# Injury Location Prediction in Professional Athletes: A Demographic and Positional Analysis

**Ranveer Patel**

*Received February 4, 2025*
*Accepted February 12, 2026*
*Electronic access March 15, 2026*

**Background / Objective**
Injuries in sports impose substantial physical and economic burdens on professional athletes and sporting organizations. While prior research has predominantly focused on predicting injury risk or injury occurrence, little attention has been given to predicting the anatomical location of injuries once an injury has occurred. This study addresses this gap by examining whether injury location can be predicted using demographic and positional features alone.

**Methods**
Injury records from four major North American professional sports leagues—Baseball, Basketball, Football, and Hockey—spanning 2005–2024 were aggregated from publicly available databases. Raw injury descriptions were manually reclassified into four mutually exclusive anatomical categories. A Random Forest classifier was subsequently trained using age, height, weight, sport, and player position as input features. Model performance was evaluated under multiple validation regimes using balanced accuracy and macro-averaged F1 score.

**Results**
In a global four-class classification task, the model achieved a balanced accuracy of roughly 0.52 and a macro F1 score of 0.51. Performance substantially improved under sport-specific and hierarchical classification settings, with balanced accuracy exceeding 0.70 in some cases. However, group-aware validation by athlete resulted in a marked performance decline, with balanced accuracy near 0.33, indicating limited generalization to unseen players.

**Conclusions**
These results demonstrate that injury location prediction is feasible under constrained conditions but fundamentally limited by sparse feature sets. The findings emphasize the importance of richer biomechanical and workload-related data for practical injury modeling applications.

**Keywords:** Sports injuries; injury location prediction; machine learning; random forest; sports analytics; biomechanics

## Introduction

### Background and Context

Sports injuries are a persistent challenge in professional athletics, affecting athlete performance, career longevity, and organizational finances[1].

As data availability has increased, predictive modeling approaches have become more common in sports medicine and analytics, primarily targeting injury risk estimation and prevention strategies[2,3].

These approaches often leverage detailed workload, exposure, and physiological data to forecast injury occurrence.

### Problem Statement and Rationale

In contrast to injury risk prediction, comparatively little research has examined whether the anatomical location of an injury can be predicted once an injury has already occurred.

Injury location prediction constitutes a distinct analytical task, focusing on characterizing injury patterns rather than forecasting injury onset.

Understanding likely injury locations may support injury surveillance, rehabilitation planning, and post-injury resource allocation.

### Significance and Purpose

Identifying whether injury location can be inferred from widely available demographic and positional features is valuable for establishing predictive limits in sports injury modeling.

This study contributes by explicitly evaluating injury location prediction under constrained informational conditions, rather than proposing a preventative or causal framework.

### Objectives

The primary objective of this study is to assess whether injury location can be predicted using demographic and positional variables alone, conditional on injury occurrence. A secondary objective is to evaluate how performance varies under different validation schema, including sport-specific and group-aware evaluation.

### Scope and Limitations

This study is limited to professional athletes in four North American sports and relies exclusively on publicly available demographic and positional features. Workload, biomechanical, and exposure-related variables are not included. As a result, the analysis is not intended to support injury prevention or clinical decision-making.

### Theoretical Framework

The study is grounded in a constrained predictive modeling framework, which evaluates the extent to which patterns can be learned from limited feature spaces. This framework emphasizes identifying structural limits to prediction rather than optimizing predictive performance at all costs.

### Methodology Overview

A supervised machine learning approach was employed, using a Random Forest classifier trained on professional sport injury data. Multiple evaluation strategies were used to assess both predictive performance and generalization behavior.

## Literature Review

Research on sports injury modeling has substantially expanded over the past two decades, driven by the increased availability of athlete performance data and advances in statistical and machine learning methodologies. A majority of prior work has focused on injury risk prediction, aiming to estimate the likelihood that an athlete will sustain an injury within a given time frame. These models are typically motivated by injury prevention and load management strategies in professional and elite sports.

