Author(s): Jauhiainen, Susanne; Kauppi, Jukka-Pekka; Leppänen, Mari; Pasanen, Kati; Parkkari, Jari; Vasankari, Tommi; Kannus, Pekka; Äyrämö, Sami

Title: New Machine Learning Approach for Detection of Injury Risk Factors in Young Team Sport Athletes

Year: 2021

Version: Accepted version (Final draft)

Please cite the original version:

Jauhiainen, S., Kauppi, J.-P., Leppänen, M., Pasanen, K., Parkkari, J., Vasankari, T., Kannus, P., & Äyrämö, S. (2021). New Machine Learning Approach for Detection of Injury Risk Factors in Young Team Sport Athletes. International Journal of Sports Medicine, 42(02), 175-182. https://doi.org/10.1055/a-1231-5304

New machine learning approach for detection of injury risk factors in young team sport athletes

Abstract

The purpose of this article is to present how predictive machine learning methods can be utilized for detecting sport injury risk factors in a data-driven manner. The approach can be used for finding new hypotheses for risk factors and confirming the predictive power of previously recognized ones. We used three-dimensional motion analysis and physical data from 314 young basketball and floorball players (48.4% males, 15.72±1.79yr, 173.34±9.14cm, 64.65±10.4kg). Both linear (L1-regularized logistic regression) and non-linear methods (random forest) were used to predict moderate and severe knee and ankle injuries (N=57) during three-year follow-up. Results were confirmed with permutation tests and predictive risk factors detected with Wilcoxon signed-rank-test ($p<0.01$). Random forest suggested twelve consistent injury predictors and logistic regression twenty. Ten of these were suggested in both models; sex, body mass index, hamstring flexibility, knee joint laxity, medial knee displacement, height, ankle plantar flexion at initial contact, leg press one-repetition max, and knee valgus at initial contact. Cross-validated areas under receiver operating characteristic curve were 0.65 (logistic regression) and 0.63 (random forest). The results highlight the difficulty of predicting future injuries, but also show that even with models having relatively low predictive power, certain predictive injury risk factors can be consistently detected.

Keywords: Sport medicine, Predictive methods, Machine learning, Knee injuries, Ankle injuries, Basketball and floorball

1. Introduction

Sport injuries are very common across different sports, among both elite and recreational athletes [1–3]. They can have significant effects on the health and performance and may even cause prolonged problems in persons life [3]. Sport injuries can lead to, for example, pain, loss of playing or working time, and decreased motility and stability [3]. The incidence rate of some injuries, such as the anterior cruciate ligament (ACL) injury, is a growing case of concern [4]. Effective prevention of injuries presumes that the most relevant risk factors are found. Even though many intrinsic and extrinsic risk factor have been identified, there is no clear consensus with the findings [5].

A large majority of existing sport injury studies rely on explanatory analysis approach [6, 7]. Explanatory methods have played an important role in the development of sport injury research and will be needed in future research as well. They are used when the purpose is to explain or understand data or phenomena of interest. However, high explanatory power does not necessarily imply high predictive power [8]. Therefore, risk factors that are identified by explanatory methods only demonstrate a statistically significant association with injuries, but might not have predictive power on them [6, 7].

Another limitation of explanatory analysis is that they often focus on a small number of variables and their linear associations with injuries in isolation. However, underlying causes behind sport injuries have been considered to be multifactorial, indicating that a high number of variables and their inter-relationships should be considered [9, 10]. It has also been suggested that using cut-off values and studying only linear interactions between isolated variables can not successfully identify injury predictors, but more complex models should be applied [11]. To overcome these limitations, predictive analysis should be utilized alongside explanatory methods. This has been previously suggested specifically for sport injury research as well [12].

Predictive analysis focuses on predicting new or future observations from data [8]. By exploiting computational power, predictive methods are able to analyze a larger set of variables including their interactions and nonlinear relationships as well as to efficiently remove redundant variables from a model. Therefore, they can be used for generating new hypotheses for sport injury risk factors in a data-driven manner.

In predictive analysis, the generalization ability of a model should always be assessed on independent test data, i.e., data that have not been used in the training phase. This measures how accurate the trained model will be on new unseen observations and only after this validation can any conclusions about the predictive power be drawn [8]. In addition, when constructing a predictive model it is necessary to confirm that the prediction results were significantly above the random chance level. This kind of confirmatory analysis is especially relevant with smaller sample sizes. If this issue is not considered, in the worst case it can lead to false interpretations and conclusions. For example, in neuroscience the problem has been widely recognize [13]. One way to confirm significance of the models and relevance of the chosen predictors are permutation tests [13].

Another important issue related to predictive analysis is the explainability of a model. Explainability means that the model somehow explains its predictions, for example, gives information on how individual variables contribute to the prediction outcome, and does not only predict as black box [14]. Explainable models and their predictions are more informative, easier to trust, and therefore can provide more practical benefits. A term widely used with sophisticated machine learning methods is Explainable Artificial Intelligence (XAI) [14]. In some domains, such as medicine, model explainability is considered highly important [15] and should be pursued in sport science and medicine as well.

During the last couple of years, the first studies using predictive analysis in sport injury research have been conducted [6, 9, 16, 17]. The previous studies have, however, focused solely on the prediction task without paying attention to the explainability of the models. In addition, two of the studies also used a very low number of variables (from three to eleven), although a larger set might have increased the accuracy [9, 16]. The need and potential of predictive machine learning methods in sport injury prediction have been recognized but more research is needed [12, 17].

Therefore, the aim of this study is to utilize predictive machine learning methods to detect variables with predictive power on sport injuries. We present a framework that can be used to detect consistent injury predictors in a data-driven manner and validate their predictive power on independent test data. Consistent means that the variable is constantly chosen as an important predictor in the used model. Our framework utilizes both linear and non-linear classification methods, namely L1-regularized logistic regression and random forests, to predict moderate and severe knee and ankle injuries. Generalization ability of these models is assessed with 10-fold cross-validation. A reference model based on randomized labels is constructed to confirm that the observed prediction performance is not achieved by chance. Consistent injury predictors are detected with Wilcoxon signed-rank test. This approach can be used for finding new hypotheses for injury risk factors as well as confirming the predictive power of previously recognized risk factors. Our secondary aim is to compare linear and non-linear methods for the task.

