BMJ Open

Relationship between a daily injury risk estimation feedback (I-REF) based on machine learning techniques and actual injury risk in athletics (track and field): protocol for a prospective cohort study over an athletics season

Pierre-Eddy Dandrieux [iD],[1,2] Laurent Navarro [iD],[2] David Blanco [iD],[3] Alexis Ruffault [iD],[4,5] Christophe Ley [iD],[6] Antoine Bruneau,[7] Joris Chapon [iD],[1] Karsten Hollander [iD],[8] Pascal Edouard [iD][1,9]

For numbered affiliations see end of article.

Correspondence to
Pierre-Eddy Dandrieux;
pierre.eddy.dandrieux@univ-st-etienne.fr

ABSTRACT

Introduction Two-thirds of athletes (65%) have at least one injury complaint leading to participation restriction (ICPR) in athletics (track and field) during one season. The emerging practice of medicine and public health supported by electronic processes and communication in sports medicine represents an opportunity for developing new injury risk reduction strategies. Modelling and predicting the risk of injury in real-time through artificial intelligence using machine learning techniques might represent an innovative injury risk reduction strategy. Thus, the primary aim of this study will be to analyse the relationship between the level of **I**njury **R**isk **E**stimation **F**eedback (I-REF) use (average score of athletes' self-declared level of I-REF consideration for their athletics activity) and the ICPR burden during an athletics season.

Method and analysis We will conduct a prospective cohort study, called **I**njury **Predict**ion with **A**rtificial **I**ntelligence (IPredict-AI), over one 38-week athletics season (from September 2022 to July 2023) involving competitive athletics athletes licensed with the **F**rench **F**ederation of **A**thletics. All athletes will be asked to complete daily questionnaires on their athletics activity, their psychological state, their sleep, the level of I-REF use and any ICPR. I-REF will present a daily estimation of the ICPR risk ranging from 0% (no risk for injury) to 100% (maximal risk for injury) for the following day. All athletes will be free to see I-REF and to adapt their athletics activity according to I-REF. The primary outcome will be the ICPR burden over the follow-up (over an athletics season), defined as the number of days lost from training and/or competition due to ICPR per 1000 hours of athletics activity. The relationship between ICPR burden and the level of I-REF use will be explored by using linear regression models.

Ethics and dissemination This prospective cohort study was reviewed and approved by the Saint-Etienne University Hospital Ethical Committee (Institutional Review Board: IORG0007394, IRBN1062022/CHUSTE). Results of the study will be disseminated in peer-reviewed journals

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study will assess the interest and efficacy of daily monitoring of physical, psychological and contextual variables combined with artificial intelligence in reducing sports injury risk.

⇒ This is the first study in athletics, and in other sports, using a machine learning model to predict injury in a real-time athletics field context.

⇒ This study will use an online tool (web application) to collect data based on daily assessment.

⇒ This study requires daily participation and may therefore be a limiting factor for athletes' involvement.

⇒ As this is an exploratory study, the sample size is unknown, and consequently, the amount of data collected may be limited for the purposes of developing a machine learning model.

and in international scientific congresses, as well as to the included participants.

INTRODUCTION

Athletics (track and field) activity leads to a risk of injury.[1–4] Indeed, epidemiological studies showed that about two-thirds of athletes presented at least one injury during an athletics season.[1 2 5 6] Specifically, in a French one-season athletics study, 65% of athletes reported at least one injury complaint leading to participation restriction (ICPR).[6] Another way to quantify the impact of injuries is the injury burden, defined as the number of sports activity days lost due to injury per 1000 hours of activity.[7] Regarding athletics activity, a previous study showed an ICPR burden equal to 285.6±619.6 days per 1000 hours of athletics activity.[6] Injuries negatively

affect athletes' training, performance, career and health. Therefore, reducing the risk of injury is fundamental to promoting healthy and sustainable athletics activity.

Injury risk reduction interventions in athletics, as in other sports, can vary because of the multifactorial nature of injuries. These interventions can target the athletes' physical or psychological condition, training, equipment, rules, lifestyle, medical organisation or sports organisation.[3 8] The emerging practice of medicine and public health supported by electronic processes and communication (e-Health) in sports medicine[9] represents an opportunity to develop new injury risk reduction strategies. Providing athletes with individualised health status feedback[10] could be one of these opportunities and could involve more domains in a bio-psycho-social approach. Artificial intelligence (AI) approaches using machine learning (ML) techniques make it possible to provide more in-depth feedback provided to athletes, by individualising their health status, and especially by taking into account individual multifactorial data that are used by the predictive models.[11–13] ML is a field of AI consisting of the development of algorithms that can automatically learn from data to make decisions.[12] In sports medicine, supervised ML algorithms are often used to direct the model's prediction towards a defined goal or target phenomenon (eg, injury events).[12 14] ML models have the potential to act as an automated data analysts that are able to provide insight into the athlete's condition.[15] Therefore, ML models may help the clinical decision-making process for sports scientists, team physicians and athletic trainers when the data acquired from different sources (eg, wearable sensors and even questionnaires from smartphones) are transformed into accurate decisions regarding the health, safety and performance of athletes.[15] Indeed, ML methods can be used to identify athletes at high risk of injury and can help detect the most important injury risk factors.[12] Consequently, ML techniques could also provide individualised injury risk estimations for athletes.