A dominant theme in the literature is the importance of workload-related variables, including training volume, match exposure, and intensity. Multiple studies have demonstrated that abrupt changes in workload—rather than absolute workload levels—are strongly associated with injury risk[4,5]. This relationship has been observed across a range of professional team sports and has motivated load-management frameworks[6]. Machine learning approaches, including Random Forests and gradient boosting methods, have been applied to these data with moderate success when rich longitudinal features are available[7,8]. Beyond workload, demographic and anthropometric variables such as age, height, and weight have been examined as contributors to injury susceptibility. Aging has been linked to increased injury risk and prolonged recovery times in professional athletes[9], while body mass and stature influence collision forces and biomechanical loading, particularly in contact sports. However, these features are generally weaker predictors of injury risk when considered in isolation and are typically used as complementary inputs in injury prediction models[10,11]. Player position has consistently been identified as a significant determinant of injury patterns, reflecting differences in biomechanical demands, exposure frequency, and contact intensity[12,13]. For example, injury profiles differ substantially between pitchers and batters in baseball, linemen and skill-position players in football, and guards versus forwards in basketball[14]. As a result, positional information is commonly incorporated into injury risk models, either directly or through sport-specific subgroup analyses.

Despite extensive research on injury risk, injury location prediction has received comparatively little attention as a standalone modeling task. Existing studies typically report injury locations descriptively or aggregate injury sites into broad categories for reporting purposes, rather than attempting to predict anatomical location directly. When injury location is modeled, it is often treated as a secondary outcome within broader risk prediction frameworks and is rarely evaluated independently[15]. Moreover, most prior work relies on sport-specific datasets enriched with detailed tracking, workload[16], or medical information that is not consistently available across leagues or historical time spans. This limits cross-sport comparability and reproducibility, particularly for studies aiming to analyze long-term trends or publicly accessible data. As a result, there remains a gap in understanding whether injury location patterns can be learned from widely available demographic and positional features alone, and how such models generalize across athletes and sports. The present study addresses this gap by framing injury location prediction as a constrained predictive modeling problem, explicitly conditioning on injury occurrence and limiting the feature set to broadly available demographic and positional variables. By evaluating performance under global, sport-specific, hierarchical, and group-aware validation settings, the study seeks to quantify both the feasibility and structural limitations of injury location prediction under sparse informational conditions.

## Methods

### Research Design

This study adopts an observational design using historical professional sports injury records.

### Sample

The sample consists of professional athletes from Baseball, Basketball, Football, and Hockey leagues between 2005 and 2024. Each sample observation corresponds to a recorded injury event.

### Data Collection

Injury records were aggregated from publicly available professional sports injury databases and league-specific statistical repositories. All data sources are publicly available—no proprietary datasets were used[17].

### Variables and Measurements

The target variable is injury location, categorized into four mutually exclusive anatomical regions: Knee, Lower Body (Non-Knee), Torso and Head, and Upper Extremities. Predictor variables include age, height, weight, sport, and player position.

### Injury Reclassification

Raw injury descriptions vary widely in specificity and frequently reference overlapping anatomical regions. To reduce label ambiguity and enforce mutual exclusivity, injuries were manually reclassified into four anatomically distinct categories according to a predefined schema (Table 1). Although injuries within the same anatomical region may have different biomechanical etiologies, the objective of this study is anatomical localization rather than injury mechanism.

**Table 1** Injury reclassification schema mapping raw injury descriptions to four anatomical categories.

| Meta Class | Original Classes |
|---|---|
| Upper Extremity | Arm, Shoulder, Elbow, Wrist, Hand |
| Lower Body (Non-Knee) | Leg, Foot, Ankle, Thigh, Groin, Hamstring |
| Torso/Head | Rib, Chest, Abdomen, Oblique, Head, Neck |
| Knee | Knee, Patella |

### Missing Data and Preprocessing

Missingness across key variables was assessed prior to model training (Figure 1). Missingness was minimal (<0.5%), with only the General Injury field exhibiting missing values (0.332%). Rows containing missing entries were removed without imputation. Numeric features were retained without normalization due to the tree-based nature of the model, and categorical variables were one-hot encoded. Player identifiers were retained exclusively for group-aware validation and were not used as predictive features. After filtering, class distributions remained moderately imbalanced. To reduce dominance by frequent classes, class counts were capped per category. Class-weighted training was also evaluated as an alternative approach. Synthetic resampling techniques such as SMOTE were considered but avoided to prevent generating anatomically implausible injury instances from sparse demographic features.