2. Methods

2.1. Participants

The data were collected in the Predictors of Lower Extremity Injuries in Team Sports (PROFITS) study [18]. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Pirkanmaa Hospital District, Tampere, Finland (ETL-code R10169). The authors declare that this study meets the ethical requirements of the journal [19]. Altogether 175 basketball and 139 floorball youth (12-21 years) players, including 162 females (15.44±1.95 years, 167.92±6.44 cm, 60.86±8.58 kg) and 152 males (16.03±1.59 years, 179.13±8.00 cm, 68.68±10.76 kg) from the two highest junior league levels of the Tampere city district, Finland, were recruited. To be included they had to be official team members (i.e., have valid playing contract and licenses), 21 years old or younger at baseline, and free from injury at baseline. Information about previous injuries, their treatment, and whether the player was fully recovered were assessed with a baseline questionnaire. The players entered the study during the preseason in 2011, 2012, or 2013.

They signed a written informed consent form before inclusion (including parental consent for players aged ≤18 years).

2.2. Data collection

At baseline, each player participated in physical tests including a vertical drop jump (VDJ) (3D motion analysis), height, weight, isokinetic concentric quadriceps and hamstring strength, isometric hip abductor strength, one repetition maximum (1RM) leg press, knee joint laxity (KT-1000), generalized joint laxity (Beighton scale), genu recurvatum, navicular drop, hip anteversion, and hamstring flexibility (for more details see Supplementary Table 1 and online supplementary appendices in [18]).

The VDJ was performed from a 30-cm box. Players were instructed to drop off the box and perform a maximal jump upon landing with their feet on two separate force platforms (BP6001200; AMTI). The 3D motion analysis was carried out using sixteen reflective markers placed over anatomic landmarks on the lower extremities according to the Plug-In Gait Marker set (Vicon Nexus v.i.7; Oxford Metrics) and eight highspeed cameras (Vicon T40). Kinetics and kinematics variables were extracted using the Vicon Nexus Plug-in Gait model. Medial knee displacements were extracted using a custom MATLAB script (MathWorks Inc). For more detailed description of the motion data collection and variable extraction see [18, 20].

The injury definition was based on the time-loss definition by Fuller et al. [21]. We focused on moderate to severe acute non-contact knee and ankle injuries that resulted in an athlete being unable to fully participate in training or match play for at least 8 days. Non-contact injury was defined as an injury which occurred without direct contact to the injured body part. Injuries were recorded by a team coach or another designated team member. For injury registration, the study physicians contacted the team coach or designate on a weekly basis by phone or email. Designate was someone who was always present at practice and matches, e.g., head, assistant, or strength and conditioning coach, team manager, or physiotherapist. The study physicians contacted the athlete after each injury and collected information about the injury time, place, cause, type, location, and the time-loss due to the injury in a standardized phone interview. For exposure registration, the team coaches recorded player participation in team practice and game play and emailed the records to the study group at the end of each month.

2.3. Data preprocessing

All data analysis was performed with MATLAB R2016b (MathWorks Inc) and classification methods run with the *Statistics and machine learning toolbox 11.0*. For classification, the players with moderate and severe acute ankle and knee injuries formed the first group (group A, n=57) and players with no injuries formed the other (group B, n=257). Athletes with mild injuries (time-loss ≤7days, n=21) were excluded from the analysis. Altogether 58 variables were chosen for further analysis by a group of experts in sport medicine, including a sports medicine researcher and four clinical researchers (one physiotherapist and three physicians). Four variables had more than 50% of missing values (iliopsoas and quadriceps extensibility from both legs) as they were added to the test patterns only in the second year of testing and these were excluded from the analysis, resulting into 54 variables. The chosen variables are described in the Supplementary Table 1.

After dropping out irrelevant and sparse variables, 22 variables with missing data remained and were imputed with K-nearest neighbour imputation with k value of 10. On average, each of these 22 variables had five missing values (1.6% of the 314 observations). Data was normalized to have mean of zero and standard deviation of one for each column. The variables that had been measured separately for both right and left legs were transformed to dominant (leg used for kicking a ball) and non-dominant leg variables.

2.4. Choice of classification methods

Two commonly used methods, random forest and L1-regularized logistic regression, were chosen for the binary classification task in our framework. These methods were selected because of their inbuilt variable importance features. Random forest is a nonlinear classification and regression method that has become a standard data analysis tool in different fields such as medicine and bioinformatics [22] and has been used in sport injury research as well [23]. It is based on building an ensemble of multiple decision trees [24]. The model was trained with a hundred trees [24] and the minimum number of observations per tree leaf and the number of predictors to sample at each split were chosen with Bayesian optimization. To estimate the predictive power of the variables, we recorded and analyzed the out-of-bag estimates of variable importance [24].

L1-regularized logistic regression, in turn, is a linear classification method that has been used to model sport injury outcomes [23]. A benefit of this method is that it is capable of automatically discarding redundant and/or irrelevant variables from the model. This is done by penalizing the model with the L1 norm and as a result, some of the variable coefficients tend to shrink to exactly zero. The optimal amount of penalization was estimated with stratified 10-fold cross-validation.

Variable importance for logistic regression was based on the variable coefficient values. We analyzed whether a variable was chosen as a predictor in the model, i.e., the variable coefficient was not shrunk to zero. Variable importance was then the number of times the variable was chosen over the ten CV folds (a value between zero and ten). The sign of each variable coefficient was also assessed in order to perceive whether the variable decreased or increased the injury risk.

2.5. Validation

Generalization ability of our models was assessed with 10-fold cross-validation (CV). K-fold CV is based on randomly splitting the data into K sets and leaving each set at a time for testing while the rest of the sets are used to train a model. Test performance was assessed with Area Under the Receiver Operating Characteristics Curve AUC-ROC [25]. It is based on both true positive and false positive rates and it can be used with imbalanced class distributions which is the case in our data. AUC-ROC provides a value 1.0 for perfect prediction and 0.5 for purely random prediction.

AUC-ROC and variable importance values were estimated by ten-fold cross-validation. Normalization and imputation of the training data were done separately inside each fold and the test data were then normalized using coefficients estimated from the training data. Because K-fold CV is based on random splitting of the data, there is variation in the K-fold validation estimates [26].

Therefore, the analysis was repeated a hundred times and results were averaged over the runs to obtain a more reliable estimate for the generalization ability.