A recent systematic review of the methodology and performance of existing musculoskeletal injury prediction models in sports found that 98% of these models have a high risk of bias.[16] The main reasons for this finding were the inappropriate and incomplete evaluation procedure of the models[16] and the lack of transparency when reporting the models (ie, full equations or complete code not appropriately described).[17] This can be considered a poor methodology and lead to a lack of generalisation. ML predictive model analysis is principally made retrospectively at the end of the sports season.[18–25] Such retrospective developments could lead to difficulties in reusing the models with new data sets or for performing external validations in different contexts (eg, environmental differences, professional vs amateur athletes; geographical, club vs national sports federations; or temporal situations, the same population at different period vs different populations at different periods). In a real sports medicine context, using geographical or domain data to externally validate the predictive model could be highly challenging (eg, collaboration, environmental constraints). Thus, the quality and nature of each new observation will influence model performance, making it more complex to compare modelling procedures or validate predictions.[26] On the contrary, using temporal input data to provide a real-time prediction could help close the gap between the injury risk reduction programmes and athletics activity, strengthening the model validation process without the need for study replication.[27] As most athletes now have access to digital tools (smartphones, computers), there is an opportunity to implement real-time semi-automated data acquisition through online software (eg, web or smartphone application). Such real-time semi-automated data acquisition can be an opportunity to develop real-time ML approaches. In addition, both techniques, ML and real-time semi-automated data acquisition make it possible to individualise injury risk estimation by using adaptative tree-based questionnaires (eg, individualised questions based on prior answers) and personal features such as the data related to injury risk estimation at a given time. To sum up, providing an individual **I**njury complaint leading to participation restriction **R**isk **E**stimation **F**eedback (I-REF) based on ML could be a relevant injury risk reduction approach.

Study hypothesis and objectives
Hypothesis
We hypothesise that providing a daily individual I-REF presented to each athlete will be related to a reduced ICPR burden during an athletics season.

Primary goal
The primary aim of this study will be to analyse the relationship between the level of I-REF use (calculated as the average score of the athletes' self-declared level of consideration of the I-REF for their athletics activity) and the ICPR burden during an athletics season (ie, 34-week follow-up).

Secondary goal
The secondary aims of this study will be to analyse the association between the level of I-REF use and:
► The percentage of athletes with at least one ICPR during an athletics season.
► The time to the first ICPR during an athletics season.
► The number of ICPR per 1000 hours of athletics activity during an athletics season.

The tertiary objectives will be to analyse the association between the frequency of I-REF view (calculated as the ratio between the total number of days that an athlete checks their I-REF and the number of days that the I-REF is generated) and:
► The ICPR burden during an athletics season.
► The percentage of athletes with at least one ICPR during an athletics season.
► The time to the first ICPR during an athletics season.
► The number of ICPR per 1000 hours of athletics activity during an athletics season.
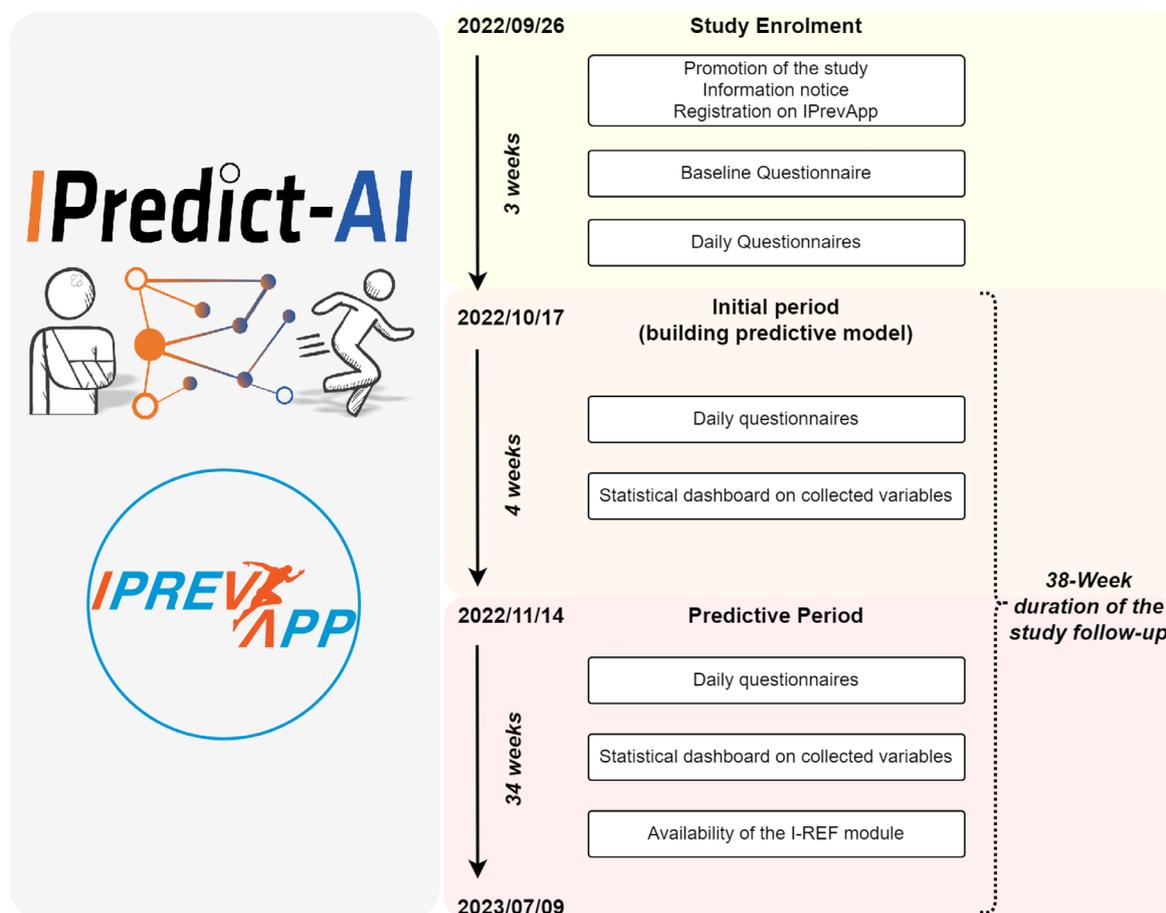
Figure 1 Study design overview. IPredict-AI, Injury Prediction with Artificial Intelligence; IPrevApp, Injury Prevention Application; I-REF, Injury Risk Estimation Feedback.