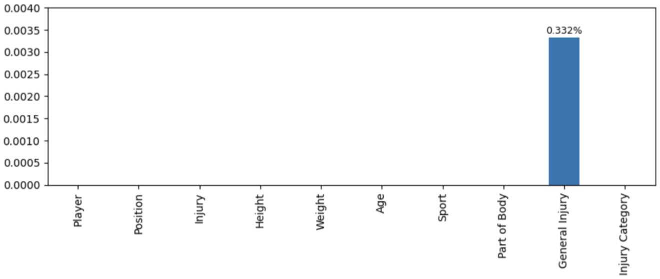

**Fig. 1** Missingness across variables in the raw dataset.

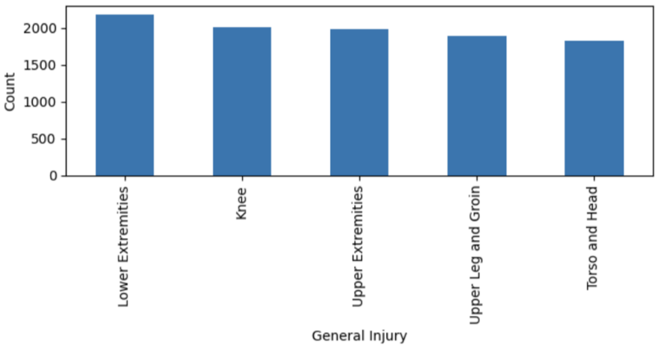

**Fig. 2** Class distribution of injury locations after reclassification.

### Model Selection and Training

A Random Forest[18] classifier was selected due to its suitability for tabular data, ability to model nonlinear feature interactions, and robustness to multicollinearity among predictors such as height and weight[9]. Class weights were applied inversely proportional to class frequency to mitigate the effects of class imbalance. Gradient boosting methods (e.g., XGBoost) can provide strong performance on tabular data; however, this study prioritized interpretability, reproducibility, and minimal dependency complexity suitable for a student-led, fully open workflow. Random Forests also offer robust baselines under sparse feature settings and allow direct confusion-matrix-based error analysis and permutation importance.

The model was trained using 400 decision trees with a maximum depth selected to balance expressiveness and overfitting risk. Hyperparameters were chosen conservatively based on empirical performance stability rather than exhaustive optimization. Hyperparameters were selected using a constrained search over a small grid of candidate values to balance runtime and overfitting risk. Candidate values included $n_{\text{estimators}} \in \{200, 400\}$, maximum depth $\in \{20, 40, \text{None}\}$, and minimum samples per split $\in \{2, 5, 10\}$. Models were compared using balanced accuracy under stratified validation, and the final configuration was chosen based on stable validation performance and interpretability. Class weighting ("balanced") was used throughout. Alternative model families (e.g., neural networks and linear classifiers) were explored during preliminary experimentation but were excluded from the final analysis due to consistently inferior performance under identical validation settings.

To further mitigate overfitting, multiple evaluation regimes were employed, including group-aware validation and hierarchical classification. These approaches provide insight into generalization behavior beyond conventional random train–test splits.

### Evaluation Metrics and Validation

Model performance was evaluated using balanced accuracy and macro-averaged F1 score. Balanced accuracy was selected to account for class imbalance by averaging recall across classes, while macro F1 score provides equal weighting to each class regardless of frequency. Four evaluation settings were employed:

1. Global evaluation using a stratified random train–test split
2. Sport-specific evaluation with models trained and tested within individual sports
3. Group-aware validation using player-level splits to assess generalization to unseen athletes
4. Hierarchical classification, in which injuries were first classified as upper or lower body before subclass prediction

Confusion matrices were generated for each setting to analyze misclassification patterns and identify sources of ambiguity between anatomically adjacent classes.

### Ethical Considerations

All data used in this study were publicly available and de-identified. No human subjects research or private medical records were involved, and no ethical approval was required.

## Results

### Global Performance

Under a global four-class classification setting, the Random Forest model achieved a balanced accuracy of approximately 0.52 and a macro F1 score of 0.51. The confusion matrix (Figure 3) reveals substantial overlap between anatomically adjacent classes, particularly between knee injuries and non-knee lower-body injuries. This pattern reflects shared biomechanical mechanisms and overlapping injury descriptions that persist despite reclassification.

Figure 3 presents the confusion matrix for the global four-class injury location prediction task. The matrix exhibits clear diagonal dominance across all classes, indicating that the model captures nontrivial anatomical structure rather than collapsing to majority-class prediction. Knee and lower-body injuries demonstrate the highest true positive counts, reflecting both their higher prevalence and clearer biomechanical signatures. Misclassifications occur predominantly between anatomically adjacent categories. Knee injuries are most frequently confused with non-knee lower extremity injuries, while torso and head injuries are commonly misclassified as upper extremity injuries. These error patterns are consistent with overlapping injury descriptions and shared biomechanical mechanisms, suggesting that classification errors are driven by intrinsic ambiguity rather than model instability. Importantly, confusion between upper- and lower-body regions is comparatively rare, indicating that the model effectively learns coarse anatomical separation even under a constrained feature set.