2.6. Confirmatory data analysis

To confirm the significance of our results, permutation tests were used [13]. A reference model was constructed by randomly shuffling the class labels in the training data. By comparing the outcome of the true models to the distribution of values from the random models we confirmed that the performance was not observed by chance. In addition, we can detect significantly consistent injury predictors by comparing the variable importance of the true and the random reference models. If a variable is consistently important in the true model, but not in the reference model, that confirms its significance in the prediction.

To confirm the significance of obtained performance, a paired comparison between AUC-ROC values of the true and random model from a hundred repeated 10-fold CV runs was conducted based on a Wilcoxon signed-rank test. In each CV run, the fold divisions were kept the same for random and true models to allow fair pairwise comparison.

To detect significantly consistent injury predictors, we compared the variable importance values. Again, the values from the hundred repetitions were compared between the random and true models but with Wilcoxon signed-rank test. The limit of significance was set to α=0.01 and corrected with Bonferroni correction. The used framework is summarized in Figure 1.

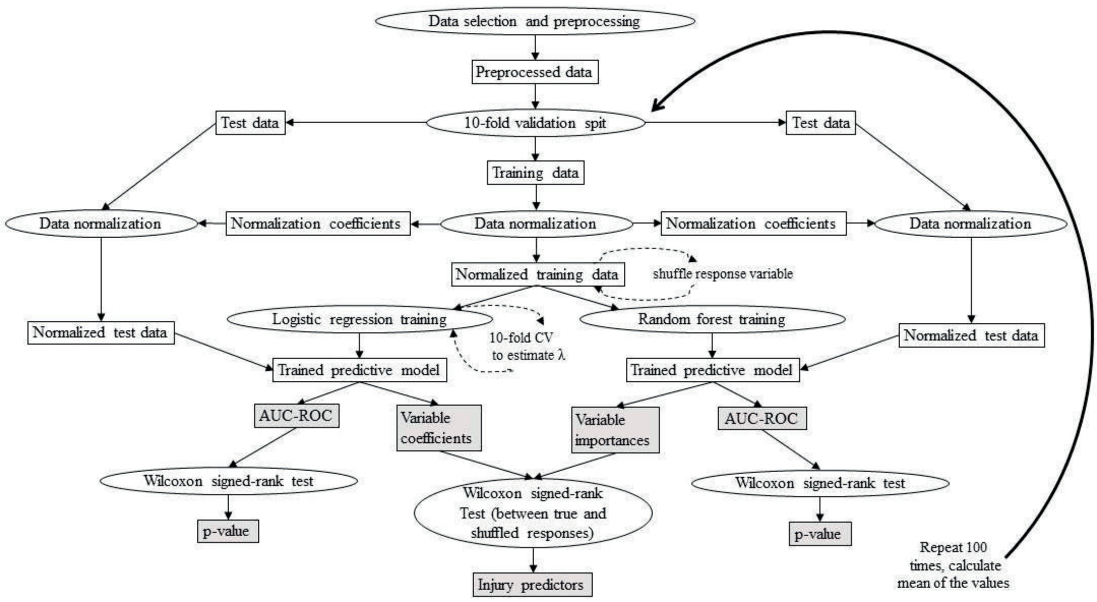

Figure 1. Framework of the proposed predictive analysis approach.

3. Results

3.1. Random forest

Random forest suggested twelve consistent injury predictors ($p < 0.01$). The variable importance values averaged over the CV folds and hundred repeated runs can be seen in Figure 2. The larger the importance value, the greater the importance of the variable is for the prediction task. By comparing the values between true and randomized results, variables with true predictive power can be detected. If the value of true model is significantly larger than the value of random model, its predictive power is not likely result of chance or noise in data. Negative values indicate the variable was not important in prediction.

As seen in the figure, sex, hamstring flexibility (both dominant and non-dominant legs), body mass index (BMI), KT1000 (dominant leg), and height show the highest random forest importance values. Other suggested predictors include leg press 1RM, knee valgus at IC (dominant leg), knee flexion peak (non-dominant leg), medial knee displacement (dominant leg), ankle flexion at IC (dominant leg), and navicular drop (non-dominant leg).

The mean AUC-ROC value for random forest was 0.63 (0.94 for the training data). The AUC-ROC values were higher (p<0.001) with real responses than the randomized ones (mean AUC-ROC 0.48), which confirms the significance of the random forest models.

