Deviance Explained are likelihood-based measures formulated to assess the model's goodness of fit. As described in `?mgcv::gamObject`, BIC = $k \ln(n) - 2 \ln(\mathcal{L}(\text{model}|\text{data}))$ and deviance explained = (1−residual deviance/null deviance). A smaller BIC indicates better performance, while a higher Deviance Explained suggests a better fit.

C.1.5 Simulation results

Next, we present the simulation study results:

- Figures C2-C7 display the estimated weight functions for each model across all simulation settings, with the true weight function represented by the thick black curve and the mean of the estimated curves depicted by the thick red curve.

- Tables C2-C3 present summary statistics for mean RMSE and mean Coverage$_\alpha$ of the estimated $h(t, t_z, z(t_z))$ and σ_b across all simulation settings.

- Figures C8-C10 show boxplots of the distribution of the RMSE of $h(t, t_z, z(t_z)$, the distribution of 95% point-wise coverage of $h(t, t_z, z(t_z))$ and the squared error of σ_b across all simulation settings.

- Table C4 displays the proportion of times each model is considered the "best", determined by either BIC or Deviance Explained in each simulation run across all settings. This represents the number of times a specific model has the lowest BIC or the greatest Deviance Explained among the three candidate models.

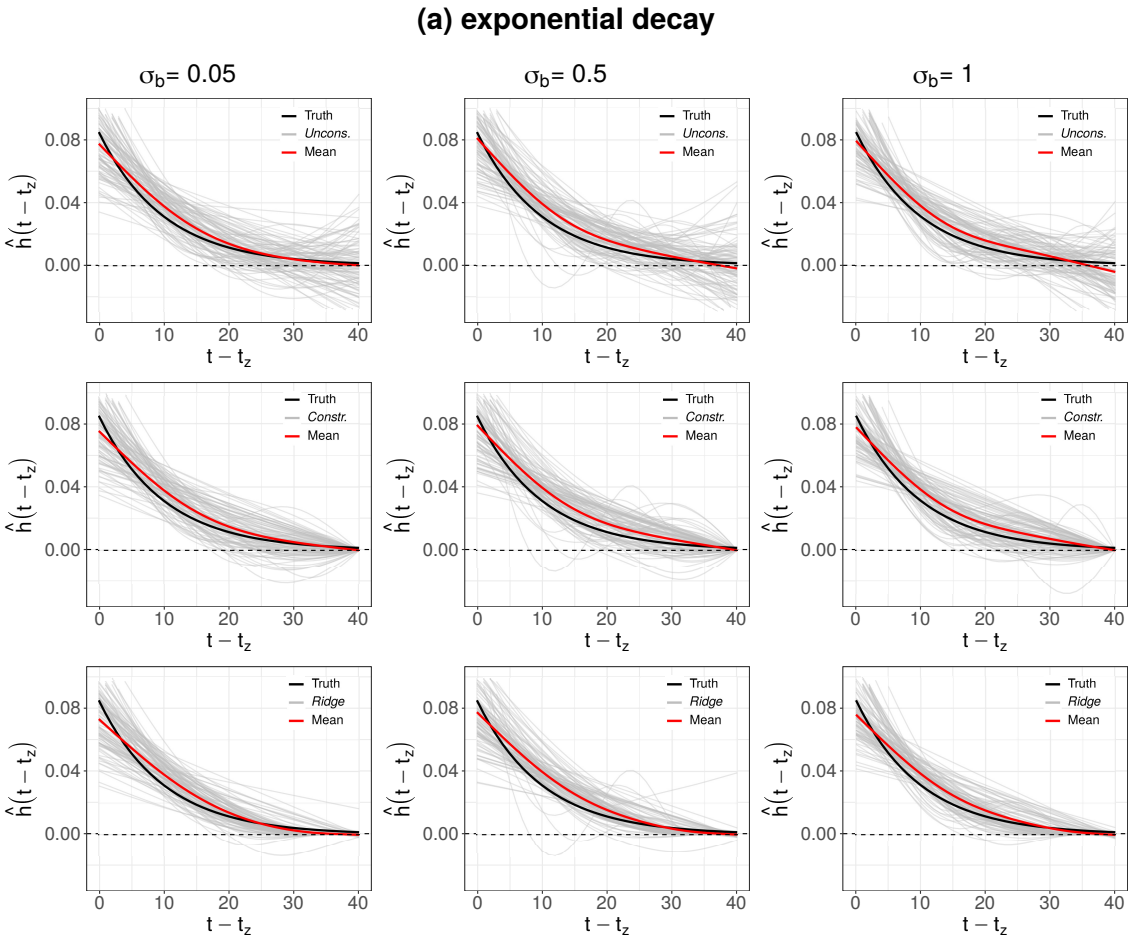


Figure C2: A random sample of 100 estimated weight functions (in grey) for the *Uncons.* model (1st row), *Constr.* (2nd row) and *Ridge* model (3rd row), for scenarios $\sigma_b = 0.05$ (left), $\sigma_b = 0.5$ (middle) and $\sigma_b = 1$ (right), together with the true weight function (in black), shape (**a**), and the mean curve (in red).

