**Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing sport injury prediction models in sport**

Garrett S. Bullock, PT, DPT, DPhil,1,2 Joseph Mylott,1,3 Tom Hughes, PT, PhD,4,5 Kristen F. Nicholson, PhD,1 Richard D. Riley, PhD,6 Gary S. Collins, PhD7,8

1. Department of Orthopaedic Surgery, Wake Forest School of Medicine, United States
2. Centre for Sport, Exercise and Osteoarthritis Research *Versus Arthritis,* University of Oxford, United Kingdom
3. Baltimore Orioles Baseball Club
4. Manchester United Football Club, Manchester, United Kingdom
5. Department of Health Professions, Manchester Metropolitan University, Manchester, United Kingdom
6. Centre for Prognosis Research, School of Medicine, Keele University, United Kingdom.
7. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, United Kingdom
8. Oxford University Hospitals NHS Foundation Trust, United Kingdom

**Corresponding author:**

Garrett S. Bullock PT, DPT, DPhil

Department of Orthopaedic Surgery

Wake Forest School of Medicine

475 Vine St

Bowman Gray Medical Building

Winston-Salem, NC, 27101

(336) 716-4264

gbullock@wakehealth.edu

**Abstract**

**Objective:** To evaluate the methodological conduct and completeness of reporting of musculoskeletal injury prediction models in sport.

**Study Design:** Systematic Review

**Data Sources:** Search of three databases in June, 2021.

**Eligibility Criteria for Selecting Studies:** Studies were included if they: 1) Predicted sport injury; 2) Used regression, machine learning, or deep learning models; 3) Written in English; 4) Peer reviewed. Studies were excluded if they: 1) Investigated risk factors without subsequent prediction; 2) Evaluated a new method; 3) Reviewed the literature; 4) Predicted concussion exclusively.

**Results:** Twenty-nine studies (200 models) were included. 10% of studies utilized only regression methods, 49% only machine learning, and 41% both regression and machine learning approaches. All models developed a prediction model and no models externally validated a prediction model. 2% of models (7% of studies) were low risk of bias and 98% of models (93% of studies) were high or unclear risk of bias. 2% of models performed an *a priori* sample size calculation. 63% of models performed internal validation. 95% reported discrimination and 2% reported calibration. 7% of models reported model equations for statistical predictions and no machine learning studies reported code or hyperparameters.

**Conclusion:** Existing sport injury prediction models were poorly developed and have a high risk of bias. No models could be recommended for use in practice.The majority of models were developed with small sample sizes, inadequate assessment of model performance, and poorly reported. To create clinically useful sports prediction models, considerable improvements in methodology and reporting are urgently required.

**Registration:** Open Science Framework:<https://osf.io/52mzn/>

**Key Words:** Musculoskeletal,Risk, Regression, Machine Learning, Calibration, Discrimination

**What is Already Known:**

* An increasing number of prediction models are being developed and implemented in sports medicine.
* The overall methodological conduct and transparency of these models is currently weak. This means that it is difficult for clinicians to identify appropriate and useful models for practice. It is also difficult for researchers to select models suitable for further evaluation, external validation, or those that could inform the development of new ones.

**What are the New Findings:**

* 98% of sport injury prediction models (and 79% of studies) were rated as high risk of bias.
* Current injury prediction models in sport are poorly developed and evaluated nor applicable for further model validation by independent researchers due to high risk of bias and lack of reported equations and/or code.
* To create clinically useful sports injury prediction models, improving methodological conduct and transparent reporting is a priority.

**Introduction**

Participating in sport carries inherent risk, resulting in musculoskeletal injuries1 with repercussions of increased healthcare service burden,2 persistent joint pain,3 4 post traumatic osteoarthritis,5 and decreased quality of life.6 Sports medicine clinicians perform periodic health examinations7 to inform the prescription of universal, selective, or individualized interventions that are designed to mitigate injury risk.8 However, while these types of management decisions are reliant upon clinical knowledge, skills and intuition,9 it is not always clear whom is at risk or will sustain an injury.10 As a result, sport organizations and clinicians implement heuristics (known as ‘rules of thumb’)11 and clinical prediction models12 to provide a more objective insight into an athlete’s injury risk.9 13

The development of clinical prediction models in sports medicine has been low compared to those developed in general medical fields, but this is now increasing.14 15 In particular, clinical prediction models use multiple variables (also known as predictors) to calculate the risk or probability of a health outcome (such as musculoskeletal injury) for an individual.16 Prediction models have traditionally been developed through statistical approaches such as regression, but more increasingly they are being developed using machine learning methods.15 17 Machine learning methods generally impose minimal modelling assumptions,18 potentially offering increased flexibility to analyze higher order interactions and non-linear associations,19 but at the expense of requiring larger sample sizes.20 However, in terms of predictive accuracy there is often no discernible difference between the two broad class of approaches.21 Regardless of the modelling method, poor methodological conduct and reporting leading to high risk of bias has been observed in studies in other healthcare fields.21 22

Clinical prediction model quality and transparency are essential to have any chance of further evaluation and implementation in to patient care.9 13 23 As prediction models may perform differently in diverse athlete or sport settings,24 depending on the intended athlete population, models need to be evaluated in individuals from different geographical regions or competition levels.24 25 Models should also undergo regular evaluation to maintain predictive accuracy in contemporary populations. Further, the full model must be reported (model availability or code provided in the context of machine learning), to allow independent researchers to evaluate and understand if a prediction model is clinically useful in their target population, and can be implemented clinically. While sports injury prediction models have been previously evaluated,14 15 these reviews were limited in their data extraction and evaluation of methodology, preventing a clinician’s ability to fully critically appraise these models and potential usefulness within their clinical setting.26 Therefore existing models may not be fit for purpose, potentially even harming patient care, and often lack the transparency to facilitate evaluation by independent research groups. To address this, this article provides a systematic review to evaluate the methodological conduct and completeness of reporting of musculoskeletal injury prediction models in sport. Our aim is to provide sports medicine clinicians and scientists a critical evaluation of the current state of sports injury prediction models, to better inform their conduct and clinical usefulness.