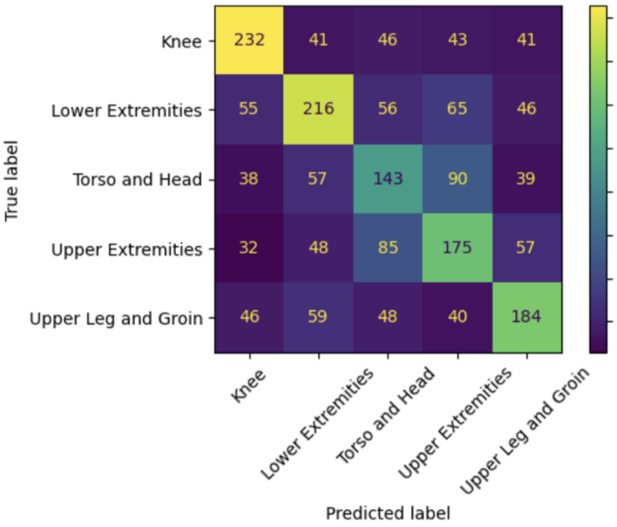

**Fig. 3** Global confusion matrix for four-class injury location prediction.

### Sport-Specific Evaluation

When models were trained and evaluated within individual sports, performance improved across several metrics. Balanced accuracy exceeded 0.70 in certain sports (Figure 4), indicating that sport-specific biomechanics and exposure patterns provide meaningful contextual constraints. These improvements suggest that aggregating across sports introduces heterogeneity that obscures injury patterns unique to each sport. However, sport-specific gains varied considerably, and performance remained inconsistent across injury classes. These results indicate that while sport context improves discrimination, demographic and positional features alone remain insufficient for robust injury localization across all anatomical regions.

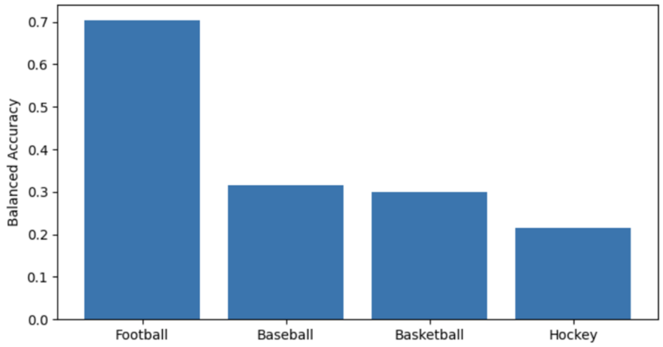

**Fig. 4** Sport-specific balanced accuracy.

### Group-Aware Validation

Group-aware validation was performed by withholding entire athletes during testing to assess generalization to unseen players. Under this regime, balanced accuracy declined sharply to approximately 0.33, only modestly above chance. The confusion matrix (Figure 5) demonstrates increased misclassification across all injury categories. This performance degradation indicates that the global model captures athlete-specific patterns rather than generalized injury mechanisms. Such reliance limits practical applicability and highlights the importance of incorporating features that generalize across individuals, such as workload histories or biomechanical measurements.

### Hierarchical Classification

Hierarchical classification improved performance by first separating injuries into upper versus lower body categories before predicting specific anatomical locations. This approach reduced ambiguity between adjacent classes and improved balanced accuracy relative to the global model (Figure 6). The results suggest that coarse anatomical separation reduces classification complexity and aligns more closely with biomechanical distinctions. However, even under this constrained framework, performance remained sensitive to athlete-specific patterns and did not fully resolve generalization limitations.