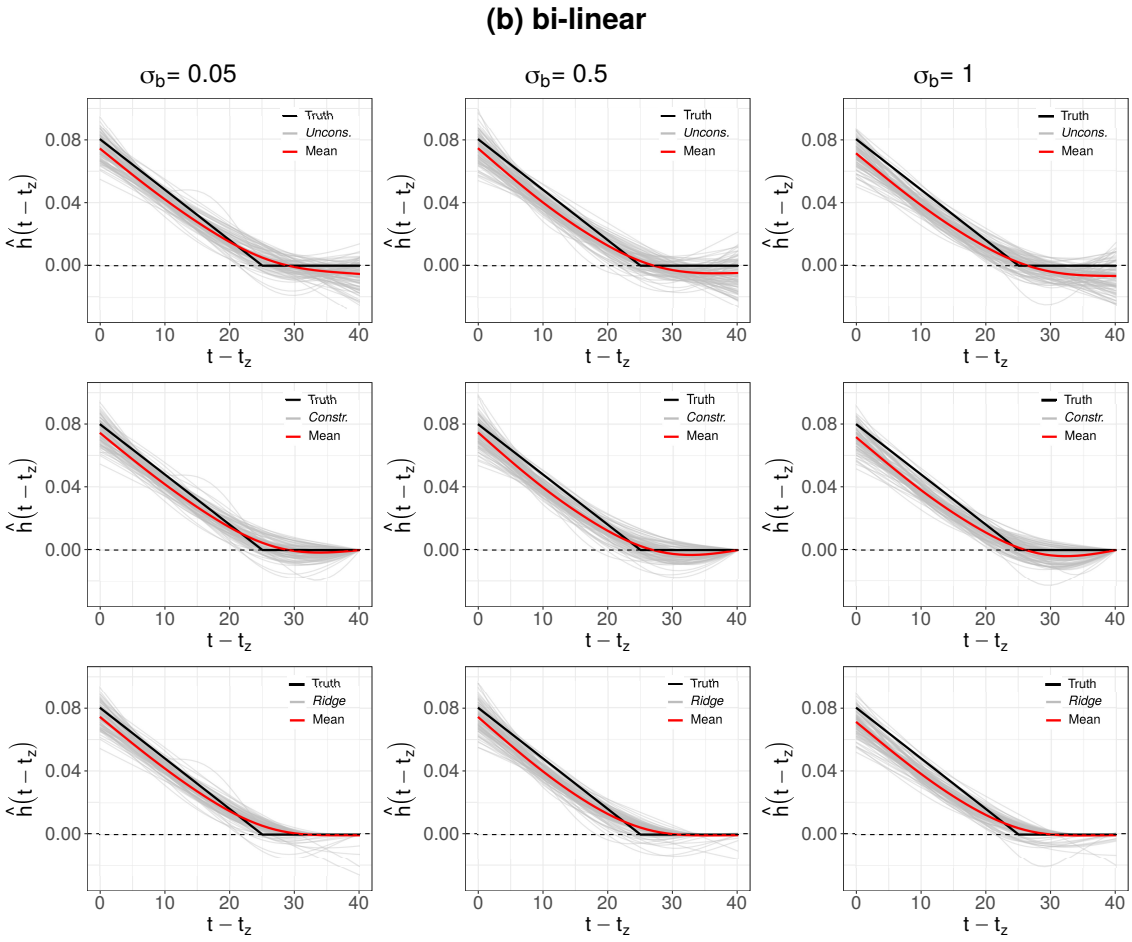


Figure C3: A random sample of 100 estimated weight functions (in grey) for the *Uncons.* model (1st row), *Constr.* (2nd row) and *Ridge* model (3rd row), for scenarios $\sigma_b = 0.05$ (left), $\sigma_b = 0.5$ (middle) and $\sigma_b = 1$ (right), together with the true weight function (in black), shape (**b**), and the mean curve (in red).

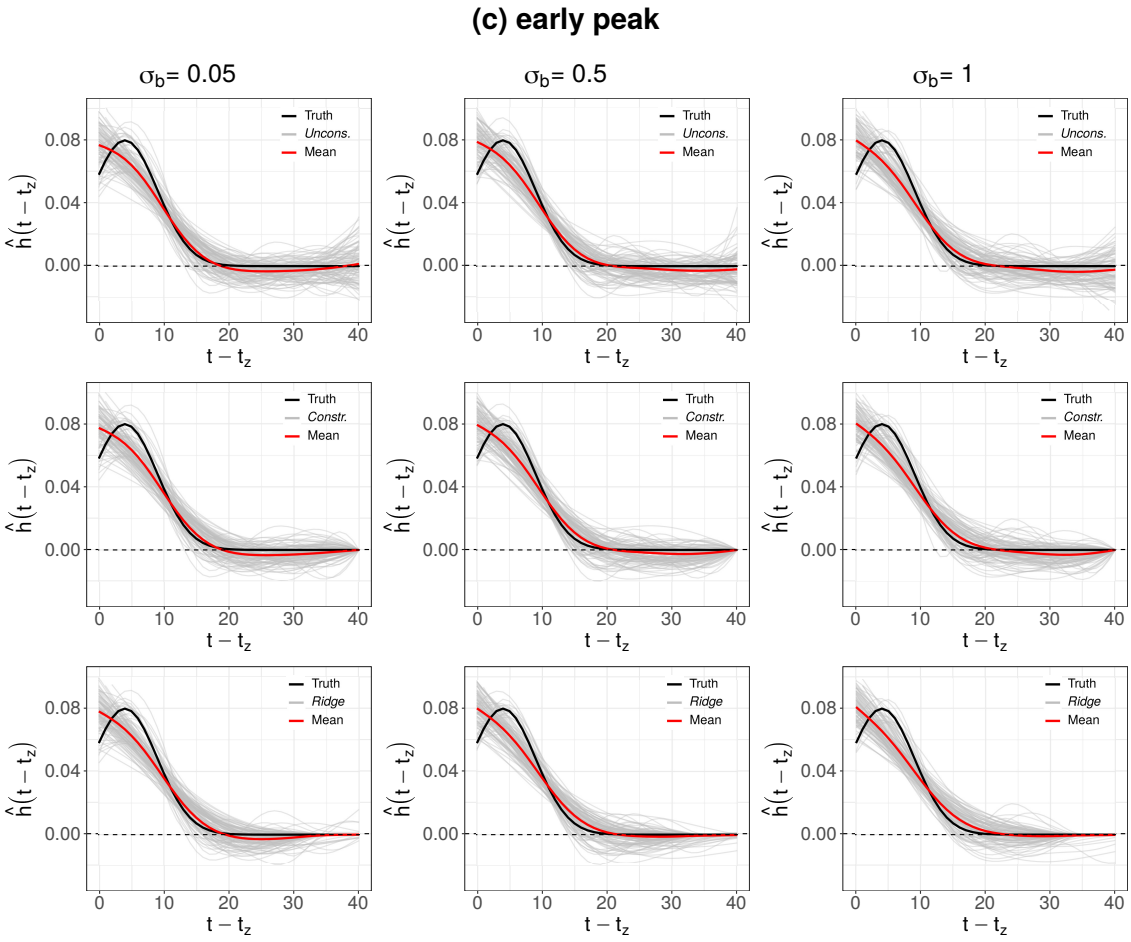


Figure C4: A random sample of 100 estimated weight functions (in grey) for the *Uncons.* model (1st row), *Constr.* (2nd row) and *Ridge* model (3rd row), for scenarios $\sigma_b = 0.05$ (left), $\sigma_b = 0.5$ (middle) and $\sigma_b = 1$ (right), together with the true weight function (in black), shape **(c)**, and the mean curve (in red).