**Methods**

*Study Design*

The Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies Checklist (CHARMS) 27 was used to frame the systematic review and develop data the extraction protocol. The *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* guideline15 was used to guide the reporting of this systematic review. Please refer to Appendix 1, for framing of the review through the CHARMS Checklist.27 This review was prospectively registered on the Open Science Framework (OSF), <https://osf.io/52mzn/>.

*Inclusion and Exclusion Criteria*

Studies were included if they developed or evaluated a model that predicted injury in sport. Injury prediction had to refer to predicting either the occurrence, the severity, or the type of musculoskeletal injuries.14 A musculoskeletal injury was defined as an injury to a tendon, ligament, nerve, muscle, or bone that occurs during any sport training, matches, competition, or event.7

Inclusion criteria consisted of:

* Described the development or validation of a clinical prediction model for individualized musculoskeletal injury prediction using two or more predictors in a sport setting
* Prediction models developed using regression (e.g., logistic or cox regression), machine learning (e.g., tree-based methods, support vector machines, ensemble methods), or deep learning (e.g., artificial neural networks)
* English language only
* Peer reviewed articles (i.e., preprints not included)

Exclusion criteria consisted of:

* Studies that investigated associations of individual factors
* A new modeling approach was introduced (hence a methodological focus)
* Reviews of the literature
* Studies for which we were unable to obtain the full text.
* Studies that predict concussion only

*Search Strategy*

Databases: Medline (Ovid), EMBASE, CENTRAL, Sportdiscus and CINAHL (Cumulative Index for Allied Health Literature) electronic databases were searched from inception to June 24, 2021 to identify relevant articles. All searches were documented including terms used and number of hits or articles obtained. Bibliographies of selected studies were hand searched to identify relevant articles not found by the search strategies. Medical subject headings (MeSH) and keywords were utilized including, “ROC Curve”, “risk scores”, “predict”, “logistic”, “regression”, “injury”, and “sport” (Appendix 2)

*Study Selection*

Title and abstracts were first screened by two blinded reviewers (GB, JM) using inclusion and exclusion criteria. Two blinded reviewers (GB, JM) performed full-text review following title and abstract screening. Any conflicts were first discussed between both reviewers. If a consensus could not be reached, another reviewer (TH), was utilized to determine final study eligibility. Following full-text review, a hand search was performed for any studies missed within the initial search (Figure 1).

**<Figure 1.** PRISMA Flow Diagram>

Data Extraction

Data were extracted into a customized Excel spreadsheet (Version 2013, Microsoft, Redmond, Washington, United States) by two independent reviewers (GB, JM) (Table 1). Disagreements concerning data domain placement were resolved by a third reviewer (TH). Raw data were uploaded to the Open Science Framework.

**<Table 1.** Variables extracted, according to Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist**27>**

# *Risk of Bias*

# Risk of bias is defined as shortcomings in the study design, conduct and analysis that can result in the findings being flawed or distorted28. We assessed risk of bias for each model (and also summarized at the study level) with the Prediction model Risk Of Bias ASsessment Tool (PROBAST)28. PROBAST identifies sources of bias in four domains: participants, predictors, outcomes and analysis. A PROBAST assessment was carried out independently by two reviewers (GB, JM). A third reviewer resolved any conflicts (TH).

*Statistical Analyses*

Data were summarized using counts (percentages) and medians (ranges). Depending whether the extracted item is fixed at the study level or varied at the model level (depending on the modelling approach), data were summarized accordingly. All analyses were performed in R version 4.02 (R Core Team (2013).

**Result**

A total of 10,723 titles and abstracts were screened, with 29 studies (reporting the development of 200 models) included (Figure 1).29-57 The 29 studies collectively included 9,067 participants, with 6,474 males (Median: 96; Range: 25-2,322), and 860 females (Median: 105; Range: 26-187). One hundred and fifty-seven (79%) models included soccer athletes,30 35 40 41 43 44 46-48 95 (48%) models included athletes from baseball, floorball, football, gymnastics, handball, korfball, ice hockey, rugby, and/or swimming,31 33 34 37 41 42 45 53-57 seven (4%) models of Australian rules football athletes,32 49 six (3%) models of basketball athletes,37 40 50 52 54 five (3%) models with athletes from undefined sports,29 38 39 45 51 and two (1%) models of volleyball athletes.40 54 One hundred and thirty-four (67%) models included professional or elite athletes,29 30 32 34 35 41 42 48-50 52-55 62 (31%) youth athletes,33 36 37 44 47 11 (6%) college athletes,31 40 43 45 46 51 53 56 57 two (1%) recreational adult athletes,39 50 and two (1%) high school athletes.38 51 One hundred and seventy-seven (89%) models used a prospective cohort design,29 30 33 34 36-38 40 41 43 44 46-50 53 54 56 57 17 (9%) models were retrospective cohorts,31 32 35 42 52 and six (3%) models used a case control study design.39 45 51 55 Eighteen (62%) studies developed twenty (10%) models using only regression,29 31-36 39 40 43 45 46 50-52 54 56 57 four (14%) studies developed 97 (49%) models using only machine learning methods,30 41 47 48 six (21%) studies used single data sets to develop 77 (38%) models using both regression and machine learning models,37 38 42 44 53 55 and one study (3%) developed five models (3%) using regression, machine learning, and deep learning methods in a single data set.49 All 29 studies focused on model development,29-57 and no studies carried out an external validation or model updating (Table 2).

**<Table 2.** Description of Included Studies>

*Risk of Bias of Each Model*

Two (1%) models were rated overall as low risk of bias33 35 and 198 (98%) models were rated as high or unclear risk of bias (Table 3).29-32 34 36-57 For the participants domain, 194 models30-44 46-50 52-54 56 57 were rated as low risk of bias and six (3%) models were rated high risk of bias29 45 51 55. For the predictor domain, 192 (96%) models were rated low risk of bias29 30 32-44 46-50 52 53 57 and eight (4%) models were rated as unclear risk of bias.31 45 51 54-56 For the outcome domain, 181 (91%) models were rated low risk of bias,30 32-41 43 44 46-50 52 53 56 57 seven (3%) models were rated as unclear risk of bias, 29 31 45 51 54 55 and 12 (6%) models rated as high risk of bias.42 For the analysis domain, two (1%) models were rated as low risk of bias33 35 and 198 (98%) models were rated as high or unclear risk of bias.29-32 34 36-57