METHODS AND ANALYSIS

Study design and overall procedure

We will conduct a prospective cohort study, called **I**njury **Predict**ion with **A**rtificial **I**ntelligence (IPredict-AI), over one athletics season (38 weeks), from September 2022 to July 2023 (figure 1). This study will involve competing athletics athletes licensed with the **F**rench **F**ederation of **A**thletics (FFA). This study protocol is reported according to the **ST**rengthening the **R**eporting of **OB**servational studies in **E**pidemiology (STROBE) items[28] and following the **S**tandard **P**rotocol **I**tems: **R**ecommendations for **I**nterventional **T**rials-**A**rtificial **I**ntelligence (SPIRIT-AI).[26]

Patient and public involvement

Participants and the public were involved in the design and development of the study protocol.

Materials

The inclusion procedure and data collection will be carried out online on a mobile and computer website application (IPrevApp, https://iprevapp.emse.fr) hosted on a local host server AlmaLinux V.8.6 (Sky Tiger) with a database (MariaDB V.10.1.48). A virtual Python environment (through an open-source Python V.3.10.2)[29]

is implemented to perform the ML analysis using the open-source scikit-learn libraries V.1.1.2,[30] on high-performance computing hardware (2 CPU Intel Xeon Gold 6132, 14-core, 2.6 GHz). All statistical analyses regarding the primary and secondary aims will be carried out using the Statistical Software R.[31]

Population recruitment and inclusion/exclusion criteria

At the beginning of the 2022–2023 athletics season, the FFA will disseminate information about the IPredict-AI study through their website, social media and by email to all athletes licensed at the FFA for competition. These emails will include an invitation to participate in this study at an individual level, as well as a description of the study purpose and procedure, the participation right, a letter of information and a link to register on the website application. Once the athlete will be registered, the athlete's account will open automatically and will offer access to the IPredict-AI study. The inclusion period will be 3 weeks from 26 September 2022 to 17 October 2022.

The inclusion of athletes will be based on the following criteria: athletes must be licensed at the FFA for competition (sprints, hurdles, jumps, throws, combined events

and endurance disciplines), without any counterindications for competitive athletics activity attested by the licence at the FFA, be aged between 15 and 60, have daily access to a digital device (smartphone, computer, tablet) with a network connection (public or private). We will not exclude athletes based on their baseline injury status or history.[6]

Injury definition

Following Edouard *et al*,[6] we chose the term 'injury complaint' since it refers to self-reported information without medical diagnosis.[32] Here, an injury is defined as: 'a pain, physical complaint or musculoskeletal lesion sustained by an athlete regardless of whether it received medical attention or its consequences with respect to impairments in connection with competition or training'.[33] Athletes reporting an injury complaint will be asked to provide the following information: the circumstance of injury occurrence (training, competition, not related to athletics activity), mode of onset (sudden or gradual),[33] injury location[33] and consequence on athletic participation classified into four categories: (1) full participation with no discomfort, (2) full participation with discomfort, (3) reduced participation due to injury complaint and (4) full absence due to injury complaint.[6 34] Injury complaints will be differentiated between those related to athletics activity and those outside of athletics activity. For the study outcome, we will restrict the analysis to injury complaints related to athletics activity (training and competition). The term 'injury complaint leading to participation restriction' (ICPR) corresponds to the last two categories (c and d).

Data collection

All data will be collected through questionnaires on the website application. To minimise filling errors, we will use only close-ended questions (slider, single-choice, multiple-choice, drop-down list). To avoid missing values during the filling process of a new observation, all questions in the questionnaires will be mandatory.

After being included in the study, athletes will have to fulfil a baseline questionnaire that contains their demographic characteristics (age, weight, height, sex, primary athletics discipline) and their history of injuries and illnesses during the previous athletics season (2021/2022) (table 1).

During the athletics season and throughout the study (from the end of the inclusion to the end of the study follow-up, ie, 38 weeks), the daily follow-up will consist of two questionnaires: one in the morning and one in the evening. These questionnaires will collect parameters related to the athletics activity (eg, volume, intensity, types), athletes' psychological state (eg, anxiety, stress, self-regulation, emotions), sleep, injuries, illnesses and, only for the 34 last weeks, the level of I-REF use and the frequency of I-REF view. All these parameters are shown in table 1. The daily questionnaires will take between 0.5 and 5 min to be filled out and will be displayed automatically

in the website application. They will be accessible during the whole day from 3:00 a.m. to 24 hours. Every night at 3:00, the answers will be reset for the next day.

In this study, missing data will necessarily result from athletes not responding to questions. When omissions occur, no imputation strategy will be used, the questionnaire will be considered incomplete. In addition, some questions and thus variables (eg, competition, training session, stress event) will depend on responses to other questions. Sometimes there will be no response to a certain question and thus no value for a related variable, but this cannot be considered a missing observation. These variables will take a predefined default value within their initial response range (eg, if the athlete replies that he/she did not go to training, the database will record 0 hours of training). The variables which cannot take a default value will be replaced by a constant one (eg, '−1'), which reflects the absence of a daily life event.