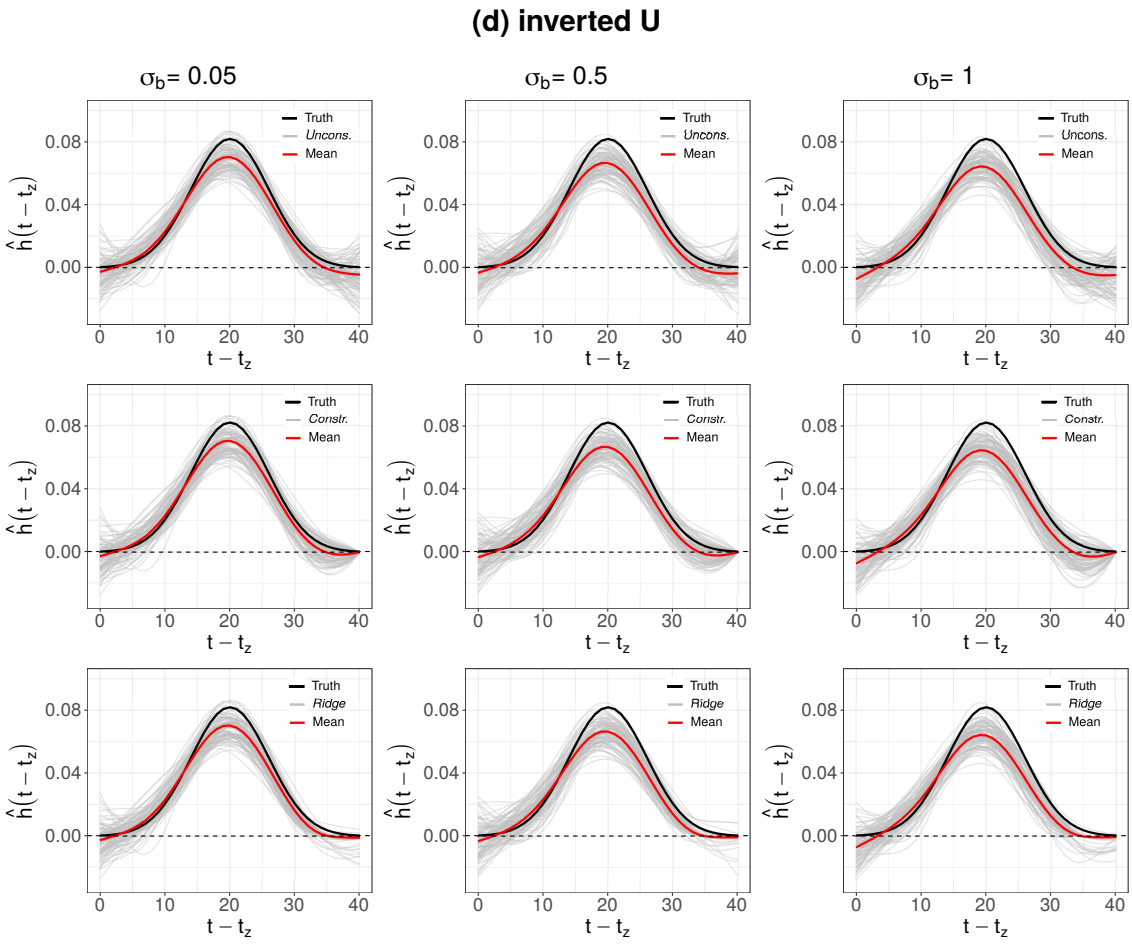


Figure C5: A random sample of 100 estimated weight functions (in grey) for the *Uncons.* model (1st row), *Constr.* (2nd row) and *Ridge* model (3rd row), for scenarios $\sigma_b = 0.05$ (left), $\sigma_b = 0.5$ (middle) and $\sigma_b = 1$ (right), together with the true weight function (in black), shape **(d)**, and the mean curve (in red).

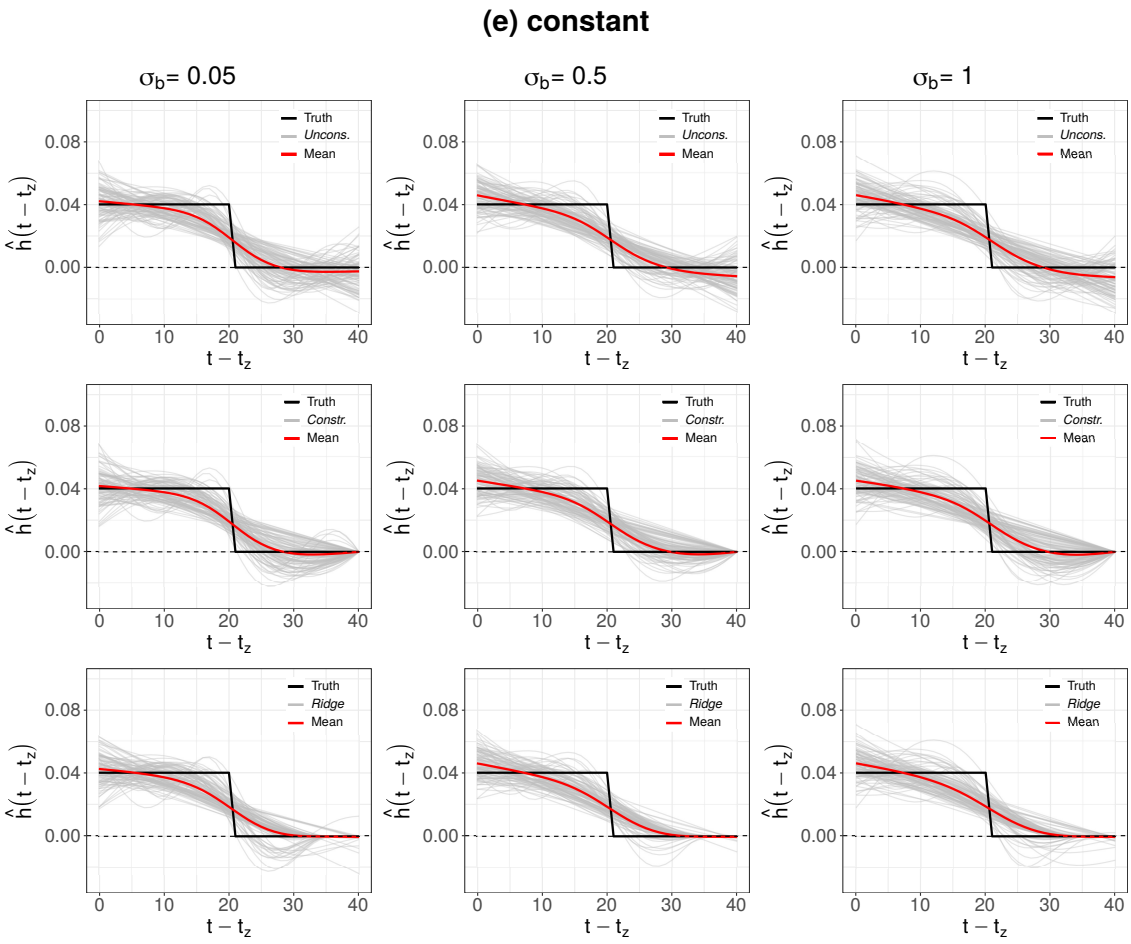


Figure C6: A random sample of 100 estimated weight functions (in grey) for the *Uncons.* model (1st row), *Constr.* (2nd row) and *Ridge* model (3rd row), for scenarios $\sigma_b = 0.05$ (left), $\sigma_b = 0.5$ (middle) and $\sigma_b = 1$ (right), together with the true weight function (in black), shape **(e)**, and the mean curve (in red).