*Risk of Bias of Each Study*

Two (7%) studies were rated overall as low risk of bias33 35 and 27 (93%) studies were rated as high or unclear risk of bias (Table 3).29-32 34 36-57 For the participants domain, 25 studies30-44 46-50 52-54 56 57 were rated as low risk of bias and four (14%) studies were rated high risk of bias29 45 51 55. For the predictor domain, 23 studies were rated low risk of bias29 30 32-44 46-50 52 53 57 and six studies were rated as unclear risk of bias,31 45 51 54-56. For the outcome domain, 22 studies were rated low risk of bias,30 32-41 43 44 46-50 52 53 56 57 six studies were rated as unclear risk of bias, 29 31 45 51 54 55 and one study rated as high risk of bias42. For the analysis domain, two (7%) studies were rated as low risk of bias33 35 and 27 studies were rated as high or unclear risk of bias.29-32 34 36-57

**<Table 3.** Risk of Bias Using Prediction Model Risk Of Bias ASsessment Tool (PROBAST) Tool>

*Sample Size and Predictors*

The median sample size used to develop the prediction models was 152 (Range: 25 to 2322), and the median number of injuries (outcome to be predicted) was 57 (Range: 5 to 6673). The median number of events per parameter was 2.7 (Range: 0.3 to 78.6). Three (2%) models33 35 36 performed an *a priori* sample size calculation, and one (<1%) of these35 performed a sample size calculation using the newly developed guidance.58 Only one regression model explored whether continuous variables needed to be modelled to account for any non-linearity,35 and others did not justify why they did not. One hundred and thirty-five (68%) models dichotomized continuous predictors prior to developing the model.32 38-41 43-45 56 57 and an additional 20 (10%) models were unclear whether continuous predictors were dichotomized prior to developing the model.29 30 46 Ninety-one (46%) models selected predictors through univariable analyses (i.e., based on their unadjusted association with the outcome) before model development, 32 33 39 44 45 50 52 53 55 57 84 (42%) models were unclear on how predictors were selected before or during model development,34 36 41 42 46 54 nine (5%) studies used recursive cross-validation methods (sic),48 eight (4%) included all predictors,40 43 47 49 three (2%) used backwards selection,29 31 35 three (2%) used least absolute shrinkage and selection operator (LASSO),37 38 and two (1%) included both univariable (before model development) and backwards step wise selection (during model development).51 56 Sixty (36%) machine learning models48 49 41 over sampled to reportedly control for class imbalance, and twenty two (13%) models41 under sampled to control for class imbalance.

*Missing Data*

Five (17%) studies [73 (37%) models] reported missing data33 35 37 38 41 and 24 (83%) studies [127 (63%) models] did not mention missing data. 29-32 34 36 39 40 42-57 Of the five (17%) studies that reported missing data,33 35 37 38 41 two (66%) studies [two (3%) models] performed multiple imputation,33 35 one (3%) study [two (3%) models] performed k-nearest neighbors imputation,37 one (3%) study [68 (93%) models] imputed the mean for each predictor, 41 and one (3%) study [one (<1%) model] did not report having any missing data, thus a complete case analysis was performed.38 Both studies that performed multiple imputation also imputed on the outcome.33 35

*Algorithms Implemented*

In the 18 (62%) studies that developed models only using regression,29 31-36 39 40 43 45 46 50-52 54 56 57 15 (75%) models were developed with logistic regression,29 31 33-35 39 40 43 45 46 50 51 54 56 57 two (10%) models were developed through Poisson regression,32 52 one (5%) model was developed with a latent classification model,36 and one (5%) model was developed with Cox regression.57 In the four (14%) studies that only used machine learning methods,30 41 47 48 27 (28%) models were developed using random forest or decision trees, 30 48 one (1%) model was developed using gradient boosting machines,47 and 68 (71%) models were unclear on the proportion of different machine learning methods employed. 41 In the seven (24%) studies that implemented both regression and machine learning based approaches,37 38 42 44 49 53 55 six (8%) models were logistic regression37 38 42 44 49 55 and one (1%) model was a generalized estimating equation.53 Concerning the machine and deep learning models, 62 (81%) models were random forest or decision trees,37 44 49 53 55 four (57%) models were naïve bayes,38 42 49 55 two (3%) models were k-nearest neighbors,38 42 two (3%) models were support vector machines,49 55 two (3%) models were gradient boosting machine,42 one (1%) model was linear discriminant,38 and one (1%) model was neural network.49

*Validation*

125 (63%) models performed internal validation.30 32 33 35 37 38 41 42 47-49 53 55 Two (2%) models33 35 performed internal validation using bootstrapping, 111 (89%) models used cross-validation,30 32 37 38 41 42 49 55 and 12 (10%) models used split sample methods.47 48 53 The median split sample split was 70% (Range: 70-80) for model development, and 20% (Range: 15-30) for model evaluation. Two (1%) models performed LASSO for shrinkage of coefficients.37 38

*Performance*

Twenty three (79%) studies [194 (97%) models] reported measures of prediction model performance (Table 4).29-39 41-44 46-49 53 55-57 Six (21%) studies [six (3%) models] did not report any prediction model performance measures.40 45 50-52 54 Eighteen (62%) studies [185 (95%) models] reported discrimination,30-33 35 37-39 41-44 46 48 49 53 56 57 three (10%) studies [four (2%) models] reported calibration,33 35 57 one (3%) study [one (<1%) model] reported the Hosmer and Lemeshow test,57 four (10%) studies [four (2%) models] reported R2,33 35 39 46 one (3%) study [one (<1%) model] reported clinical decisions curves and net benefit,35 seven (24%) studies [38 (20%) models] reported overall accuracy,30 31 34 47-49 55 and one (3%) study [one (<1%) model] reported root mean square error approximation.36

<**Table 4.** Prediction Model Performance Statistics for Each Included Study>

*Model Presentation*

Four (14%) studies [four (2%) models] reported model equations for the regression based approaches,29 31 35 50 and one (3%) study [one (<1%) model] partially reported model equation, without the intercept.33 No machine learning studies or models reported tuned hyperparameters within the manuscript, in an appendix, or provided a web link to code or logic. Out of the 28 studies29-49 51-57 (199 models) that published following the release of the TRIPOD statement, two (7%) studies [two (1%) models] cited or discussed the TRIPOD guidelines.35 38 One (3%) study [one (<1%) model] was published prior to TRIPOD development.50

**Discussion**

Only two models were rated at low risk of bias, with 198 models rated as high risk of bias. Following PROBAST, studies are considered to be at high risk of bias overall when one or more appraisal domains are judged to be at high risk.28 When stratified according to PROBAST domains, the poor overall quality of the current evidence could be explained because crucially, the analyses were considered at high risk of bias for all 198 models.