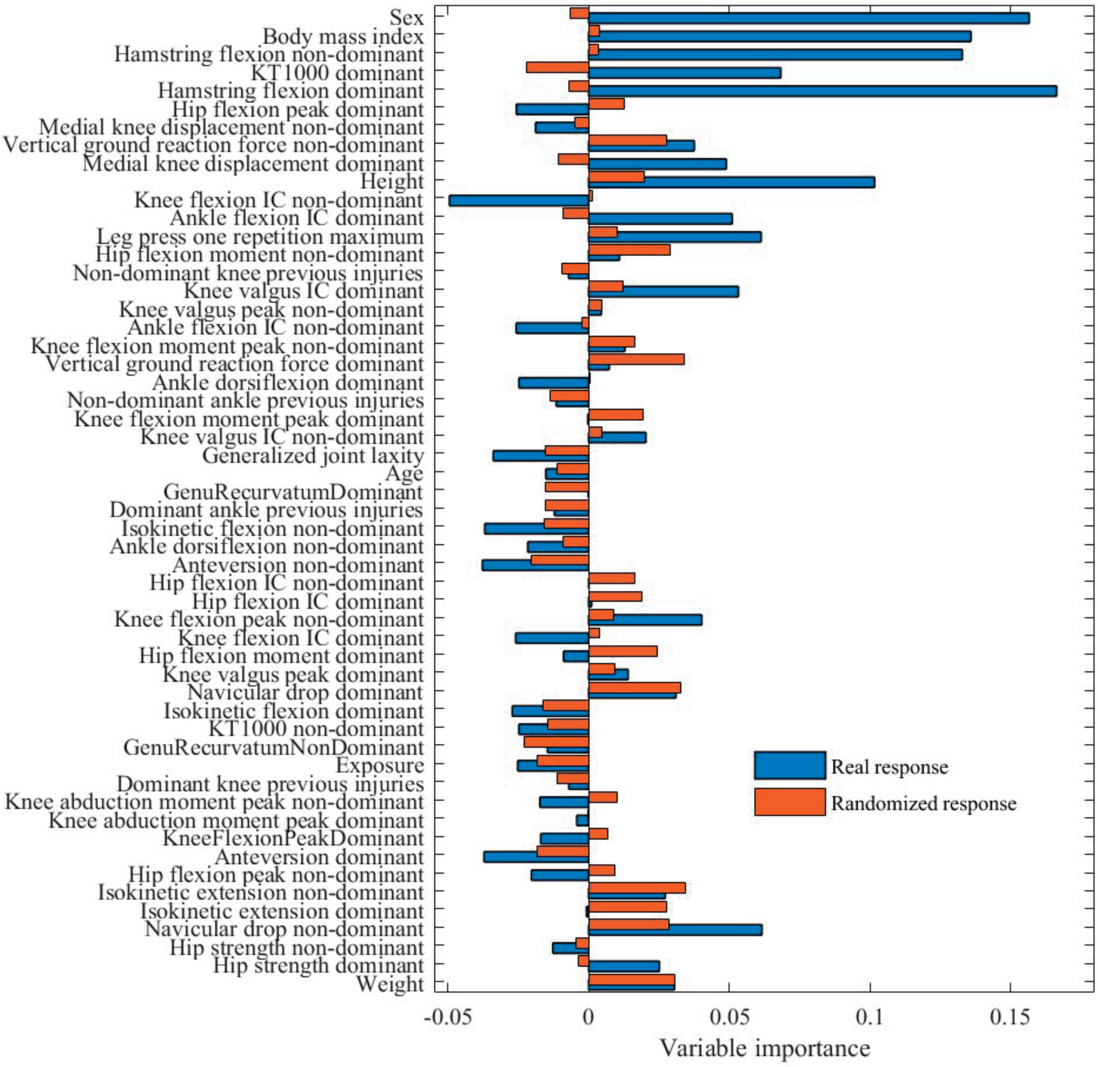

Figure 2. Variable importance values from random forest. Blue bars correspond to the results with real response, red ones with randomized response.

3.2. Logistic regression

Figure 3 shows the variables chosen most frequently as predictors in the L1-regularized logistic regression. The bars represent the number of CV folds where a variable was chosen for the predictive model (i.e., its coefficient was not shrunk to zero). As can be seen in the figure, a part of variables were chosen for prediction in almost every CV split , whereas the others were regarded as not important and their coefficients shrunk to exactly zero. Twenty variables were suggested as consistent injury predictors ($p<0.01$) with the logistic regression model.

The suggested variables were sex, BMI, hamstring flexibility (both legs), KT1000 (dominant leg), hip flexion peak (dominant leg), medial knee displacement (both legs), vertical ground reaction force (vGRF) (both legs), height, knee flexion at IC (non-dominant leg), ankle flexion at IC (both legs), leg press 1RM, hip flexion moment peak (non-dominant leg), previous injuries of non-dominant knee, knee valgus at IC (dominant leg), knee valgus peak (non-dominant leg), and knee flexion moment peak (non-dominant leg). In the figure, these are the twenty variables with the highest frequency value.

The mean AUC-ROC value for logistic regression models was 0.65 (0.76 for the training data). The AUC-ROC values were higher ($p<0.001$) with real responses than the randomized ones (mean AUC-ROC 0.50).

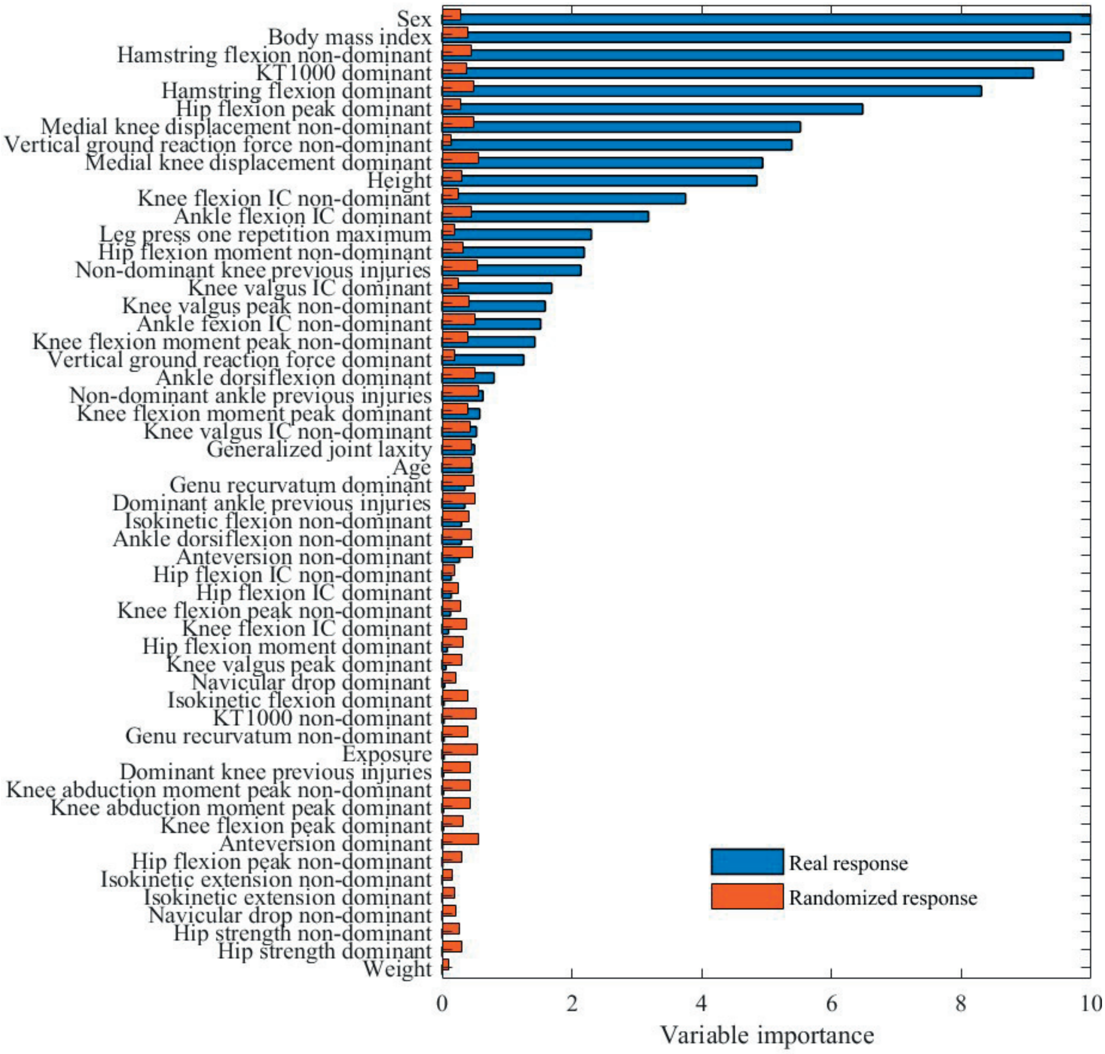

Figure 3. Variable importances for L1-regularized logistic regression. Measured as the number of times each variable was chosen over the ten CV folds. Blue bars correspond to the results with real response, red to the randomized response

3.3. Logistic regression coefficients

Whenever a variable was chosen to the logistic regression model, the direction of the coefficient was extremely consistent, always either positive or negative. Therefore, over all the folds and a hundred runs, the variable always had a similar effect on the prediction, i.e., it either increased or decreased the risk of injury. Directions of variable coefficients for the ten most often selected variables, as well as those that were found by both models, can be seen in Table 1.