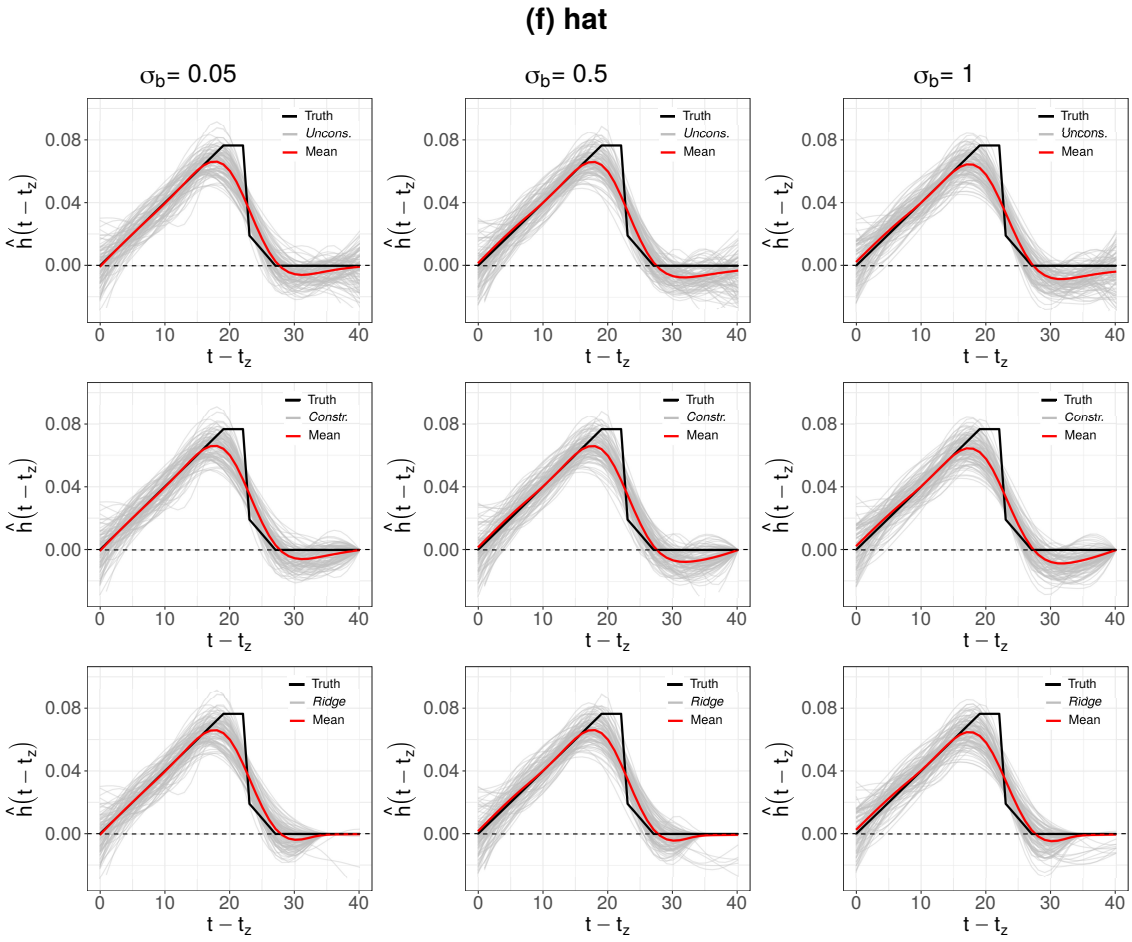


Figure C7: A random sample of 100 estimated weight functions (in grey) for the *Uncons.* model (1st row), *Constr.* (2nd row) and *Ridge* model (3rd row), for scenarios $\sigma_b = 0.05$ (left), $\sigma_b = 0.5$ (middle) and $\sigma_b = 1$ (right), together with the true weight function (in black), shape **(f)**, and the mean curve (in red).

Table C2: Simulation results for $N_{\text{sim}} = 500$ in each scenario setting, presented in terms of mean RMSE and mean coverage ($\alpha = 0.05$) of $h_{t,t_z,z(t_z)}$, and RMSE of σ_b.

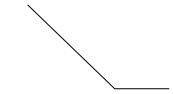
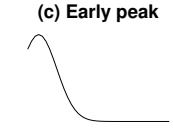
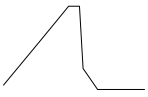
Data generation mechanism		**Model**	**Mean RMSE**		**Mean Coverage**
WCE shape	**Heterogeneity**		$h_{t,t_z,z(t_z)}$	σ_b	$h_{t,t_z,z(t_z)}$
(a) Exponential decay 	$\sigma_b = 0.05$	*Uncons.*	0.012	0.255	**0.912**
	$\sigma_b = 0.05$	*Constr.*	0.010	0.254	0.837
	$\sigma_b = 0.05$	*Ridge*	**0.009**	**0.251**	0.753
	$\sigma_b = 0.5$	*Uncons.*	0.012	**0.208**	**0.889**
	$\sigma_b = 0.5$	*Constr.*	0.011	**0.208**	0.804
	$\sigma_b = 0.5$	*Ridge*	**0.009**	0.209	0.755
	$\sigma_b = 1$	*Uncons.*	0.011	**0.232**	**0.904**
	$\sigma_b = 1$	*Constr.*	0.010	**0.232**	0.830
	$\sigma_b = 1$	*Ridge*	**0.008**	0.234	0.774
(b) Bi–linear 	$\sigma_b = 0.05$	*Uncons.*	0.007	0.129	0.863
	$\sigma_b = 0.05$	*Constr.*	0.006	0.128	**0.898**
	$\sigma_b = 0.05$	*Ridge*	**0.005**	**0.127**	0.889
	$\sigma_b = 0.5$	*Uncons.*	0.007	**0.185**	0.847
	$\sigma_b = 0.5$	*Constr.*	0.007	0.187	**0.874**
	$\sigma_b = 0.5$	*Ridge*	**0.006**	0.186	0.869
	$\sigma_b = 1$	*Uncons.*	0.009	**0.379**	0.819
	$\sigma_b = 1$	*Constr.*	0.008	0.381	**0.851**
	$\sigma_b = 1$	*Ridge*	**0.007**	0.380	0.849
(c) Early peak 	$\sigma_b = 0.05$	*Uncons.*	**0.009**	**0.142**	**0.865**
	$\sigma_b = 0.05$	*Constr.*	**0.009**	**0.142**	0.839
	$\sigma_b = 0.05$	*Ridge*	**0.009**	**0.142**	0.853
	$\sigma_b = 0.5$	*Uncons.*	**0.009**	0.161	**0.868**
	$\sigma_b = 0.5$	*Constr.*	**0.009**	0.161	0.844
	$\sigma_b = 0.5$	*Ridge*	**0.009**	**0.160**	0.855
	$\sigma_b = 1$	*Uncons.*	0.010	**0.338**	**0.862**
	$\sigma_b = 1$	*Constr.*	**0.009**	**0.338**	0.844
	$\sigma_b = 1$	*Ridge*	**0.009**	**0.338**	0.851