While the performance metrics across these reported 198 models varied, some reported high levels of discrimination,42 accuracy,31 sensitivity,34 39 43 47 and specificity,29 39 44 46 so it would be tempting to assume that some models may offer clinical value. Importantly for clinicians however, our risk of bias assessment suggests that the reported performance metrics for all 198 models are likely to be optimistic and considered as unreliable. Indeed, even if any of these existing models were to be deployed in practice with other athlete groups, it is likely that prediction performance would deteriorate, thus increasing the uncertainty regarding their clinical utility.59

Of the two models judged to be of low risk of bias,35 38 their performance measures were modest, and it is unclear whether they would offer any clinical benefit. Consequently, this means that there are no existing injury prediction models that could be confidently recommended for use in sports medicine practice, and our review has clearly shown that a drive to improve research methodology (with particular emphasis on the conduct of analyses) is urgently required in this potentially important area.

The median sample size events per parameter for these prediction models was 2.7 outcomes (i.e., injury) per predictor parameter. While prediction sample size calculations have moved beyond the ‘events-per-variable’ rule of thumb,60 this calculation still provides insight on the number of predictors in relation to the number of events (i.e., injury). Further, only three models included an *a priori* sample size calculation. Sample size guidance for developing clinical prediction models using regression have recently been proposed and one model in the review followed these principles.58 61 There are no equivalent guidance for sample size of machine learning studies, so until guidance is made available, the aforementioned studies should be considered when calculating the required sample size, bearing in mind these will likely be an underestimate as studies have shown machine learning methods require larger sample sizes compared to regression.20 The small sample sizes, and lack of sample size calculations suggest these models are at risk for overfitting. Overfitting is defined as when the model is too closely fitted to the idiosyncrasies in the data, decreasing the model’s generalizability and usefulness in other samples.62 Greater care is needed to perform *a priori* sample size calculations, to gain appropriate sized datasets and increase prediction generalizability for future athlete injury risk assessment.

Over two thirds of models dichotomized continuous predictors prior to incorporating into the model. This is where, prior to being incorporated into models, predictors derived from continuous measurements are converted into low and high risk groups using arbitrarily chosen or data-driven thresholds. Dichotomizing continuous predictors is discouraged as this method discards information, decreases prediction accuracy and is biologically implausible.63 Continuous predictors should remain continuous and modelled appropriately (e.g., when developed a regression model to consider fractional polynomials or splines).64 Many machine learning methods (e.g., random forests) handle nonlinearity directly, but dichotomizing or categorizing prior to fitting the machine learning method should clearly be avoided. Linearity should also not be assumed as many biological relationships are non-linear in nature (e.g., body mass index and mortality risk),63 and including non-linear transformations has demonstrated improved prediction in sport.65 However, only one model35 assessed for non-linear relationships between continuous predictors and the outcome. Further, almost one in two models used univariable screening for predictor selection prior to any model building. While there may be more predictors available than can be included in the model, univariable selection methods should be avoided as certain predictors may only become significant for inclusion after controlling for other predictors or spurious univariable associations may cause other predictors to be included.66 If predictor selection is required for regression based models to control for overfitting, then approaches based on penalization such as LASSO or elastic net may be considered,67 though recent studies have shown these methods to be unreliable when the sample size is small.68

Just over half of the prediction models performed internal validation, with one in ten models that performed internal validation used a split sample approach. As prediction models are anticipated to have improved prediction performance during development compared to new data,69 it is important to control for this optimism bias through internal validation.70 Internal validation can be performed through bootstrapping or cross-validation methods.13 Split sampling methods have been advised against in the regression framework as inefficient, as well as decreasing the development dataset size (thus increasing the risk of overfitting),70 but are common in machine learning (where data are split into training, tuning and testing). As almost all of the included studies involved small sample sizes, split sampling therefore increased the risk of overfitting to an even greater extent. It is advised to use all available data for model development, and use internal validation methods such as bootstrapping or cross-validation to assess model performance. If sample size permits and multicenter data are available, then internal-external cross-validation can be considered to investigated heterogeneity in model performance across different hospitals (for example).71

Almost two thirds of studies reported discrimination. Discrimination is the ability of a model to differentiate between athletes who will and who will not sustain an injury.72 While this is an important measure of prediction performance, it is only one piece, and does not assess the reliability of these risk estimates. This is important because unreliable predictions could lead to potentially misleading clinical decisions.73 Risk estimates may be too high for all athletes, informing clinicians that all athletes will sustain an injury, or all risk estimates may be too low, informing clinicians that all athletes will not sustain an injury. Therefore, it is important (and widely recommended) that calibration is assessed. Calibration evaluates the accuracy of the predictions (as recommended in the TRIPOD reporting guideline).72 However, of the included models, only three models assessed calibration, with one study using the Hosmer-Lemeshow test, which should be avoided as it provides no information on direction or magnitude of any miscalibration, and is highly influenced by sample size.73

Only 14% of regression studies reported model equations and no machine learning models reported tuned hyperparameters, provided a link code, or logic. Further, only 7% of eligible studies reported or adhered to TRIPOD guidelines.22 Complete reporting is essential to understand the research question and defined outcome, how the model was developed (including data used), and how the model performed.22 Without these key details, it is difficult to critically appraise the methods, leading to unclear risk of bias and an inability to assess the quality of the prediction model. Further, without model transparency, through reporting equations for statistical models, or tuned hyperparameter for machine learning models, models cannot be externally validated or tested. Independent and externally validating prediction models is crucial to evaluating model usefulness,74 and also reducing research waste.22 The lack of reporting of equations, code, and/or hyperparameter tuning is consistent compared to prediction model literature in other healthcare fields, such as oncology and diabetes.21 22 75 These findings suggest that poor prediction model development is inherent to the medical field, and not specific to sports medicine.