Based on the coefficients, female sex contributes to bigger risk than male (male=1, female=2 in data) as well as larger BMI, lower height, and higher leg press 1RM result. Higher hamstring flexibility and vGRF of both legs increase the risk of injury. The higher value of KT1000 of dominant leg as well as higher hip flexion peak and knee flexion at IC of non-dominant leg also contribute to the injury risk. Less ankle plantar flexion (negative values) and larger knee valgus angle (negative values) of the dominant leg contribute to the higher risk. Interestingly for medial knee displacement, the direction was different between the legs. For non-dominant leg, higher medial knee displacement increased the risk but for dominant leg, a lower value increased it.

Table 1. The number of coefficients with positive, negative and zero values over the ten folds and hundred runs.

Variable	Positive	Negative	Zero
Sex	0	999	1
Body mass index	968	0	32
Hamstring flexion non-dominant	957	0	43
KT1000 dominant	911	0	89
Hamstring flexion dominant	831	0	169
Hip flexion peak dominant	648	0	352
Medial knee displacement non-dominant	552	0	448
Vertical ground reaction force non-dominant	539	0	461
Medial knee displacement dominant	0	494	506
Height	0	485	515
Knee flexion IC non-dominant	375	0	625
Ankle flexion IC dominant	318	0	682
Leg press one repetition maximum	230	0	770
Knee valgus IC dominant	0	169	831
Vertical ground reaction force dominant	126	0	874

3.4. Consistent injury predictors chosen by both methods

The following ten variables were suggested as consistent injury predictors ($p<0.01$) by both models: sex, body mass index, hamstring flexibility (non-dominant leg), KT1000 (dominant leg), hamstring flexibility (dominant leg), medial knee displacement (dominant leg), height, ankle (plantar) flexion at IC (dominant leg), leg press one repetition maximum (1RM), and knee valgus at IC (dominant leg).

4. Discussion

The purpose of this study was to utilize predictive machine learning methods to detect variables with predictive power on sport injuries. Multiple injury risk factors have been recognized in previous explanatory studies, but the predictive power of these variables remains unclear until tested on independent data. We presented a framework that detects consistent injury predictors in a data-driven manner and validates their predictive power on independent test data. This approach can be used for finding new hypotheses for injury risk factors as well as confirming the predictive power of previously recognized risk factors. Any new hypotheses should then be confirmed by domain experts in future studies, utilizing explanatory methods as well.

Despite the low predictive accuracy (AUC=0.65), a set of ten consistent injury predictor variables was detected by both models. The obtained AUC score is in line with the previous studies [6, 9, 16, 17] and confirms the difficulty of predicting sport injuries. A recently published predictive analysis study that compared different methods and their injury prediction accuracies, obtained an AUC score of 0.747 when predicting lower extremity muscle injuries in 132 male professional soccer and handball players [9]. A paper by Dower and colleagues [17] utilized time series data and artificial neural networks, achieving AUC scores between 0.75 and 0.80 on average when predicting soft tissue injuries in Australian football players.

Another study found that previously detected risk factors with explanatory power had a very poor predictive performance (median AUC scores 0.57 and 0.52) on hamstring strain injuries in 362 elite Australian footballers [16]. However, this study used a small number of variables in the prediction (three and eight). In addition, previous studies have focused solely on the prediction task, without considering the explainability of the predictive model. Explainable models, assessing, for example, the effect of each variable in prediction, are easier to trust and provide more practical information to the domain experts.

Most of the injury predictor variables suggested in our study are supported by previous research. Our results suggest that female sex, larger BMI, and lower height increased the risk of acute non-contact knee and ankle injury. Previous explanatory research has detected similar associations with lower extremity sport injuries [2, 5, 27, 28]. For muscle flexibility, there are contradicting findings [5, 29]. Our results propose that increased hamstring flexibility of both dominant and non-dominant leg contribute to larger risk of acute non-contact knee and ankle injury.

Concerning the association between muscle strength and sport injury risk, the findings are conflicting [30, 31]. Our study found higher leg press 1RM to be associated with higher injury risk. This could be, for example, because stronger athletes exert greater forces and moments to the joints and muscles during activity; are more mature; and tend to train more and perform at higher levels. Also, our findings that increased knee laxity (KT-1000) and less ankle plantar flexion at IC of the dominant leg contribute to higher injury risk have been previously recognized [32, 33].

Our results suggest that larger knee valgus and medial knee displacement of non-dominant leg increase the risk of acute non-contact knee and ankle injury. Associations between knee valgus loading and risk of lower extremity injuries have been found previously [34]. However, our results also suggested that smaller medial knee displacement of the dominant leg increased the risk, which is contradictory to the results of the non-dominant leg. In the group of non-injures athletes, the medial knee displacement of dominant leg is notably larger than with the non-dominant leg. In the injured group, there is no such difference (see Supplementary Table 1). This side difference is causing the conflicting regression coefficients inside the framework. However, such side differences were not observed in the knee valgus angles. This might be due to the medial knee displacement being more sensitive towards the athlete rotating during landing. In our data, approximately 74% of the athletes rotated towards the side of their dominant leg during VDJ. Another possible explanation might simply be differences in the use of dominant and non-dominant leg.

Our secondary aim was to assess differences between linear and non-linear methods. In our prediction task, the predictive accuracy of the linear L1-regularized logistic regression was slightly better (AUC=0.65) than the accuracy of the non-linear random forest model (AUC=0.63). The difference is, however, negligible for drawing conclusion of their mutual superiority. The suggested injury risk factors were largely the same for both models, but logistic regression suggested a larger set of predictors. Generally, we believe it can be beneficial to utilize a combination of methods to detect the most relevant injury risk factors.