Table C3: Simulation results for $N_{\text{sim}} = 500$ in each scenario setting, presented in terms of mean RMSE and mean coverage ($\alpha = 0.05$) of $h_{t,t_z,z(t_z)}$, and RMSE of σ_b (continuation).

Data generation mechanism		**Model**	**Mean RMSE**		**Mean Coverage**
WCE shape	**Heterogeneity**		$h_{t,t_z,z(t_z)}$	σ_b	$h_{t,t_z,z(t_z)}$
	$\sigma_b = 0.05$	*Uncons.*	0.009	0.130	**0.888**
	$\sigma_b = 0.05$	*Constr.*	**0.008**	0.130	0.870
	$\sigma_b = 0.05$	*Ridge*	**0.008**	**0.128**	0.750
(d) Inverted U 	$\sigma_b = 0.5$	*Uncons.*	0.010	**0.181**	**0.839**
	$\sigma_b = 0.5$	*Constr.*	0.010	0.182	0.819
	$\sigma_b = 0.5$	*Ridge*	**0.009**	0.184	0.684
	$\sigma_b = 1$	*Uncons.*	**0.011**	**0.385**	**0.818**
	$\sigma_b = 1$	*Constr.*	**0.011**	0.386	0.797
	$\sigma_b = 1$	*Ridge*	**0.011**	0.387	0.648
	$\sigma_b = 0.05$	*Uncons.*	**0.009**	**0.137**	**0.740**
	$\sigma_b = 0.05$	*Constr.*	**0.009**	0.138	0.728
	$\sigma_b = 0.05$	*Ridge*	**0.009**	0.138	0.722
(e) Constant	$\sigma_b = 0.5$	*Uncons.*	0.010	0.164	0.700
	$\sigma_b = 0.5$	*Constr.*	**0.009**	**0.162**	0.698
	$\sigma_b = 0.5$	*Ridge*	**0.009**	0.163	**0.714**
	$\sigma_b = 1$	*Uncons.*	0.010	0.343	**0.718**
	$\sigma_b = 1$	*Constr.*	0.010	**0.341**	0.704
	$\sigma_b = 1$	*Ridge*	**0.009**	0.342	**0.718**
	$\sigma_b = 0.05$	*Uncons.*	0.011	0.134	0.854
	$\sigma_b = 0.05$	*Constr.*	**0.010**	**0.133**	0.855
	$\sigma_b = 0.05$	*Ridge*	**0.010**	0.135	**0.861**
(f) Hat	$\sigma_b = 0.5$	*Uncons.*	0.011	0.167	0.850
	$\sigma_b = 0.5$	*Constr.*	0.011	0.167	0.850
	$\sigma_b = 0.5$	*Ridge*	**0.010**	**0.163**	**0.861**
	$\sigma_b = 1$	*Uncons.*	0.012	0.346	0.858
	$\sigma_b = 1$	*Constr.*	**0.011**	0.346	0.860
	$\sigma_b = 1$	*Ridge*	**0.011**	**0.343**	**0.869**

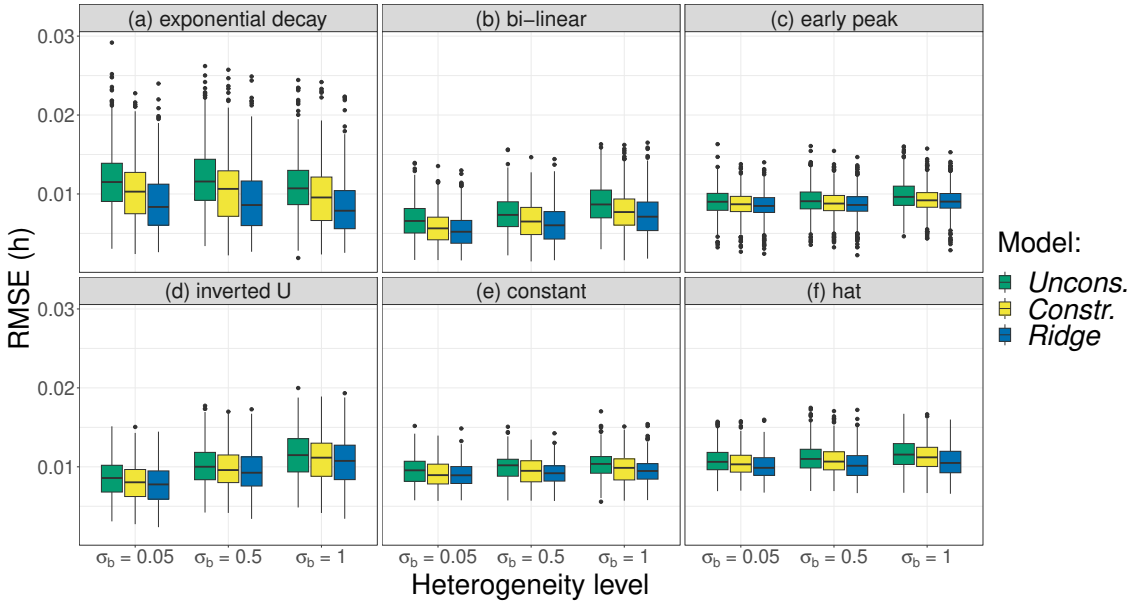


Figure C8: Distribution of the RMSE of $h(t, t_z, z(t_z))$ across all simulation settings ($N_{\text{sim}} = 500$).