Almost half of the machine and deep learning models over or under sampled the outcome to control for class imbalance. Data are indicated to be imbalanced when the proportion of outcome events is unequal. Class imbalance can be an issue when calculating improper performance measures like accuracy as the larger group without the outcome will be disproportionately predicted, distorting prediction accuracy.76 However, concerning prediction, which assesses risk, over or under sampling to control for class imbalance distorts the outcome prevalence, altering predictions and thus if left unaltered (e.g., via recalibration) the model will be miscalibrated. This method will decrease prediction performance in other data sets and ultimately hinders model generalizability and application into clinical practice.26

As with all studies there were limitations. As only three databases were included, some publications may have been missed. However, our search included over 10,000 titles and abstracts, with additional hand searching, suggesting that these overall findings would not have been altered with further study inclusion. Only English studies were included, causing a single language bias. Some studies included only machine learning models, where authors may have included different reporting and methodological parameters compared to regression models. However, there are machine learning reporting guidelines available, such as Minimum information about clinical artificial intelligence modelling (MI-CLAIM) checklist77 which were not cited for any of the machine learning models, suggesting poor transparency and reporting for these studies. The TRIPOD for Machine Learning78 (TRIPOD-AI) reporting guideline is in developmental stages, which will further aid machine learning reporting.

*Conclusions*

The majority of sports injury prediction models were rated at high risk of bias and we found that currently, there are no sports injury prediction models that could be recommended for use in practice. Typically, these models were developed with small sample sizes, and model performance was inappropriately and incompletely evaluated. Further, none of these models have been externally validated. An improvement in the methodological conduct in sports prediction model studies is required with an emphasis upon study analyses, with international collaborative efforts needed to acquire data of sufficient sample size for model development, external validation, and clinical impact assessment.

**Contributions:** GSB, TH, GSC conceived the study idea. GSB, TH, GSC were involved in design and planning. GSB, GSC wrote the first draft of the manuscript. GSB, JM, TH, KFN, RDR, GSC critically revised the manuscript. GSB, JM, TH, KFN, RDR, GSC approved the final version of the manuscript.

**Funding:** GSC was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294).

**Competing Interests:** None

**Ethics Approval:** Not Applicable

**Data Availabilit**y: Extracted data are available through open science framework: <https://osf.io/52mzn/>

**Patient Public Involvement:** A series of symposiums are planned at various sports medicine conferences to help further educate clinicians on this topic.

**Table 1.** Variables extracted, according to Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist**27**

|  |  |
| --- | --- |
| Domain | Key Items |
| Source of Data  | Study design and source |
| Participants | Eligibility; Recruitment; Participant characteristics; sport, gender, country, level of sport |
| Outcome to be predicted | Type of musculoskeletal injury, body part(s), recurrence, follow up time of the outcome |
| Candidate predictors | Number and type of predictors; timing of predictors; whether or not predictors were measured and examiners were blinded to the outcome; handling method of any continuous predictors |
| Sample size | Details of any sample size calculation. The overall sample size and number of outcome events. |
| Missing data | Number of missing data; handling of missing data  |
| Model Development | Modeling method (e.g., logistic regression, random forest, neural network); approach for any predictor selection before model development; method for predictor selection during model development; whether shrinkage performed (for regression models) and if so how; Machine learning tuning hyperparameters |
| Model performance | Calibration and how it was reported (e.g., calibration slope, graphical calibration plot), discrimination, net benefit, other performance measures |
| Model Evaluation | Whether internal validation was performed, if so, what type (e.g., bootstrap, cross-validation, split-sample); details of any external validation performed.In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibration, predictor adjustment for regression models) |
| Results | Whether the final model equation was reported, with intercept and predictor coefficients (for regression models);Whether the code was made available for machine learning models, and were all hyperparameters reported?Whether alternative methods were used to present the model and if so, what types (e.g., nomogram)?Whether a comparison of performance was performed, and distribution of predictors with and without missing data or handling of missing data (e.g., imputation)Whether study was reported according to TRIPOD Guidelines79  |