The response proportion to the questionnaire will be calculated at the individual athlete's level and expressed as the number of responses obtained in the questionnaires divided by the total number of expected responses.[6] Several strategies will be used to promote regular participation and maintain high response rates throughout the study. First, in order to help athletes to integrate the questionnaires into their daily lives, a calendar file in universal '.ics' format will be made available to athletes directly from the website application and will contain all IPredict-AI study reminders for easy addition into a personal calendar. In addition, automatic reminders will be generated if an athlete has not yet completed a questionnaire before the questionnaire closing time. Furthermore, social media, emails and newsletters within the website application will be used to disseminate information regarding the study (eg, tips and tricks regarding the use of the website application, and key details), thus promoting the use of the study's daily questionnaires throughout the study period. Participating athletes will have access to a visual dashboard on the platform displaying their personal data for the collected variables over time, allowing an individualised real-time summary of the athlete's answers.

Daily individualised feedback on ICPR risk estimation (I-REF)

The daily feedback on individual athlete ICPR risk estimation (ie, I-REF) will be calculated by a dynamic model (ie, predictive model) using the collected data. More precisely, the feedback presented by the website application will reflect the athlete's probability of ICPR and will be called I-REF (figure 2). The I-REF value will range between 0% (no risk of injury) and 100% (maximal risk of injury) for the following day (figure 2A). The I-REF value will be displayed on the website application together with the receiver operator characteristic area under the curve (ROC-AUC), which measures the ability of the predictive classifier model to distinguish between the two classes of injury/non-injury (figure 2B). This ROC-AUC ranges from 0 (bad performance of the predictive model) to 1 (perfect performance of the predictive model). The

Table 1 Summary of data collected during the IPredict-AI study

Questionnaires details					Machine learning model	
					Input features X	Output data Y
Questionnaire	Occurrence	Theme	Variables	Period (figure 1.)	Input features X	Output data Y
Collected variables						
Registration	Unique	Demographic	Age	SE	X	
Baseline	Unique	Demographic	Sex	SE, IP, PP	X	
Baseline	Unique	Demographic	Weight	SE, IP, PP	X	
Baseline	Unique	Demographic	Height	SE, IP, PP	X	
Baseline	Unique	Athletics activity	Number of years of athletics activity	SE, IP, PP	X	
Baseline	Unique	Athletics activity	Main athletics discipline	SE, IP, PP	X	
Baseline	Unique	Athletics activity	Mean hours of athletics per week	SE, IP, PP	X	
Baseline	Unique	Athletics activity	Mean hours of sports per week, outside athletics	SE, IP, PP	X	
Baseline	Unique	Injury	Injury(ies) last season	SE, IP, PP	X	
Baseline	Unique	Illness	Illness(es) last season	SE, IP, PP	X	
Morning	Daily	Sleep details	Time to fall asleep (Ts)	SE, IP, PP		
Morning	Daily	Sleep details	Time to wake up (Tw)	SE, IP, PP		
Morning	Daily	Sleep details	Quality	SE, IP, PP	X	
Morning	Daily	State of fitness	Fatigue	SE, IP, PP	X	
Morning	Daily	State of fitness	Pain	SE, IP, PP	X	
Morning	Daily	Anxiety	Concern	SE, IP, PP		
Morning	Daily	Anxiety	Tension	SE, IP, PP		
Morning	Daily	Anxiety	Confidence	SE, IP, PP		
Morning	Daily	Training	Event	SE, IP, PP		
Morning	Daily	Motivation to train	Intrinsic (i)	SE, IP, PP		
Morning	Daily	Motivation to train	Introjected (ii)	SE, IP, PP		
Morning	Daily	Motivation to train	Extrinsic (iii)	SE, IP, PP		
Morning	Daily	Motivation to train	Amotivation (iv)	SE, IP, PP		
Evening	Daily	Training session	Number (n)	SE, IP, PP		
Evening	Daily	Training session (n)	Duration (D_{train_i})	SE, IP, PP		
Evening	Daily	Training session (n)	Intensity (I_{train_i})	SE, IP, PP		
Evening	Daily	Training session (n)	Type of training	SE, IP, PP		
Evening	Daily	Training session (n)	Number of sprints	SE, IP, PP		
Evening	Daily	Training self-efficacy	Performance assessment (TSE_1)	SE, IP, PP		
Evening	Daily	Training self-efficacy	Performance belief assessment (TSE_2)	SE, IP, PP		
Evening	Daily	Competition session	Event	SE, IP, PP		
Evening	Daily	Competition session	Duration (D_{comp})	SE, IP, PP		
Evening	Daily	Competition session	Intensity (I_{comp})	SE, IP, PP		
Evening	Daily	Competition self-efficacy	Performance assessment (CSE_1)	SE, IP, PP		
Evening	Daily	Competition self-efficacy	Performance belief assessment (CSE_2)	SE, IP, PP		
Evening	Daily	I-REF	I-REF level of use	PP		
Evening	Daily	State of fitness	Fatigue	SE, IP, PP	X	
Evening	Daily	State of fitness	Pain	SE, IP, PP	X	
Evening	Daily	Emotions	Positive emotion	SE, IP, PP	X	
Evening	Daily	Emotions	Negative emotion	SE, IP, PP	X	