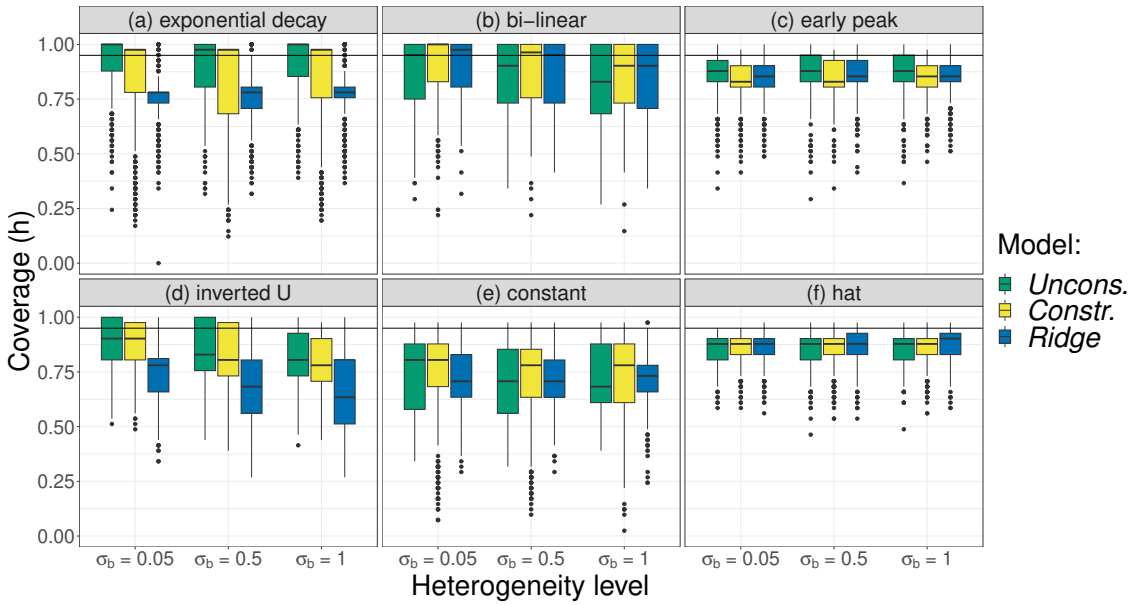


Figure C9: Distribution of the 95% point-wise Coverage of $h(t, t_z, z(t_z))$ across all simulation settings ($N_{\text{sim}} = 500$). The horizontal black line is set at a 0.95 nominal coverage value.

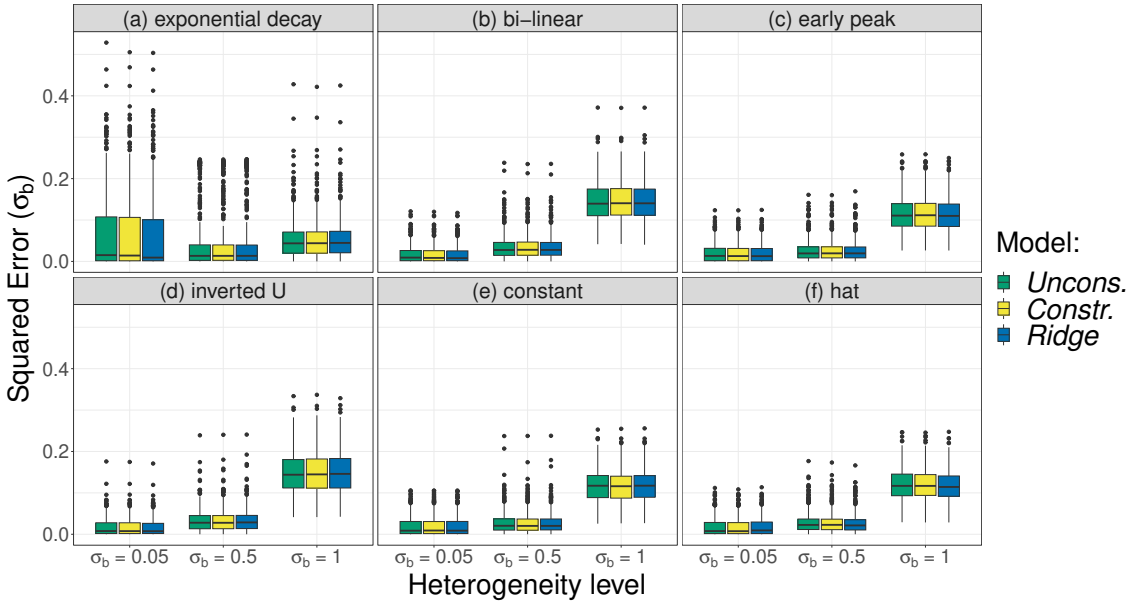


Figure C10: Distribution of the squared error of σ_b across all simulation settings ($N_{\text{sim}} = 500$).

Table C4: Frequency of the models that yield best BIC and Deviance Explained in each replication of the simulation, $N_{\text{sim}} = 500$, across all scenarios.

Data generation mechanism							
WCE shape	**Heterogeneity**	**Lowest BIC** *Uncons.* (%)	**Lowest BIC** *Constr.* (%)	**Lowest BIC** *Ridge* (%)	**Largest Dev.** *Uncons.* (%)	**Largest Dev.** *Constr.* (%)	**Largest Dev.** *Ridge* (%)
(a) Exponential decay 	$\sigma_b = 0.05$	3	18	**79**	**76**	15	9
	$\sigma_b = 0.5$	4	20	**76**	**78**	16	6
	$\sigma_b = 1$	5	18	**77**	**77**	21	2
(b) Bi–linear 	$\sigma_b = 0.05$	5	17	**78**	**68**	16	16
	$\sigma_b = 0.5$	2	35	**63**	**62**	8	29
	$\sigma_b = 1$	2	34	**64**	**65**	8	27
(c) Early peak 	$\sigma_b = 0.05$	2	18	**80**	**79**	7	14
	$\sigma_b = 0.5$	3	29	**68**	**66**	6	28
	$\sigma_b = 1$	3	24	**73**	**66**	9	25
(d) Inverted U 	$\sigma_b = 0.05$	6	18	**76**	**70**	16	14
	$\sigma_b = 0.5$	4	18	**77**	**72**	14	14
	$\sigma_b = 1$	2	14	**83**	**73**	15	12
(e) Constant 	$\sigma_b = 0.05$	13	24	**64**	**50**	29	21
	$\sigma_b = 0.5$	18	25	**57**	37	**39**	24
	$\sigma_b = 1$	21	23	**56**	34	**51**	15
(f) Hat 	$\sigma_b = 0.05$	4	31	**65**	**60**	5	36
	$\sigma_b = 0.5$	9	**51**	41	30	7	**63**
	$\sigma_b = 1$	5	**48**	46	36	5	**59**