**Table 2.** Description of Included Studies

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Study | Year | Study Design | Sample Size | Number of Events | Sport | Competition Level | Country | Injured Body Part(s) | Total Number of Models | Prediction Methods |
| Abalo-Nunez | 2018 | Prospective Cohort | 73 | 12 | Gymnastics, "Other" | Elite | Spain | Lower Extremity | 1 | Logistic regression |
| Ayala | 2019 | Prospective Cohort | 96 | 18 | Soccer | Professional | Spain | Hamstring | 18 | Random Forest, Gradient Boost Machine |
| Carbuhn | 2020 | Retrospective Cohort | 27 | 5 | Cross Country | College | USA | Long Bone | 1 | Logistic regression |
| Colby  | 2018 | Retrospective Cohort | 60 | 58 | Australian Rules Football | Professional | Australia | Lower Extremity | 2 | Poisson regression |
| Feijen | 2021 | Prospective Cohort | 201 | 42 | Swimming | Youth | USA | Shoulder | 1 | Logistic regression |
| Gabbet | 2010 | Prospective Cohort | 91 | 159 | Rugby | Professional | Australia | Trunk, Core, Lower Extremity | 1 | Logistic regression |
| Hughes | 2020 | Retrospective Cohort | 317 | 138 | Soccer | Professional | UK | Lower Extremity | 1 | Logistic regression |
| Ivarsson | 2014 | Prospective Cohort | 101 | 67 | Soccer | Elite Junior | Sweden | Any MSK Injury | 1 | Latent Classification |
| Jauhiainen | 2020 | Prospective Cohort | 314 | 57 | Basketball, Floorball | Youth | Finland | Ankle, knee | 2 | Logistic regression,Random Forest |
| Karuc | 2021 | Prospective Cohort | 556 | 90 | Athletic, Non-Athletic | Adolescents | Croatia | Unclear | 1 | Logistic regression,Naïve Bayes, Linear Discriminant,k-nearest Neighbors |
| Khayambashi | 2015 | Case Control | 501 | 15 | Athletes | Recreational | Iran | ACL | 1 | Logistic regression |
| Landis | 2018 | Prospective Cohort | 187 | 17 | Soccer, Volleyball, Basketball | College | USA | Lower Extremity | 1 | Logistic regression |
| Lopez-Valencia | 2018 | Prospective Cohort | 132 | 32 | Soccer, Handball | Professional | Spain | Lower Extremity | 68 | Random Forest, Boosting  |
| Luu | 2020 | Retrospective Cohort | 2322 | 6673 | Ice Hockey | Professional | USA | Any MSK Injury | 12 | Logistic regression,k-nearest neighbors,Naïve Bayes,Boosting |
| McCann | 2018 | Prospective Cohort | 43 | 8 | Soccer | College | USA | Ankle Sprain | 1 | Logistic regression |
| Oliver | 2020 | Prospective Cohort | 335 | 99 | Soccer | Youth | UK | Lower Extremity | 57 | Logistic regression,Decision Tree  |
| Pontillo | 2021 | Case Control | 278 | 16 | Athletes | College | USA | ACL | 1 | Logistic regression |
| Powers | 2018 | Prospective Cohort | 43 | 8 | Soccer | College | USA | Ankle | 1 | Logistic regression |
| Rommers  | 2020 | Prospective Cohort | 734 | 368 | Soccer | Youth | Belgium | Any MSK Injury | 9 | Gradient Boosting Machine |
| Rossi | 2018 | Prospective Cohort | 26 | 23 | Soccer | Professional | Italy | Lower Extremity | 5 | Decision Tree, Random Forest |
| Ruddy | 2018 | Prospective Cohort | 362 | 53 | Australian Rules Football | Professional | Australia | Hamstring | 1 | Logistic regression,Naïve Bayes,Random Forest,Support Vector Machine, Neural Network |
| Shambaugh | 1991 | Prospective Cohort | 45 | 15 | Basketball | Recreational | USA | Foot, Ankle, Knee, Hip | 1 | Logistic regression |
| Sturnick | 2015 | Case Control | 176 | 88 | Athletes | High School; College | USA | Knee | 1 | Logistic regression |
| Teramoto | 2017 | Retrospective Cohort | 1443 | 681 | Basketball | Professional | USA | Any MSK Injury | 1 | Poisson regression |
| Thornton | 2017 | Prospective Cohort | 25 | 156 | Rugby | Professional | Australia | Any MSK Injury | 2 | Generalized Estimated Equations,Random Forest |
| van der Does | 2015 | Prospective Cohort | 75 | 11 | Basketball, Volleyball, Korfball | Elite or Sub Elite | Netherlands | Ankle and Knee | 1 | Logistic regression |
| Whiteside | 2016 | Case Control | 208 | 104 | Baseball | Professional | USA | UCL | 3 | Logistic regression,Naïve Bayes,Support Vector Machine |
| Wiese | 2014 | Prospective Cohort | 144 | 93 | Football | College | USA | Upper or Lower Extremity | 1 | Logistic regression |
| Wilkerson | 2015 | Prospective Cohort | 152 | 132 | Football | College | USA | Core or Lower Extremity | 2 | Logistic regression, Cox regression |

ACL = Anterior Cruciate Ligament

MSK = Musculoskeletal

UCL = Ulnar Collateral Ligament

USA = United States of America

**Table 3.** Risk of Bias Using Prediction Model Risk Of Bias ASsessment Tool (PROBAST) Tool

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Study | Overall | Analyses | Outcome | Predictors | Participants |
| Abalo-Nunez | High | High | Unclear | Low | High |
| Ayala | High | High | Low | Low | Low |
| Carbuhn | High | High | Unclear | Unclear | Low |
| Colby | High | High | Low | Low | Low |
| Feijen | Low | Low | Low | Low | Low |
| Gabbet | High | High | Low | Low | Low |
| Hughes | Low | Low | Low | Low | Low |
| Ivarsson | High | High | Low | Low | Low |
| Jauhiainen | Unclear | Unclear | Low | Low | Low |
| Karuc | High | High | Low | Low | Low |
| Khayambashi | High | High | Low | Low | Low |
| Landis | High | High | Low | Low | Low |
| Lopez-Valencia | High | High | Low | Low | Low |
| Luu | Unclear | Unclear | High | Low | Low |
| McCann | High | High | Low | Low | Low |
| Oliver | High | High | Low | Low | Low |
| Pontillo | High | High | Unclear | Unclear | High |
| Powers | High | High | Low | Low | Low |
| Rommers | High | High | Low | Low | Low |
| Rossi | High | High | Low | Low | Low |
| Ruddy | Unclear | Unclear | Low | Low | Low |
| Shambaugh | Unclear | Unclear | Low | Low | Low |
| Sturnick | High | High | Unclear | Unclear | High |
| Teramoto | High | High | Low | Low | Low |
| Thornton | High | High | Low | Low | Low |
| van der Does | High | High | Unclear | Unclear | Low |
| Whiteside | High | High | Unclear | Unclear | High |
| Wiese | High | High | Low | Unclear | Low |
| Wilkerson | High | High | Low | Low | Low |

Risk of bias was assessed at the model level. Risk of bias is reported for all models at the study level unless differences in model risk of bias within specific studies were reported.