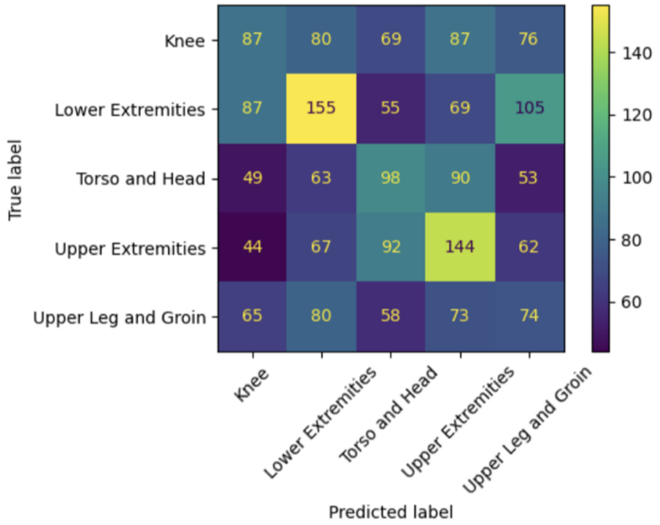

**Fig. 5** Group-aware confusion matrix for unseen athletes.

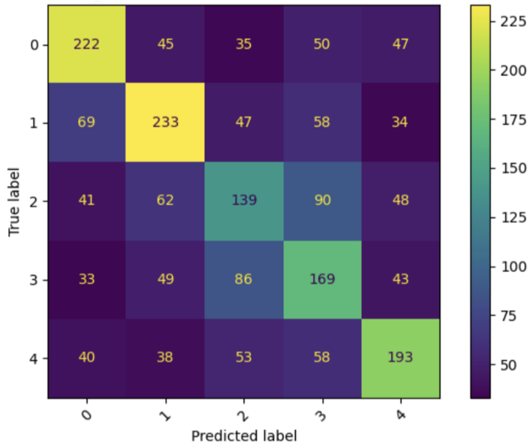

**Fig. 6** Hierarchical confusion matrix.

### Feature Importance

Permutation feature importance analysis revealed that sport and player position contributed more predictive signal than age, height, or weight (Figure 7). Anthropometric features exhibited relatively low importance, consistent with prior findings that demographic variables alone are weak predictors of injury outcomes. The overall magnitude of feature importance values was modest, reinforcing the conclusion that the feature space is limited in informational richness. These results support the interpretation that improved performance would require inclusion of workload, exposure, or biomechanical variables rather than further tuning of demographic inputs.

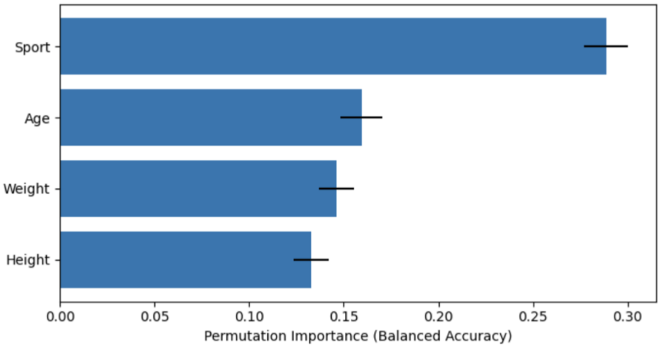

**Fig. 7** Permutation feature importance grouped by base feature.

### Performance Summary

A comparison of balanced accuracy and macro F1 score across evaluation settings is shown in Figure 8. Performance improved as contextual constraints increased (sport-specific and hierarchical settings) but declined under group-aware validation. This trade-off illustrates the tension between model expressiveness and generalization, emphasizing the limitations of sparse feature sets for injury location prediction.

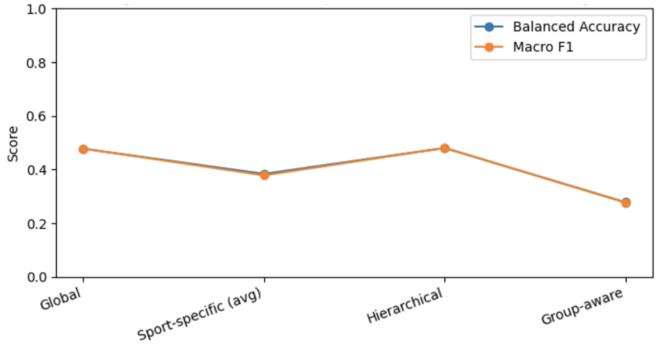

**Fig. 8** Performance comparison across evaluation settings.

## Discussion

The results demonstrate that injury location prediction using demographic and positional features alone is feasible but limited[19]. Improvements observed under sport-specific and hierarchical constraints indicate that biomechanical context plays a meaningful role, while group-aware validation underscores the difficulty of generalizing across athletes. Importantly, the model does not provide causal explanations[20] or actionable prevention strategies. Instead, it illustrates the upper bound of predictive performance achievable using coarse, publicly available features. These findings reinforce prior literature emphasizing the importance of workload, movement patterns, and physiological data in injury modeling.