C.2 Application to external training load data

C.2.1 Model specification and results

The log-hazard rate of player l of the model we fit is expressed as:

$$\log(\lambda(t|\boldsymbol{z}_l(t), b_l, i)) = \beta_0 + f_0(t_j) + z_l^{\text{type session}}(t_j)\beta_1 + g_1(\boldsymbol{z}_l^{\text{Speed}}(t), t) + g_2(\boldsymbol{z}_l^{\text{Dist}}(t), t) + b_l$$
$$\forall t \in (\kappa_{j-1}, \kappa_j],\ t_j := \kappa_j \text{ and } b_l \sim N(\mathbf{0}, \sigma_b),$$

where:

- $\beta_0 + f_0(t_j)$ indicates the log-baseline hazard rate,
- $z_l^{\text{type session}}(t_j)$ the type of session undertaken by player l at t_j (whether match or training session),
- g_1 and g_2 are non-linear time-varying effects of the training load variables, i.e. the cumulative effects defined as, $\int_{\tau_{\text{Speed}}(t)} h(t - t_z) z_l^{\text{Speed}}(t_z) dt_z$ and $\int_{\tau_{\text{Dist}}(t)} h(t - t_z) z_l^{\text{Dist}}(t_z) dt_z$,
- and b_l a Gaussian random intercept term associated with player l.

The lag-lead windows are defined to be large enough to identify relevant past exposure effects by fitting a PAMM with a ridge penalization. In this regard, we define $\tau_{\text{Speed}}(t) = \tau_{\text{Dist}}(t) = \{t_z : t > t_z \ \wedge \ t < t_z + 11\}$, meaning that all *Speed* and *Dist* values recorded in the last 10 sessions prior to t (i.e. before three weeks approximately) can affect the hazard of injury at time t.

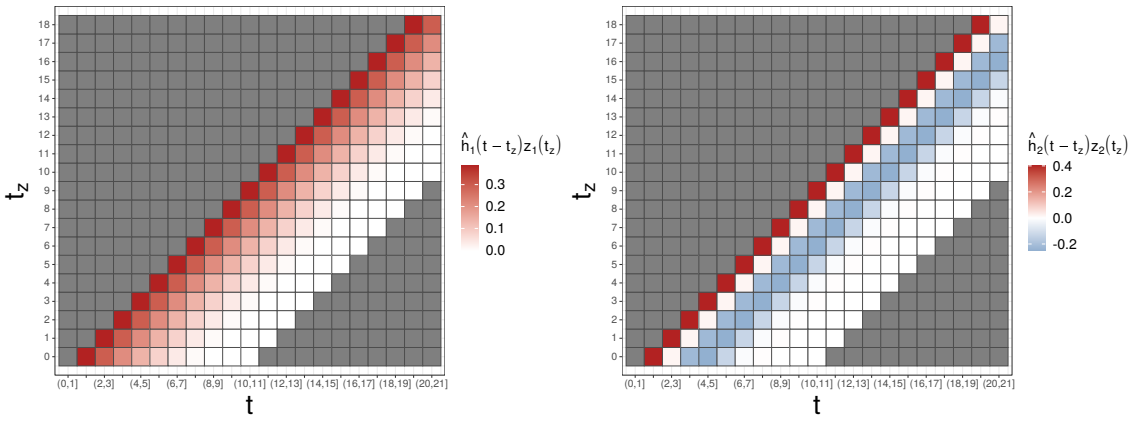


Figure C11: Estimated partial effects for covariates *Speed*, $\hat{h}_1(t - t_z)z_1(t_z)$, (left panel) and for *Dist*, $\hat{h}_2(t - t_z)z_2(t_z)$, (right panel) displayed over their respective lag-lead windows. Note: lag-lead windows are cut for the sake of clarity.

The estimated partial effects for different combinations of t and t_z, with $\boldsymbol{z}^{\text{Speed}}(t_z) = 3.9\ \forall t_z$ and $\boldsymbol{z}^{\text{Dist}}(t_z) = 4700\ \forall t_z$, and the resulting cumulative effects, $\hat{g}(\boldsymbol{z}^{\text{Speed}}(t), t)$ and $\hat{g}(\boldsymbol{z}^{\text{Dist}}(t), t)$, are shown in Figures C11 and C12, respectively. In Table C5 the summary of the estimated model coefficients is shown, and in Figure C13, the shapes of the estimated smooth baseline and player random effect are shown.

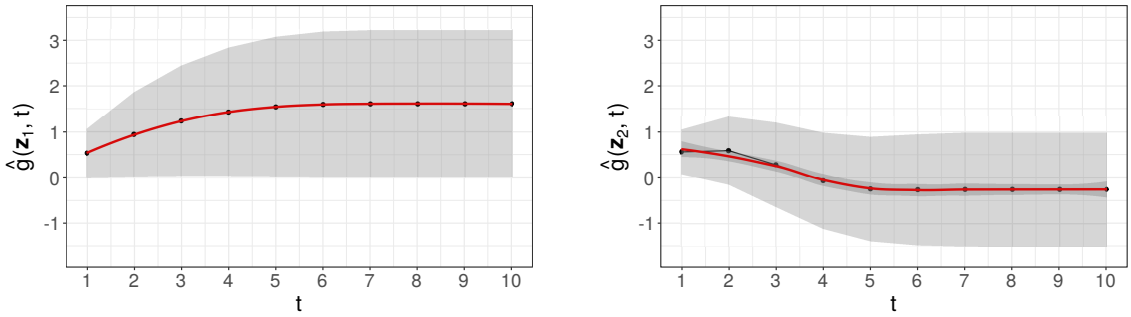


Figure C12: Estimated cumulative effects, $\hat{g}(\boldsymbol{z}_1(t), t) = \hat{g}(\boldsymbol{z}^{\text{Speed}}(t), t)$ and $\hat{g}(\boldsymbol{z}_2(t), t) = \hat{g}(\boldsymbol{z}^{\text{Dist}}(t), t)$, for $\boldsymbol{z}^{\text{Speed}}(t_z) = 3.9\ \forall t_z$ and $\boldsymbol{z}^{\text{Dist}}(t_z) = 4700\ \forall t_z$, respectively.