**Table 4.** Prediction Model Performance Statistics for Each Included Study

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Study | Discrimination | Accuracy | Calibration | R2 | Sensitivity | Specificity |
| Abalo-Nunez |  |  |  |  | 0.58 | 0.97 |
| Ayala | 0.84 |  |  |  |  |  |
| Carbuhn | 0.87 | 0.96 |  |  |  |  |
| Colby | 0.54-0.59 |  |  |  |  |  |
| Feijen | 0.71(0.60-0.94) |  | Calibration plot | 0.14 |  |  |
| Gabbet |  | 0.62 |  |  | 0.87(0.81-0.92) | 0.99(0.98-0.99) |
| Hughes | 0.59(0.53-0.65) |  | 0.72(0.28-1.61) | 0.06 |  |  |
| Jauhiainen | 0.65 |  |  |  |  |  |
| Karuc | Logistic regression: 0.56(0.47-0.63)Naïve Bayes: 0.58 |  |  |  | Logistic regression: 0.50Naïve Bayes: 0.70 | Logistic regression: 0.55Naïve Bayes: 0.49 |
| Khayambashi | 0.78(0.70-0.86) |  |  | 0.11 | 0.93(0.68-0.99) | 0.87(0.60-0.98) |
| Lopez-Valencia | 0.75 |  |  |  | 0.66 | 0.79 |
| Luu | Logistic regression: 0.94XGBoost: 0.95 | Logistic regression: 0.95XGBoost: 0.95 |  |  |  |  |
| McCann | 0.73(0.58-0.89) |  |  |  | 0.88 | 0.51 |
| Oliver | 0.66(0.60-0.73) |  |  |  | 0.15 | 0.97 |
| Powers | 0.67(0.56-0.78) |  |  | 0.07 | 0.48(0.28-0.69) | 0.82(0.75-0.88) |
| Rommers |  | 0.85 |  | 0.85 | 0.85 |  |
| Rossi | Decision Tree: 0.76 |  |  |  | Decision Tree: 0.87 | Decision Tree: 0.96 |
| Ruddy | 0.53 |  |  |  |  |  |
| Thornton | 0.64 |  |  |  |  |  |
| Whiteside |  | Support Vector Machine: 0.75 |  |  | Support Vector Machine: 0.74 | Support Vector Machine: 0.75 |
| Wiese | Unclear Discrimination Performance |  |  |  |  |  |
| Wilkerson |  |  | Homes & Lemeshow Test |  | 0.56 | 0.80 |

95% Confidence intervals are reported in parentheses when available

When multiple machine learning models were developed, the best performing model is reported

**References**

1. Hootman JM, Dick R, Agel J. Epidemiology of collegiate injuries for 15 sports: summary and recommendations for injury prevention initiatives. *J Athl Train* 2007;42(2):311.

2. Knowles SB, Marshall SW, Miller T, et al. Cost of injuries from a prospective cohort study of North Carolina high school athletes. *Inj Prevent* 2007;13(6):416-21.

3. Cai H, Bullock GS, Sanchez-Santos MT, et al. Joint pain and osteoarthritis in former recreational and elite cricketers. *BMC Musculoskelet Disord* 2019;20(1):596.

4. Bullock GS, Collins GS, Peirce N, et al. Playing sport injured is associated with osteoarthritis, joint pain and worse health-related quality of life: a cross-sectional study. *BMC Musculoskelet Disord* 2020;21(1):1-11.

5. Whittaker JL, Woodhouse LJ, Nettel-Aguirre A, et al. Outcomes associated with early post-traumatic osteoarthritis and other negative health consequences 3-10 years following knee joint injury in youth sport. *Osteoarthritis Cartilage* 2015;23(7):1122-9. doi: 10.1016/j.joca.2015.02.021 [published Online First: 2015/03/01]

6. Filbay S, Culvenor A, Ackerman I, et al. Quality of life in anterior cruciate ligament-deficient individuals: a systematic review and meta-analysis. *Br J Sport Med* 2015;49(16):1033-41.

7. Levy AS, Lintner S, Kenter K, et al. Intra-and interobserver reproducibility of the shoulder laxity examination. *Am J Sport Med* 1999;27(4):460-63.

8. Emery CA, Roy T-O, Whittaker JL, et al. Neuromuscular training injury prevention strategies in youth sport: a systematic review and meta-analysis. *Br J Sport Med* 2015;49(13):865-70.

9. Bullock GS, Hughes T, Sergeant JC, et al. Clinical Prediction Models in Sports Medicine: A Guide for Clinicians and Researchers. *J Orthop Sport Phys Ther* 2021;51(10):517-25.

10. Bahr R. Why screening tests to predict injury do not work—and probably never will…: a critical review. *Br J Sport Med* 2016;50(13):776-80.

11. Fischer JE, Steiner F, Zucol F, et al. Use of simple heuristics to target macrolide prescription in children with community-acquired pneumonia. *Arch Ped Adolescent Med* 2002;156(10):1005-08.

12. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults. *Chest* 1991;100(6):1619-36.

13. Bullock GS, Hughes T, Sergeant JC, et al. Methods matter: clinical prediction models will benefit sports medicine practice, but only if they are properly developed and validated. Br J Sport Med, 2021.

14. Seow D, Graham I, Massey A. Prediction models for musculoskeletal injuries in professional sporting activities: A systematic review. *Trans Sport Med* 2020;3(6):505-17. doi: 10.1002/tsm2.181

15. Claudino JG, de Oliveira Capanema D, de Souza TV, et al. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sport Med Open* 2019;5(1):1-12.

16. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart* 2012;98(9):683-90.

17. Van Calster B, Wynants L, Timmerman D, et al. Predictive analytics in health care: how can we know it works? *J Am Med Informatic Assoc* 2019;26(12):1651-54.

18. Bzdok D, Altman N, Krzywinski M. Points of significance: statistics versus machine learning. *Nature* 2018;7:1-7.

19. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016;35(2):214-26.

20. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14(1):1-13.

21. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12-22.

22. Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol* 2021;138:60-72.

23. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *J Br Surg* 2015;102(3):148-58.

24. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-98.

25. Janssen K, Moons K, Kalkman C, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61(1):76-86.

26. Bullock GS, Hughes T, Sergeant JC, et al. Improving prediction model systematic review methodology: Letter to the Editor. *Trans Sport Med 2021.*

27. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.

28. Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Int Med* 2019;170(1):51-58.

29. Abalo-Núñez R, Gutiérrez-Sánchez A, Pérez MI, et al. Injury prediction in aerobic gymnastics based on anthropometric variables. *Science & Sports* 2018;33(4):228-36.

30. Ayala F, López-Valenciano A, Martín JAG, et al. A preventive model for hamstring injuries in professional soccer: learning algorithms. *Int J Sport Med* 2019;40(05):344-53.

31. Carbuhn AF, Sanchez Z, Fry AC, et al. A simplified prediction model for lower extremity long bone stress injuries in male endurance running athletes. *Clin J Sport Med* 2020;30(5):e124-e26.

32. Colby MJ, Dawson B, Peeling P, et al. Improvement of prediction of noncontact injury in elite Australian footballers with repeated exposure to established high-risk workload scenarios. *Int J Sport Physiol Perform* 2018;13(9):1130-35.

33. Feijen S, Struyf T, Kuppens K, et al. Prediction of shoulder pain in youth competitive swimmers: the development and internal validation of a prognostic prediction model. *Am J Sport Med*2021;49(1):154-61.