Continued

Table 1 Continued

Questionnaires details					Machine learning model	
Questionnaire	Occurrence	Theme	Variables	Period (figure 1.)	Input features X	Output data Y
Evening	Daily	Stress	Event	SE, IP, PP		
Evening	Daily	Stress	Perceived demands ($P_{demands}$)	SE, IP, PP		
Evening	Daily	Stress	Perceived ressources ($P_{ressources}$)	SE, IP, PP		
Evening	Daily	Illness	Event	SE, IP, PP	X	
Evening	Daily	Illness	Detail(s)	SE, IP, PP		
Evening	Daily	Injury	Event	SE, IP, PP		
Evening	Daily	New injury	Event	SE, IP, PP		X
Evening	Daily	New injury	Place of appearance	SE, IP, PP		
Evening	Daily	New injury	Mode of onset	SE, IP, PP		
Evening	Daily	New injury	Mechanism of occurrence	SE, IP, PP		
Evening	Daily	New injury	Body laterality	SE, IP, PP		
Evening	Daily	New injury	Body locality	SE, IP, PP		
Evening	Daily	New injury	Impact of new injury on the training	SE, IP, PP		
Evening	Daily	Injury tracking	Tracking of past injury(ies)	SE, IP, PP	X	
Evening	Daily	Injury tracking	Impact of past injury(ies) on training	SE, IP, PP		
Processed variables						
Morning	Daily	Night duration	$Score = Ts - Tw$	SE, IP, PP	X	
Morning	Daily	Anxiety	$Score = \frac{Concern + Tension}{2} - Confidence$	SE, IP, PP	X	
Morning	Daily	Motivation to train	$Score = 2i + 1ii + (-1iii) + (-2iv)$	SE, IP, PP	X	
Evening	Daily	Training workload (T_w)	$Score = \sum_{i=1}^{n} \left(D_{train_i} * I_{train_i} \right)$	SE, IP, PP	X	
Evening	Daily	Training self-efficacy	$Score = \frac{TSE_1 + TSE_2}{2}$	SE, IP, PP	X	
Evening	Daily	Competition workload (C_w)	$Score = D_{comp} * I_{comp}$	SE, IP, PP		
Evening	Daily	Overall daily workload	$Score = T_w + C_w$	SE, IP, PP	X	
Evening	Daily	Competition self-efficacy	$Score = \frac{CSE_1 + CSE_2}{2}$	SE, IP, PP	X	
Evening	Daily	Stress	$Score = -P_{demands} + P_{ressources}$	SE, IP, PP	X	

IP, initial period; IPredict-AI, Injury Prediction with Artificial Intelligence; I-REF, Injury complaint leading to participation restriction Risk Estimation Feedback; PP, predictive period; SE, study enrolment.

I-REF value will be estimated individually based on the athlete's features (table 1). In order to mitigate overestimation in ROC-AUC score estimation, we will adhere to the practice of nested cross-validation, which serves to reduce bias in small data sets.[35] Specifically, 200 iterations of nested cross-validation will be conducted to calculate each ROC-AUC score. This will involve the random selection of 75% of the data as the outer training set for hyperparameters tuning, while the remaining 25% will be reserved for validation, serving as the outer test set. Then, for the inner cross-validation, half of the 75% selected for hyper-parameters tuning, that is, 50% (inner train set), will be used for predicting the remaining 50% (inner test set). To evaluate the performance of each model on the outer test set, we will conduct validation by retaining the best model from the inner cross-validations. This process will be repeated four times for each of the 200 iterations to obtain the best model for each outer test set.

The development of I-REF will be made using tree-based ensemble classifier machine learning techniques. Tree-based models are among the most popular ML predictive models in sports medicine.[12 16] Compared with deep learning models such as neural networks, they are more accurate for data sets where features (like ours) are individually meaningful and not strongly related via a temporal or spatial structure (as in the image or speech recognition). Besides accuracy, explainability is another essential aspect of ML models in healthcare.
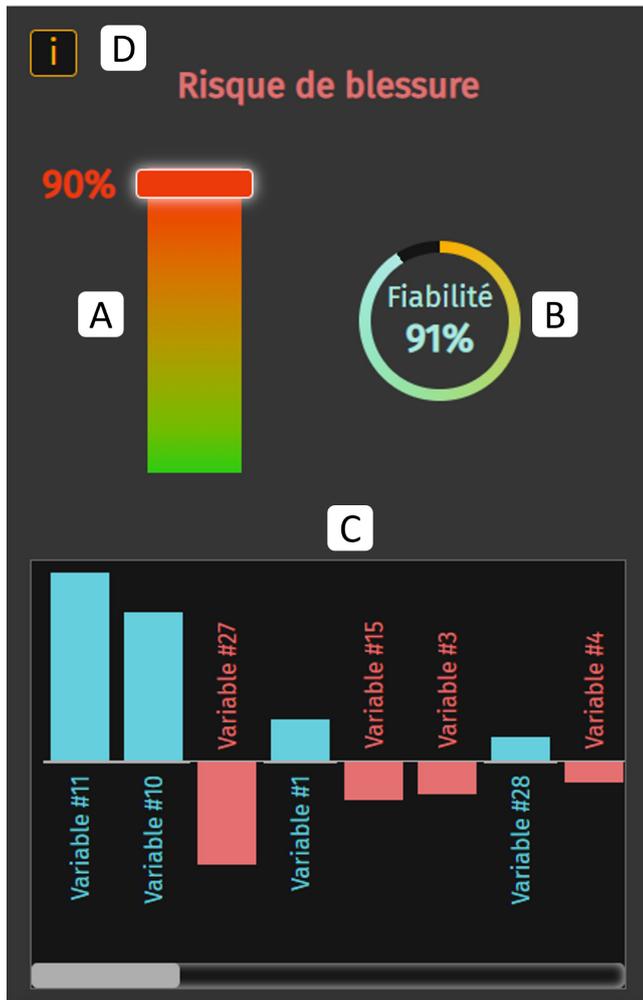
Figure 2 I-REF Module on smartphone. (A) Individual predicted class probability. (B) Ability of the predictive classifier model to distinguish between the two classes injury/non-injury. (C) Individual amount of contribution of each X to predict the Y value, where each variable will be displayed (eg, Variable #12) and will be ordered based on their absolute value influence. Blue variables decrease the risk of injury; red variables increase it. (D) Link for users to a simple explanation of A, B and C. 'Risque de blessure' means 'injury risk', 'Fiabilité' corresponds to the receiver operator characteristic area under the curve and means 'trustability', and 'Variable' means 'variable'.