The strength of our approach is that with predictive methods and confirmatory analysis, consistent injury predictors can be detected even from data with weak phenomena. For example, with small datasets the approach can help to avoid findings by chance. Thus, it can be useful in other sport science and medicine studies as well, even though the used data does not necessarily possess high predictive power or strong phenomena itself. Another strength is the prospective data collection of a large number of variables from a large cohort of athletes. Predictive methods utilize computational power and thus enable analysis of all relevant data and do not require exclusion based on prior assumptions. In addition, our study uses a well defined prediction outcome of moderate and severe knee and ankle injuries which risk factors have been established in explanatory research previously.

However, there are also limitations related to the used data. After baseline data was collected, the injury follow-up lasted for 12 months. Many of the collected variables might, however, change notably during this period, especially in young athletes [10]. In the future, more comprehensive data that observes short-term changes in variables should be collected as there can be changes, for example, based on the time in season and weekly training and game loads. Wearable technologies, for example, allow continuous monitoring of athletes. It can be expected that time series data from wearable devices combined with applicable predictive methods will increase the prediction accuracy as the study by Dower et al. indicated [17].

To conclude, in order to have practical value in the clinical assessment of injury risk, the predictive accuracy of the presented models that were trained on the prospective data should be improved. The models were, however, able to detect a set of consistent injury predictors. Thus, the approach can be useful for finding new hypotheses for injury risk factors as well as confirming the predictive power of risk factors found in previous explanatory studies. While the achieved predictive accuracy of our study remained relatively low (AUC=0.65), a set of ten consistent injury predictor variables was detected by both models (sex, body mass index, hamstring flexibility, knee joint laxity, medial knee displacement, height, ankle plantar flexion at initial contact, leg press one-repetition max, and knee valgus at initial contact). The obtained accuracy is in line with previous studies and confirms that predicting sport injuries is a cumbersome task. More research is required to find risk factors that best predict injury and to include more comprehensive data. The obtained performance was similar between the linear and non-linear methods. Future research is needed to assess the suitability and performance of linear versus non-linear methods in sport injury prediction tasks.

Funding

This study was supported by the Finnish Ministry of Education and Culture, and Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (grants 9S047, 9T046, 9U044, 9N053). This work has been carried out in two projects ”Value from health data with cognitive computing” and ”Watson Health Cloud”, funded by Business Finland. Susanne Jauhiainen was funded by the Jenny and Antti Wihuri Foundation (grant 00180121). Jukka-Pekka Kauppi was funded by the Academy of Finland Postdoctoral Researcher program (Research Council for Natural Sciences and Engineering; grant 286019).

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

1 *Jacobsson J, Timpka T, Kowalski J, et al.* Prevalence of musculoskeletal injuries in Swedish elite track and field athletes. Am J Sports Med 2012; 40: 163–169

2 *Emery CA, Rose MS, McAllister JR, Meeuwisse WH.* A prevention strategy to reduce the incidence of injury in high school basketball: a cluster randomized controlled trial. Clin J Sport Med 2007; 17: 17–24

3 *Myklebust G, Holm I, Mæhlum S, et al.* Clinical, functional, and radiologic outcome in team handball players 6 to 11 years after anterior cruciate ligament injury: a follow-up study. Am J Sports Med 2003; 31: 981–989

4 *Bahr R, Holme I*. Risk factors for sports injuries -- a methodological approach. Br J Sports Med 2003; 37: 384–392

5 *Murphy DF, Connolly DAJ, Beynnon BD*. Risk factors for lower extremity injury: a review of the literature. Br J Sports Med 2003; 37: 13–29

6 *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

7 *Bahr R*. Why screening tests to predict injury do not work—and probably never will…: a critical review. Br J Sport Med 2016; 50: 776–780

8 *Shmueli G*. To explain or to predict? Stat Sci 2010; 25: 289–310

9 *López-Valenciano A, Ayala F, Puerta JM, et al.* A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms. Med Sci Sports Exerc 2018; 50: 915–927

10 *Meeuwisse WH, Tyreman H, Hagel B, Emery C*. A dynamic model of etiology in sport injury: the recursive nature of risk and causation. Clin J Sport Med 2007; 17: 215–219

11 *Bittencourt NFN, Meeuwisse WH, Mendonça LD, et al.* Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept. Br J Sport Med 2016; 50: 1309–1314

12 *Robertson S*. Improving load/injury predictive modelling in sport: The role of data analytics. J Sci Med Sport 2014; 18: 25--26

13 *Combrisson E, Jerbi K*. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci

Methods 2015; 250: 126–136

14 *Biran O, Cotton C*. Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI). 2017

15 *Bellazzi R, Zupan B*. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008; 77: 81–97

16 *Ruddy JD, Shield AJ, Maniar N, et al.* Predictive Modeling of Hamstring Strain Injuries in Elite Australian Footballers. Med Sci Sports Exerc 2018; 50: 906–914

17 *Dower C, Rafeli A, Weber J, Mohamad R*. An enhanced metric of injury risk utilizing Artificial Intelligence. In: Proceedings of the 13th annual MIT SLOAN sports analytics conference. 2019

18 *Pasanen K, Rossi MT, Parkkari J, et al.* Predictors of lower extremity injuries in team sports (PROFITS-study): a study protocol. BMJ open Sport Exerc Med 2015; 1: e000076

19 *Harriss DJ, MacSween A, Atkinson G*. Ethical Standards in Sport and Exercise Science Research: 2020 Update. Int J Sports Med 2019; 40: 813–817

20 *Leppänen M, Pasanen K, Kujala UM, et al.* Stiff landings are associated with increased ACL injury risk in young female basketball and floorball players. Am J Sports Med 2017; 45: 386–393

21 *Fuller CW, Molloy MG, Bagate C, et al.* Consensus statement on injury definitions and data collection procedures for studies of injuries in rugby union. Br J Sports Med 2007; 41: 328–331

22 *Boulesteix A-L, Janitza S, Kruppa J, König IR*. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2012; 2: 493–507

23 *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

24 *Breiman L*. Random forests. Mach Learn 2001; 45: 5–32

25 *Fawcett T*. An introduction to ROC analysis. Pattern Recognit Lett 2006; 27: 861–874

26 *Krstajic D, Buturovic LJ, Leahy DE, Thomas S*. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 2014; 6: 10–25