34. Gabbett TJ. The development and application of an injury prediction model for noncontact, soft-tissue injuries in elite collision sport athletes. *J Strength Cond Res* 2010;24(10):2593-603.

35. Hughes T, Riley RD, Callaghan MJ, et al. The value of preseason screening for injury prediction: the development and internal validation of a multivariable prognostic model to predict indirect muscle injury risk in elite football (soccer) players. *Sport Med Open* 2020;6:1-13.

36. Ivarsson A, Johnson U, Lindwall M, et al. Psychosocial stress as a predictor of injury in elite junior soccer: a latent growth curve analysis. *J Sci Med Sport* 2014;17(4):366-70.

37. Jauhiainen S, Kauppi J-P, Leppänen M, et al. New machine learning approach for detection of injury risk factors in young team sport athletes. *Int J Sport Med* 2021;42(02):175-82.

38. Karuc J, Mišigoj-Durakovic M, Šarlija M, et al. Can injuries be predicted by functional movement screen in adolescents? The application of machine learning. *J Strength Cond Res* 2021;35(4):910-19.

39. Khayambashi K, Ghoddosi N, Straub RK, et al. Hip muscle strength predicts noncontact anterior cruciate ligament injury in male and female athletes: a prospective study. *Am J Sport Med* 2016;44(2):355-61.

40. Landis SE, Baker RT, Seegmiller JG. Non-contact anterior cruciate ligament and lower extremity injury risk prediction using functional movement screen and knee abduction moment: an epidemiological observation of female intercollegiate athletes. *Int J Sport Phys Ther* 2018;13(6):973.

41. López-Valenciano A, Ayala F, Puerta JM, et al. A preventive model for muscle injuries: a novel approach based on learning algorithms. *Med Sci Sport Exerc* 2018;50(5):915.

42. Luu BC, Wright AL, Haeberle HS, et al. Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: an analysis of 2322 players from 2007 to 2017. *Orthop J Sport Med* 2020;8(9):2325967120953404.

43. McCann RS, Kosik KB, Terada M, et al. Acute lateral ankle sprain prediction in collegiate women's soccer players. *International journal of sports physical therapy* 2018;13(1):12.

44. Oliver JL, Ayala F, Croix MBDS, 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(11):1044-48.

45. Pontillo M, Hines SM, Sennett BJ. Prediction of ACL Injuries from Vertical Jump Kinetics in Division 1 Collegiate Athletes. *Int J Sport Phys Ther* 2021;16(1):156.

46. Powers CM, Ghoddosi N, Straub RK, et al. Hip Strength as a Predictor of Ankle Sprains in Male Soccer Players: A Prospective Study. *J Athl Train* 2017;52(11):1048-55. doi: 10.4085/1062-6050-52.11.18

47. Rommers N, Rössler R, Verhagen E, et al. A machine learning approach to assess injury risk in elite youth football players. *Med Sci Sport Exerc* 2020;52(8):1745-51.

48. Rossi A, Pappalardo L, Cintia P, et al. Effective injury forecasting in soccer with GPS training data and machine learning. *PloS one* 2018;13(7):e0201264.

49. Ruddy J, Shield A, Maniar N, et al. Predictive modeling of hamstring strain injuries in elite Australian footballers. *Med Sci Sport Exerc* 2018;50(5):906-14.

50. Shambaugh JP, Klein A, Herbert JH. Structural measures as predictors of injury basketball players. *Med Sci Sport Exerc* 1991;23(5):522-27.

51. Sturnick DR, Vacek PM, DeSarno MJ, et al. Combined anatomic factors predicting risk of anterior cruciate ligament injury for males and females. *Am J Sport Med* 2015;43(4):839-47.

52. Teramoto M, Cross CL, Cushman DM, et al. Game injuries in relation to game schedules in the National Basketball Association. *J Sci Med Sport* 2017;20(3):230-35.

53. Thornton HR, Delaney JA, Duthie GM, et al. Importance of various training-load measures in injury incidence of professional rugby league athletes. *Int J Sport Physiol Perform* 2017;12(6):819-24.

54. Van Der Does H, Brink M, Benjaminse A, et al. Jump landing characteristics predict lower extremity injuries in indoor team sports. *Int J Sport Med* 2016;37(03):251-56.

55. Whiteside D, Martini DN, Lepley AS, et al. Predictors of ulnar collateral ligament reconstruction in Major League Baseball pitchers. *Am J Sport Med* 2016;44(9):2202-09.

56. Wiese BW, Boone JK, Mattacola CG, et al. Determination of the functional movement screen to predict musculoskeletal injury in intercollegiate athletics. *Athl Train Sport Health Care* 2014;6(4):161-69.

57. Wilkerson GB, Colston MA. A refined prediction model for core and lower extremity sprains and strains among collegiate football players. *J Athl Train* 2015;50(6):643-50.

58. Riley RD, Ensor J, Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. *Bmj* 2020;368

59. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *Bmj* 2020;369

60. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016;16(1):163. doi: 10.1186/s12874-016-0267-3

61. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I–Continuous outcomes. *Stat Med* 2019;38(7):1262-75.

62. Steyerberg EW. Clinical prediction models: Springer 2019.

63. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25(1):127-41.

64. Gauthier J, Wu Q, Gooley T. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Nature 2020:675-670.

65. Carey DL, Crossley KM, Whiteley R, et al. Modeling Training Loads and Injuries: The Dangers of Discretization. *Med Sci Sport Exerc* 2018;50(11):2267-76.

66. Collins GS, Ma J, Gerry S, et al. Risk prediction models in perioperative medicine: methodological considerations. *Curr Anesthesiol Reports* 2016;6(3):267-75.

67. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stati Society* 1996;58(1):267-88.

68. Riley RD, Snell KI, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol* 2021;132:88-96.

69. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453-73.

70. Steyerberg EW, Harrell Jr FE, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54(8):774-81.

71. Debray TP, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68(3):279-89.

72. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiol* 2010;21(1):128.

73. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17(1):1-7.

74. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14(1):40.

75. Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9(1):1-14.

76. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artificial Intelligence Res* 2002;16:321-57.

77. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Med* 2020;26(9):1320-24.

78. Collins GS, Dhiman P, Navarro CLA, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11(7):e048008.

79. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation* 2015;131(2):211-19.