The structured misclassification patterns observed in the global confusion matrix further clarify the limits of demographic-based injury localization. Errors occur primarily between neighboring anatomical regions rather than across coarse body divisions, indicating that the model learns meaningful spatial structure but lacks the resolution to distinguish fine-grained injury sites. This suggests that demographic and positional features encode broad biomechanical context but are insufficient for precise localization[20].

These findings imply that the observed performance ceiling is driven by feature sparsity rather than model choice. Without access to workload, movement, or exposure data, injury location prediction remains fundamentally constrained[21], regardless of classifier complexity.

## Conclusion

This study evaluated the feasibility of predicting injury location in professional athletes using demographic and positional variables, conditional on injury occurrence. While tree-based models demonstrated modest predictive ability, performance was constrained by limited feature richness and poor generalization to unseen athletes. Future work should incorporate biomechanical, workload, and temporal features to improve both predictive performance and practical relevance. Injury location prediction may serve as a complementary analytical tool, but it should not be interpreted as a standalone injury prevention solution. The primary contribution of this study is not a high-performing predictive model, but a quantitative demonstration of the structural limits of injury location prediction under sparse, publicly available feature sets.

## Acknowledgements

The author thanks the National High School Journal of Science reviewers for their constructive feedback. No external funding was received for this study.

## References

1 G. S. Bullock, E. Murray, J. Vaughan and S. Kluzek, *Scientific Reports*, 2021, **11**, 1–11.
2 T. J. Gabbett, *British Journal of Sports Medicine*, 2016, **50**, 273–280.
3 M. K. Drew and C. F. Finch, *Sports Medicine*, 2016, **46**, 861–883.

4 B. T. Hulin, T. J. Gabbett, D. W. Lawson, P. Caputi and J. A. Sampson, *British Journal of Sports Medicine*, 2016, **50**, 231–236.

5 D. L. Carey, P. Blanch, K. L. Ong, K. M. Crossley, J. Crow and M. E. Morris, *International Journal of Sports Physiology and Performance*, 2018, **13**, 1–8.

6 J. Windt and T. J. Gabbett, *British Journal of Sports Medicine*, 2017, **51**, 428–435.

7 A. Rossi, L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernández and D. Medina, *PLOS ONE*, 2018, **13**, e0201264.

8 H. V. Eetvelde, J. Mendonça, F. Ley, J. Seil and T. Tischer, *Journal of Experimental Orthopaedics*, 2021, **8**, 1–14.

9 B. T. Feeley, S. Kennelly, R. P. Barnes, M. S. Muller, B. T. Kelly, S. A. Rodeo and R. F. Warren, *The American Journal of Sports Medicine*, 2008, **36**, 1597–1603.

10 N. F. N. Bittencourt, W. H. Meeuwisse, L. D. Mendonça, A. Nettel-Aguirre, J. M. Ocarino and S. T. Fonseca, *British Journal of Sports Medicine*, 2016, **50**, 1309–1314.

11 W. H. Meeuwisse, P. E. H. Tyson, J. L. Hagel and P. A. Emery, *Clinical Journal of Sport Medicine*, 2007, **17**, 215–219.

12 J. L. Dragoo, H. J. Braun, J. L. Durham, M. R. Chen and A. H. S. Harris, *The American Journal of Sports Medicine*, 2012, **40**, 990–995.

13 J. Orchard and H. Seward, *British Journal of Sports Medicine*, 2002, **36**, 39–44.

14 N. Rommers, R. Rössler, L. Goossens, R. Vaeyens, M. Lenoir and E. Witvrouw, *The American Journal of Sports Medicine*, 2020, **48**, 1074–1084.

15 B. Clarsen and R. Bahr, *British Journal of Sports Medicine*, 2014, **48**, 510–512.

16 T. Soligard, M. Schwellnus, J.-M. Alonso *et al.*, *British Journal of Sports Medicine*, 2016, **50**, 1030–1041.

17 E. Verhagen and W. van Mechelen, *Oxford Textbook of Sports Medicine*, Oxford University Press, 2010.

18 F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

19 R. Whiteley and S. Malone, *Journal of Orthopaedic & Sports Physical Therapy*, 2020, **50**, 213–215.

20 A. Hulme, S. J. Finch, M. R. P. Cassidy *et al.*, *Sports Medicine*, 2017, **47**, 1031–1052.

21 R. Bahr and T. Krosshaug, *British Journal of Sports Medicine*, 2005, **39**, 324–329.

22 F. M. Impellizzeri, E. Rampinini and S. Marcora, *Journal of Sports Sciences*, 2005, **23**, 583–592.