The often-denounced black box effect of ML models is partly resolved by models that have good explainability.[35][36] Recent research conducted by Lundberg *et al*[37] has provided a tailor-made method for decision trees, hereby strengthening their use. Baseline and daily questionnaires data will be used as input features (X) (table 1). The output parameter (Y) will be the new injury event (ie, ICPR, as a binary outcome). Our modelling process aims to establish a binary classification based on X from the day of starting data collection to the day of the last model training in order to obtain a predicted classification value of Y (ie, 0: not injured, 1: injured) for the following day. Then, we will estimate the predicted class probabilities ranging between 0 and 100 of this binary classification, which will represent the risk of injury. Finally, to individualise the procedure we will use a recent local interpretable explanation method for ML tree algorithms[37] based on classic game-theory additive Shapley values[38] to extract the amount of contribution of each X to predict its Y value (figure 2C,D).[39]

When algorithms are intended to help real-time clinical decision-making as conditions evolve, then these algorithms should make dynamic predictions using new data as it becomes available.[40] However, there is no standardised method to set the period between each model training (ie, the time span between each retraining of the model) in connection with model performance when predictions are made in a real-time context.[40] A possibility to solve this could be to evaluate model performance metrics at several predetermined, discrete time points, achieving continued monitoring of the predictive performance of the model.[27][40][41] We will train the first version of our model within the first 4 weeks of data collection (figure 1), and we will retrain the model each week.

Athletes will be able to check the I-REF value for the first time after 4 weeks of data collection (figure 1). All athletes will be free to consult their own I-REF value, or not, and they will also be free to adapt or not their athletics activity based on it, or not. They will not receive any practical individualised recommendations regarding the risk of ICPR, so the management of this risk will be their own responsibility.

The level of I-REF use by athletes will be assessed by calculating the average score of the self-declared level of consideration that athletes give to I-REF for their athletics activity days. The frequency of I-REF viewing will be calculated as the ratio between the total number of days that athletes check their I-REF divided by the number of days that an I-REF is generated. For the frequency of I-REF viewing, we will consider that an athlete checked the I-REF when he/she enters the I-REF module in the application. This variable will be monitored automatically.

Strategies to limit the bias

We will develop strategies to limit filing errors and missing values (see the 'data collection' section). Strategies will also be implemented to limit bias in the predictive ML model and in the ROC-AUC score estimation (see the 'Daily individualized feedback on ICPR risk estimation (I-REF)' section). In order to limit the bias implied by the non-randomisation, the key variables, that have been shown to be associated with injury risk: sex, age and history of ICPR during the previous season, will be included in the analyses as independent variables. Consequently, there will be no stratification as these key variables will be already included in the analyses.

Sample size

As this study is an exploratory hypothesis-generating research project formal sample calculation is not necessary.

Study outcomes

The primary outcome will be the ICPR burden, defined as the number of athletics activity days lost due to ICPR per 1000 hours of exposure,[6 7] over the period of follow-up with the presentation of I-REF to the athletes (ie, 34 weeks). The secondary outcomes will be (i) the percentage of athletes who will present at least one ICPR during the follow-up, (ii) the time, in hours of athletics activity, to the first ICPR[6] and (iii) the number of ICPR per 1000 hours of exposure during the follow-up. All these variables will be gathered from the daily questionnaires in the website application. The follow-up window will be 34 weeks, starting when the first I-REF is calculated and provided to athletes, and finishing at the end of the follow-up (figure 1).

Statistical methods

We will perform the statistical analyses using a well-known software, namely R (V.3.6.3 (2020-02-29, Copyright 2020 The Foundation for Statistical Computing (Comprehensive R Archive Network, http://www.R-project.org)) and the R library 'survival').[31]

For the primary outcome, we will use a linear regression model where the dependent variable will be the ICPR burden and the independent variables will be the level of use of I-REF, sex, age and history of ICPR during the previous season. Using this model, we will explore the relationship between the ICPR burden and the level of use of I-REF by calculating the model coefficient for the variable 'level of use of I-REF' and its 95% CI.