Table C5: Model summary.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	-7.2664	0.9198	-7.8997	< 0.0001
Type session:match	2.4481	0.2679	9.1371	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
Baseline	2.1280	2.6029	11.1358	0.0109
Speed cumulative effect	0.8981	1.1243	4.5324	0.0522
Dist cumulative effect	2.4947	2.9676	11.2029	0.0195
Player random effect	9.2944	35.0000	14.2589	0.0373

C.2.2 Comparison to conventional training load measures

We consider two measures widely used in the sports medicine and exercise physiology literature, namely, ACWR with rolling averages and ACWR with EWMA, as well as the unweighted sum of the past training exposures.

ACWR stands for acute chronic workload ratio and was introduced to model the relationship between changes in load and injury risk (Killen et al., 2010; Gabbett et al., 2016). It is a ratio describing the acute training load (e.g. the training load of the last week) to the chronic load (e.g. the training load of the last 4 weeks).

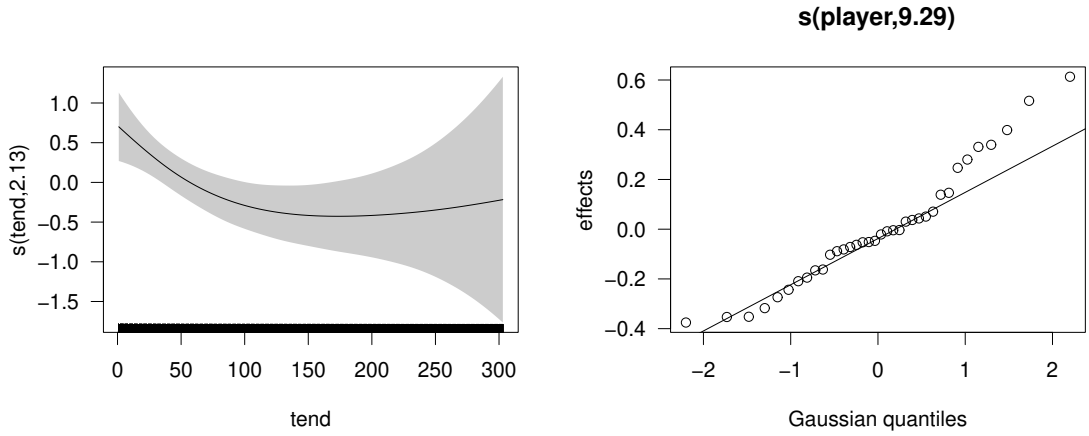


Figure C13: Estimated log-baseline hazard, the shaded area representing the point-wise 95% confidence interval (left); and a quantile-quantile plot for the player random effect (right).

The concept is based on Banister's fitness-fatigue model (Banister and Calvert, 1980) where the acute load dictates the "fatigue" state of an athlete, whereas the chronic load dictates the athlete's overall "fitness". They are intended to reflect the athlete's preparedness.

The ACWR measure compares the load the athlete has performed (acute) relative to the load the athlete has prepared for (chronic). The time frames (or windows) for acute and chronic workloads represent the time needed to dissipate the negative (fatigue) and positive (fitness) effects of training. In general terms:

$$\text{ACWR} := \frac{\text{Acute Load}}{\text{Chronic Load}}.$$

Commonly, and despite critiques, the rolling average has been the most frequently used method to account for the cumulative effects of training load, acute and chronic. The rolling average of a training load variable denoted by z, over a n-sized time-lag window, is defined by:

$$\text{RA}_n(\boldsymbol{z}(t)) = \frac{z(t_{z-n+1}) + z(t_{z-n+2}) + \ldots + z(t_z)}{n},$$

where k is the last value in the time-lag window.

Alternatively, exponentially weighted moving averages (EWMA) have been proposed to summarize the cumulative effects of training load. In this case, EWMA is defined as:

$$\text{EWMA}_n(\boldsymbol{z}(t)) = z(t_z) \cdot \lambda_n + (1 - \lambda_n) \cdot \text{EWMA}_n(\boldsymbol{z}(t_{z-1})),$$

with the time decay constant typically defined by $\lambda_n = \frac{2}{n+1}$, where n is 7 and 28 days, for acute and chronic loads, respectively. This measure acknowledges the weekly variations in load. The effect of training decays over time since it places "more weight" on recent loads and less on distant ones in the past.

Then, one can compute ACWR using either rolling averages or EWMA to quantify the acute and chronic load. Typically, 7 and 28 days are used, respectively, as the time-lag windows of the acute (numerator) and chronic (denominator).

These measures have been largely discussed and criticised in the literature, as they have several conceptual and mathematical limitations. We refer the reader to Wang et al. (2020) for a thorough review and discussion.

In summary, the models we consider differ only in the definition of the cumulative effects:

- ***ACWR (rolling avg.) model***. All the model terms of the main model remain the same except for the cumulative effects of z^{Speed} and z^{Dist}, which we replace with:

$$g(\boldsymbol{z}(t), t) = \text{ACWR}^{\text{RA}}(\boldsymbol{z}(t)) = \frac{\text{RA}_7(\boldsymbol{z}(t))}{\text{RA}_{28}(\boldsymbol{z}(t))}.$$

- ***ACWR (EWMA) model***. All the model terms of the main model remain the same except for the cumulative effects of z^{Speed} and z^{Dist}, which we replace with:

$$g(\boldsymbol{z}(t), t) = \text{ACWR}^{\text{EWMA}}(\boldsymbol{z}(t)) = \frac{\text{EWMA}_7(\boldsymbol{z}(t))}{\text{EWMA}_{28}(\boldsymbol{z}(t))}.$$

- ***Unweighted sum model***. All the model terms of the main model remain the same except for the cumulative effects of z^{Speed} and z^{Dist}, which we define as the cumulative (unweighted) sum of the past six exposures (sessions), i.e.:

$$g(\boldsymbol{z}^{\text{Speed}}(t), t) = \sum_{t_z < t}^{t_z > t-7} z^{\text{Speed}}(t_z) \text{ and } g(\boldsymbol{z}^{\text{Dist}}(t), t) = \sum_{t_z < t}^{t_z > t-7} z^{\text{Dist}}(t_z).$$

To compare the model performance, we compute likelihood-based measures. The results are shown in Table C6, ordered from the best performance to the least, according to the BIC measure.

Table C6: Likelihood-based measures regarding the goodness-of-fit of the fitted models, ordered according to BIC.

Model	**AIC**	**Deviance**	**Deviance Explained**	**BIC**
PAMM WCE ridge model	717.21	539.58	20.57	866.41
Unweighted sum model	802.74	628.49	7.48	924.92
ACWR (rolling avg.) model	804.24	625.67	7.89	950.49
ACWR (EWMA) model	796.74	616.54	9.24	951.02