27 *Vanderlei FM, Bastos FN, Tsutsumi GYC, et al.* Characteristics and contributing factors related to sports injuries in young volleyball players. BMC Res Notes 2013; 6

28 *Jones BH, Bovee MW, Harris III JM, Cowan DN*. Intrinsic risk factors for exercise-related injuries among male and female army trainees. Am J Sports Med 1993; 21: 705–710

29 *Boden BP, Dean GS, Feagin JA, Garrett WE*. Mechanisms of anterior cruciate ligament injury. Orthopedics 2000; 23: 573–578

30 *Yamamoto T*. Relationship between hamstring strains and leg muscle strength. A follow-up study of collegiate track and field athletes. J Sports Med Phys Fitness 1993; 33: 194–199

31 *Beynnon BD, Renström PA, Alosa DM, et al.* Ankle ligament injury risk factors: a prospective study of college athletes. J Orthop Res 2001; 19: 213–220

32 *Woodford-Rogers B, Cyphert L, Denegar CR*. Risk factors for anterior cruciate ligament injury in high school and college athletes. J Athl Train 1994; 29: 343–346

33 *Boden BP, Torg JS, Knowles SB, Hewett TE*. Video analysis of anterior cruciate ligament injury: abnormalities in hip and ankle kinematics. Am J Sports Med 2009; 37: 252–259

34 *Hewett TE, Myer GD, Ford KR, et al.* Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a

prospective study. Am J Sports Med 2005; 33: 492–501

Supplementary table: Variables used in this study. For more detailed description see (1) *and supplementary material therein. VDJ stands for the 3D motion analysis for vertical drop jump.*

Variable name (unit)	Test	Mean non-injured	Mean injured	Description
Height (cm)	Anthropometric	174.05±9.22	170.15±8.12	Height
Weight (kg)	Anthropometric	64.63±10.77	64.72±8.90	Weight
BMI (kg/m^2)	Anthropometric	21.26±2.62	22.31±2.15	Body mass index
Anteversion dominant (deg)	Joint anatomy	9.05±6.08	9.06±5.61	Femoral anteversion. Measured with Craig's test (2). The athlete lies in prone position while physiotherapist passively flexes the knee to 90°. The hip is passively rotated internally and externally until the most lateral portion of the greater trochanter is palpable. In this position, the angle between the true vertical and the shaft of the tibia is measured to the nearest degree with a universal goniometer (Absolute+Axis™ Baseline® Evaluation Instruments, White Plains, NY, USA). Dominant leg.
Anteversion non-dominant (deg)	Joint anatomy	9.28±6.24	8.81±5.83	Femoral anteversion (see the description above). Non-dominant leg
Knee valgus IC dominant (deg)	VDJ	6.65±8.66	3.39±7.92	Knee valgus at initial contact (negative value refers to valgus alignment and positive value to varus alignment). Dominant leg.
Knee valgus IC non-dominant (deg)	VDJ	7.61±9.09	3.80±7.55	Knee valgus at initial contact (negative value refers to valgus alignment and positive value to varus alignment). Non-dominant leg.
Knee valgus peak dominant (deg)	VDJ	-4.49±8.29	-6.84±8.71	Peak knee valgus during contact (negative value refers to valgus alignment and positive value to varus alignment). Dominant Leg.
Knee valgus peak non-dominant (deg)	VDJ	-3.98±7.85	-5.33±8.23	Peak knee valgus during contact (negative value refers to valgus alignment and positive value to varus alignment). Nondominant.
Knee flexion IC dominant (deg)	VDJ	28.71±9.82	29.84±9.22	Knee flexion at initial contact. Dominant leg
Knee flexion IC non-dominant (deg)	VDJ	28.52±10.59	30.60±10.98	Knee flexion at initial contact. Non-dominant leg
Knee flexion peak dominant (deg)	VDJ	84.19±10.15	85.13±10.44	Peak knee flexion during contact. Dominant leg
Knee flexion peak non-dominant (deg)	VDJ	84.24±10.41	85.38±9.84	Peak knee flexion during contact. Non-dominant leg

Vertical ground reaction force dominant (N)	VDJ	1182.39±330.66	1251.04±384.07	Peak vertical ground reaction force during contact. Dominant leg
Vertical ground reaction force non-dominant (N)	VDJ	1139.01±311.46	1210.93±342.44	Peak vertical ground reaction force during contact. Non-dominant leg
Knee abduction moment peak dominant (N·m)	VDJ	-32.76±20.63	-34.27±21.87	Peak knee abduction moment during contact. Dominant leg
Knee abduct moment peak non-dominant (N·m)	VDJ	-31.06±17.88	-33.29±19.83	Peak knee abduction moment during contact. Non-dominant leg
Medial knee displacement dominant (mm)	VDJ	24.57±20.93	21.93±18.36	Medial knee displacement during contact. Dominant leg
Medial knee displacement non-dominant (mm)	VDJ	17.74±18.73	22.80±20.88	Medial knee displacement during contact. Non-dominant leg
Hip flexion peak dominant (deg)	VDJ	65.50±11.66	69.12±12.52	Peak hip flexion during contact. Dominant leg
Hip flexion peak non-dominant (deg)	VDJ	65.77±11.75	68.77±12.97	Peak hip flexion during contact. Non-dominant leg
Ankle dorsiflexion dominant (deg)	VDJ	43.32±7.07	41.81±7.30	Peak ankle dorsiflexion during contact. Dominant leg
Ankle dorsiflexion non-dominant (deg)	VDJ	42.30±6.82	41.41 ±7.90	Peak ankle dorsiflexion during contact. Non-dominant leg
Ankle flexion IC dominant (deg)	VDJ	-8.53±10.10	-6.96±11.32	Ankle flexion at initial contact (negative value refers to plantar flexion and positive value to dorsiflexion)
Ankle flexion IC non-dominant (deg)	VDJ	-8.63±9.51	-7.59±9.97	Ankle flexion at initial contact (negative value refers to plantar flexion and positive value to dorsiflexion). Non-dominant leg
Hip flexion IC dominant (deg)	VDJ	43.35±10.83	45.70±10.89	Hip flexion at initial contact. Dominant leg
Hip flexion IC non-dominant (deg)	VDJ	43.50±11.42	46.00±10.62	Hip flexion at initial contact. Non-dominant leg
Knee flexion moment peak dominant (N·m)	VDJ	135.95±44.84	140.36±37.29	Peak knee flexion moment during contact. Dominant leg
Knee flexion moment peak non-dominant (N·m)	VDJ	127.40±45.18	134.01±39.59	Peak knee flexion moment during contact. Non-dominant leg
Hip flexion moment dominant (N·m)	VDJ	205.41±76.08	219.07±84.29	Peak hip flexion moment during contact. Dominant leg
Hip flexion moment non-dominant (N·m)	VDJ	205.77±70.05	219.65±72.86	Peak hip flexion moment during contact.Non-dominant leg
Hamstring flexion dominant (degree)	Muscle extensibility	136.45±15.75	146.35±14.91	Hamstring flexibility. The athlete is lying in supine position, while the hip of the testing leg is fixed at 120° flexion. Three landmarks are placed on the leg: lateral fibular malleolus, lateral femoral epicondyle and the greater trochanter of femur. The knee is extended passively with an 8kg load (a fish scale, Salter Super Samson, Taylor Precision Products, Inc., Illinois,