For the secondary outcome (i), we will adjust a logistic regression model where the dependent variable will be the occurrence (or non-occurrence) of at least one ICPR over the season. We will calculate the OR with 95% CI as a measure of association. For the secondary outcome (ii), we will use survival analysis. First, non-athletics-related ICPR will be analysed as competing risks (using R package 'cmprsk') to explore whether there were significant differences between the athletes who showed higher and lower use of I-REF in terms of the cumulative incidence of ICPR occurred during athletics and outside of athletics. Second, we will adjust a Cox proportional hazards model where the dependent variable will be the time to the first ICPR. We will calculate the HR with 95% CI as a measure of association. Participants will be considered as right-censored (ie, will have incomplete data on the study outcomes at the right side of the follow-up period) if they stop completing the questionnaires, if they have an ICPR that occurred outside of athletics activity or if they have not had an injury at the end of the follow-up. For the secondary outcome (iii), we will use a linear regression model where the dependent variable will be the number of ICPR per 1000 hours of exposure during the follow-up. For all these analyses of secondary outcomes, we will include as independent variables the ones mentioned above in relation to the primary outcome analysis.

Except for the analysis of secondary outcome (ii), which has been explained above, these analyses will be performed on athletes with 100% of the data (complete case analysis). We will also consider performing sensitivity analyses based on consideration of 'best-case' and 'worst-case' scenarios.

Ethics and dissemination

This prospective cohort study was reviewed and approved by the Saint-Etienne University Hospital Ethical Committee (Institutional Review Board: IORG0007394, IRBN1062022/CHUSTE). Participants' study information will not be released outside of the study without the written permission of the participants. During the study, participants will give their personal information and complete study questionnaires through a secured individual account on a website application. The database will be hosted on the physical server of Mines Saint-Etienne secured by a private network and individual professional access (login, password). Only P-ED, PE, LN, and the engineer who developed the application will have access to the database before, during and after the trial. Any extraction from this database will necessarily involve the anonymisation of the subjects. The raw database will not be made available online.

The results of the study will be communicated through articles in peer-reviewed journals following the STROBE items[26] to produce the manuscripts, and in international scientific congresses. Individuals who have contributed to the design and implementation of the protocol will be eligible to be included in publications as coauthors. The results of the study will also be part of the doctoral thesis of P-ED. In addition, the participants of this study will be informed about the results of the study. Dissemination of the results to the end-users, with the aim of knowledge translation will also be made.

Author affiliations
[1]Inter-university Laboratory of Human Movement Biology, EA 7424, F-42023, Université Jean Monnet Saint-Etienne, Lyon 1, Université Savoie Mont-Blanc, Saint-Etienne, Auvergne-Rhône-Alpes, France
[2]Centre CIS, F-42023, Mines Saint-Etienne, Univ Lyon, Univ Jean Monnet, INSERM, U 1059 Sainbiose, Saint-Etienne, Auvergne-Rhône-Alpes, France
[3]Physiotherapy Department, Universitat Internacional de Catalunya, Barcelona, Catalunya, Spain
[4]Laboratory Sport, Expertise, and Performance (EA 7370), French Institute of Sport (INSEP), Paris, France
[5]Unité de Recherche interfacultaire Santé & Société (URiSS), Université de Liège, Liege, Belgium
[6]Department of Mathematics, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[7]French Athletics Federation, Paris, France
[8]Institute of Interdisciplinary Exercise Science and Sports Medicine, Medical School Hamburg, Hamburg, Germany
[9]Department of Clinical and Exercise Physiology, Sports Medicine Unit, University Hospital of Saint-Etienne, Faculty of Medicine, Saint-Etienne, Auvergne-Rhône-Alpes, France

ORCID iDs
Pierre-Eddy Dandrieux http://orcid.org/0000-0001-7230-6728
Laurent Navarro http://orcid.org/0000-0002-8788-8027
David Blanco http://orcid.org/0000-0003-2961-9328
Alexis Ruffault http://orcid.org/0000-0001-6610-4169
Christophe Ley http://orcid.org/0000-0002-2290-8437
Joris Chapon http://orcid.org/0009-0004-0640-4185
Karsten Hollander http://orcid.org/0000-0002-5682-9665
Pascal Edouard http://orcid.org/0000-0003-1969-3612