				USA). A goniometer (HiRes, Baseline® Evaluation Instruments, White Plains, NY, USA) is placed to point of knee joint line and flexibility is measured as static range of motion. Dominat leg
Hamstring flexion non-dominant (degree)	Muscle extensibility	136.85±15.57	146.51±14.29	Hamstring flexibility (see the description above). Non-dominant leg
Hip strength dominant (kg)	Strength	13.00±3.49	12.35±2.74	Maximum isometric hip abductor strength, tested with a hand-held dynamometer (Hydraulic Push-Pull Dynamometer, Baseline® Evaluation Instruments, White Plains, NY, USA). Dominant leg
Hip strength non-dominant (kg)	Strength	12.74±3.56	12.23±3.30	Maximum isometric hip abductor strength (see the description above). Non-dominant leg
Isokinetic extension dominant (kg)	Strength	162.71±38.81	158.44±32.88	Maximum isokinetic strength, tested with Biodex Multi-Joint System Pro dynamometer (Biodex System 4, Biodex Medical Systems, Inc., Shirley, NY, USA), extension of dominant leg
Isokinetic extension non-dominant (kg)	Strength	156.94±36.71	156.91±30.60	Maximum isokinetic strength (see the description above), extension of non-dominant leg
Isokinetic flexion dominant (kg)	Strength	98.01±23.15	97.52±20.09	Maximum isokinetic strength (see the description above), flexion of dominant leg
Isokinetic flexion non-dominant (kg)	Strength	96.05±22.66	97.37±20.16	Maximum isokinetic strength (see the description above), flexion of non-dominant leg
Leg press one repetition maximum (kg)	Strength	170.82±50.49	174.52±45.99	One repetition maximum leg press
Navicular drop dominant (mm)		0.69±0.43	0.65±0.38	Navicular drop of dominant leg
Navicular drop non-dominant (mm)		0.71±0.45	0.66±0.36	Navicular drop of non-dominant leg
Exposure (h)	Team diary	202.53±96.88	188.08±85.26	Total exposure time from practices and games (practice and game hours). Collected individually for each participant.
Age (yr)	Baseline Questionnaire	15.68±0.50	15.93±1.99	Age
Genu recurvatum dominant (deg)	Anatomical characteristics	4.93±4.24	5.58±3.41	Genu regurvatum/knee hyperextension, dominant leg. The athlete lies in supine position and a small bolster is placed under the distal aspect of the tibia. The anterior and posterior portions of the lateral knee joint line are palpated and a mark placed at the midpoint in the sagittal plane. The most prominent aspect of the lateral malleolus and the greater trochanter are palpated and marked. A goniometer (HiRes goniometer, Baseline® Evaluation Instruments, White Plains, NY, USA) is used for measurement. The axis of the

				goniometer is positioned over the mark on the joint line, and the angle formed by a line from the lateral joint line to the greater trochanter. A line from the lateral joint line to the lateral malleolus is measured to the nearest degree with a goniometer.
Genu recurvatum non-dominant (deg)	Joint anatomy	5.17±4.08	5.33±3.56	Genu regurvatum (see the description above). Non-dominant leg
KT1000 dominant (mm)	Joint laxity	6.70±2.04	7.35±2.28	Knee joint laxity, dominant leg. The KT-1000 arthrometer (MEDmetric Corp, San Diego, California) is used to measure anterior-posterior (A-P) knee laxity (A-P displacement of the tibia relative to the femur). The athlete is in a supine position and the knee joint space line is marked medially with the knee in slightly flexed position (25° ± 5°). First, posterior-directed forces are applied to the tibia to establish a zero reference point, followed by anterior-directed forces (134 N) to measure anterior knee joint laxity (mm).
KT1000 non dominant (mm)	Joint laxity	6.75±2.12	7.04±2.30	Knee joint laxity (see the description above). Non-dominant leg
		Median non-injured	Median injured	
Generalized joint laxity (points)	Joint laxity	1	2	Generalized joint laxity, measured using the Beighton scale (3). The athlete is measured for excessive joint laxity at the trunk, the fifth fingers, thumbs, elbows, and knees. The score of four points or more on a scale of 0- 9 indicates generalized joint laxity. Two goniometers (HiRes, Baseline® Evaluation Instruments, White Plains, NY, USA) are used to measure the fifth fingers, elbows, and knees.
Dominant knee previous injuries	Baseline Questionnaire	0	0	Number of previous knee injuries in dominant leg
Non-dominant knee previous injuries	Baseline Questionnaire	0	0	Number of previous knee injuries in non-dominant leg
Dominant ankle previous injuries	Baseline Questionnaire	0	1	Number of previous ankle injuries in dominant leg
Non-dominant ankle previous injuries	Baseline Questionnaire	0	0	Number of previous ankle injuries in non-dominant leg
		Non-injured	Injured	
Sex (male-female)	Baseline Questionnaire	138-119	14-43	Sex

1. Pasanen K, Rossi MT, Parkkari J, Heinonen A, Steffen K, Myklebust G, et al. Predictors of lower extremity injuries in team sports (PROFITS-study): a study protocol. BMJ open Sport Exerc Med. 2015;1(e000076).
2. Ruwe PA, Gage JR, Ozonoff MB, DeLuca PA. Clinical determination of femoral anteversion. A comparison with established techniques. J Bone Joint Surg Am. 1992;74(6):820–30.
3. Beighton PH, Solomon L, Soskolne CL. Articular mobility in an African population. Ann Rheum Dis. 1973;32(5):413.