REFERENCES

1 Edouard P, Morel N, Serra J-M, *et al*. Prévention des lésions de l'appareil locomoteur liées à la pratique de l'athlétisme sur piste. revue des données épidémiologiques. *Science & Sports* 2011;26:307–15.
2 Jacobsson J, Timpka T, Kowalski J, *et al*. Injury patterns in Swedish elite athletics: annual incidence, injury types and risk factors. *Br J Sports Med* 2013;47:941–52.
3 Edouard P, Alonso JM, Jacobsson J, *et al*. Injury prevention in athletics: The race has started and we are on track. *New Stud Athl* 2015:69–78.
4 Edouard P, Navarro L, Branco P, *et al*. Injury frequency and characteristics (location, type, cause and severity) differed significantly among athletics ('track and field') disciplines during 14 international championships (2007–2018): implications for medical service planning. *Br J Sports Med* 2020;54:159–67.
5 Edouard P, Alonso JM. Epidemiology of track and field injuries. *New Studies in Athletics* 2013;28:85–92.
6 Edouard P, Steffen K, Peuriere M, *et al*. Effect of an unsupervised exercises-based athletics injury prevention programme on injury complaints leading to participation restriction in athletics: a cluster-randomised controlled trial. *Int J Environ Res Public Health* 2021;18:11334.
7 Bahr R, Clarsen B, Derman W, *et al*. International Olympic committee consensus statement: Methods for recording and reporting of Epidemiological data on injury and illness in sport 2020 (including STROBE extension for sport injury and illness surveillance (STROBE-SIIS)). *Br J Sports Med* 2020;54:372–89.
8 West SW, Clubb J, Torres-Ronda L, *et al*. More than a metric: how training load is used in elite sport for athlete management. *Int J Sports Med* 2021;42:300–6.
9 Verhagen E, Bolling C. Protecting the health of the @ Hlete: How Online technology may aid our common goal to prevent injury and illness in sport. *Br J Sports Med* 2015;49:1174–8.
10 Hespanhol LC Jr, van Mechelen W, Verhagen E. Effectiveness of online tailored advice to prevent running-related injuries and promote preventive behaviour in Dutch TRAIL runners: a pragmatic randomised controlled trial. *Br J Sports Med* 2018;52:851–8.
11 Edouard P, Verhagen E, Navarro L. Machine learning analyses can be of interest to estimate the risk of injury in sports injury and rehabilitation. *Ann Phys Rehabil Med* 2022;65:101431.
12 Van Eetvelde H, Mendonça LD, Ley C, *et al*. Machine learning methods in sport injury prediction and prevention: a systematic review. *J Exp Orthop* 2021;8:27.
13 Rahlf AL, Hoenig T, Stürznickel J, *et al*. A machine learning approach to identify risk factors for running-related injuries: study protocol for a prospective longitudinal cohort trial. *BMC Sports Sci Med Rehabil* 2022;14:75.
14 Ley C, Martin RK, Pareek A, *et al*. Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc* 2022;30:753–7.
15 Seshadri DR, Thom ML, Harlow ER, *et al*. Wearable technology and analytics as a complementary toolkit to optimize workload and to reduce injury burden. *Front Sports Act Living* 2021;2:630576.
16 Bullock GS, Mylott J, Hughes T, *et al*. Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport. *Sports Med* 2022;52:2469–82.
17 Bullock GS, Hughes T, Arundale AH, *et al*. Black box prediction methods in sports medicine deserve a red card for reckless practice: a change of tactics is needed to advance athlete care. *Sports Med* 2022;52:1729–35.
18 McCullagh J, Whitfort T. An investigation into the application of artificial neural networks to the prediction of injuries in sport. 2013;7:5.
19 Thornton HR, Delaney JA, Duthie GM, *et al*. Importance of various training-load measures in injury incidence of professional rugby League athletes. *Int J Sports Physiol Perform* 2017;12:819–24.
20 Carey DL, Ong K, Whiteley R, *et al*. Predictive modelling of training loads and injury in Australian football. *Int J Comput Sci Sport* 2018;17:49–66.
21 López-Valenciano A, Ayala F, Puerta JosM, *et al*. A preventive model for muscle injuries: a novel approach based on learning algorithms. *Med Sci Sports Exerc* 2018;50:915–27.
22 Ruddy JD, Shield AJ, Maniar N, *et al*. Predictive modeling of hamstring strain injuries in elite Australian footballers. *Med Sci Sports Exerc* 2018;50:906–14.
23 Oliver JL, Ayala F, De Ste Croix MBA, *et al*. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *J Sci Med Sport* 2020;23:1044–8.
24 Rommers N, Rössler R, Verhagen E, *et al*. A machine learning approach to assess injury risk in elite youth football players. *Med Sci Sports Exerc* 2020;52:1745–51.
25 Lövdal SS, Den Hartigh RJR, Azzopardi G. Injury prediction in competitive runners with machine learning. *Int J Sports Physiol Perform* 2021;16:1522–31.
26 Cruz Rivera S, Liu X, Chan AW, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *Lancet Digit Health* 2020;2:e549–60.
27 Rossi A, Pappalardo L, Cintia P, *et al*. Effective injury forecasting in soccer with GPs training data and machine learning. *PLoS ONE* 2018;13:e0201264.
28 von Elm E, Altman DG, Egger M, *et al*. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg* 2014;12:1495–9.
29 van GV, Drake FL. The python language reference. Hampton, NH: Python Software Foundation,
30 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
31 R: a language and environment for statistical computing. In: *R Foundation for Statistical Computing 2021*. Vienna, Austria, Available: https://www.R-project.org/
32 Alonso J-M, Jacobsson J, Timpka T, *et al*. Preparticipation injury complaint is a risk factor for injury: a prospective study of the Moscow 2013 IAAF championships. *Br J Sports Med* 2015;49:1118–24.
33 Timpka T, Alonso J-M, Jacobsson J, *et al*. Injury and illness definitions and data collection procedures for use in epidemiological studies in athletics (track and field): consensus statement. *Br J Sports Med* 2014;48:483–90.

34 Edouard P, Jacobsson J, Timpka T, *et al*. Extending in-competition athletics injury and illness surveillance with pre-participation risk factor screening: a pilot study. *Phys Ther Sport* 2015;16:98–106.

35 Burkart N, Huber MF. A survey on the explainability of supervised machine learning. *Jair* 2021;70:245–317.

36 Navarro L, Dandrieux P-E, Hollander K, *et al*. Digitalization in professional football: an opportunity to estimate injury risk. In: Camarinha-Matos LM, Ortiz A, Boucher X, eds. *Collaborative Networks in Digitalization and Society 5.0*. Cham: Springer International Publishing, 2022: 366–75.

37 Lundberg SM, Erion G, Chen H, *et al*. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;2:56–67.

38 Lundberg SM, Lee S-I, *et al*. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc, 2017. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

39 Lundberg SM, Nair B, Vavilala MS, *et al*. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018;2:749–60.

40 Loftus TJ, Tighe PJ, Ozrazgat-Baslanti T, *et al*. Ideal algorithms in healthcare: explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS Digit Health* 2022;1:e0000006.

41 Rossi A, Pappalardo L, Cintia P. A narrative review for a machine learning application in sports: an example based on injury forecasting in soccer. *Sports (Basel)* 2021;10:5.