Received: 16 February 2021 | Accepted: 17 February 2021

DOI: 10.1002/tsm2.240

LETTER TO THE EDITOR

WILEY

Improving prediction model systematic review methodology: Letter to the Editor

Dear Editor,

In their recently published paper, Seow et al[1] carried out a systematic review of musculoskeletal injury prediction models in professional sport and military special forces. Their review encompassed a comprehensive search that included both conference and published papers, used a standardized musculoskeletal injury definition that was informed by the literature, and included both statistical and machine learning-based models. Nevertheless, we have a number of concerns regarding the conduct and reporting of some aspects of the study that limit the usefulness of their findings.

Our first point relates to how the studies were appraised. While the authors should be commended on assessing each study for risk of bias, the Newcastle Ottawa Scale (NOS) is not the correct tool to do this. The NOS is a generic tool designed to assess the quality of non-randomized studies such as case-control and cohort studies—and while prediction model studies often use cohort design, the tool includes no specific assessment of analysis issues relating to the development or validation of a prediction model. Hence, the NOS is a blunt instrument to assess risk of bias in these studies. The tool that should have been used to assess the risk of bias in the review by Seow et al[1] is the Prediction model Risk Of Bias Assessment Tool (PROBAST),[2] which includes 20 signaling questions over four domains (participants, predictors, outcome, and analysis), to cover key aspects of prediction model studies. Furthermore, when designing a systematic review of prediction model studies, the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist[3] provides detailed guidance to help authors in developing their systematic review questions relating to prediction models, extracting pertinent prediction model data, and appraising prediction model studies.[3] Had these more relevant tools been used, and indeed, the review process outlined by the Cochrane Prognosis Methods Group followed[4]; it would have enabled the authors to better appraise and utilize the included prediction model studies in their review. In particular, it would have given more depth and clarity, and allowed enhanced identification of any strength in the existing evidence and also highlighted particular areas of conduct and reporting that should be improved upon in future studies.

While the authors extracted and reported the discrimination performance (such as area under the curve) of models that were included, we note that there was no comment on model calibration—an essential component of model performance.[4,5] Calibration is the agreement between probabilities derived from the model versus those actually observed within the data[6] and is important in understanding the accuracy of the predictions from the model.[7,8] This omission could have been addressed at the design stage using the aforementioned CHARMS checklist. Consequently, the authors have missed an important opportunity to report on this critical aspect of prediction model performance assessment and therefore presented readers with incomplete information on the usefulness of the included prediction models. Furthermore, any omission of calibration in the primary studies will have a direct and negative impact on the risk of bias assessment. A related concern is that the authors do not explain how they extracted performance estimates, and whether they used the extensive tools of Debray et al[9] to help extract estimates (eg, the area under the curve and its confidence interval) when these were not reported directly, in order to maximize the information available for review. Whether performance statistics were adjusted for optimism was also not reported,[10] and clinical utility measures (eg, net benefit[11]) were not discussed.

We were also concerned with the authors' expectations regarding the handling class imbalance using over- or under-sampling to create a more balanced data set. Data are said to be imbalanced when there are fewer individuals in the data set with the outcome (compared to those without the outcome). In the context of classification, this can indeed be a problem, for example, when evaluating classification accuracy (ie, proportion of correct classifications) in the sense that incorrectly misclassifying individuals with the outcome in a highly imbalanced data set could yield high accuracy—as the larger non-outcome group will dominate the calculation of overall accuracy.[12] However, in the context of prediction (the aim of the review by Seow et al[1]), class imbalance is a feature of the

data to be accounted for in the modeling. For example, models developed using logistic regression, and the class imbalance is handled by the intercept, balancing the data (ie, changing the outcome prevalence) by over or under sampling will affect estimation of the model intercept and thus lead to inaccurate model predictions. Artificially modifying the outcome prevalence by using such sampling approaches would inherently prohibit a prediction model from properly quantifying injury risk, thus hindering calibration and the potential for application. This could affect the health outcomes of athletes and have potentially serious implications for clinical practice. As such, there is no need to account for class imbalance.

Due to the inherent limitations of a letter to the editor, we are restricted from presenting details of further methodological and reporting concerns. Briefly, these include the absence of the extraction or reporting on: (a) missing data and relevant missing data mechanisms; (b) the handling of continuous predictors and whether non-linearity was considered; (c) a priori sample size calculations, (d) the use of number of events (rather overall sample size) in driving model sample size, (e) whether developed models were adjusted for overfitting using penalization methods, and (f) external validation processes to increase generalizability and clinical implementation of these models.[13-17]

While improved data extraction and risk of bias evaluation of primary prediction model studies would enhance the contribution of this systematic review, we also feel that it is important to highlight such issues in order to advance the understanding, development, and critical appraisal of prediction models in sport more widely. We hope that by increasing awareness of strategies for improving such methods, we can help improve prediction model performance in professional sport and athletic populations, ultimately aiding athlete health and career longevity.

Garrett S. Bullock[1]
Tom Hughes[2]
Jamie C. Sergeant[3]
Michael J. Callaghan[2,3,5]
Gary S. Collins[6]
Richard D. Riley[4]

[1]*Nuffield Department of Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[2]*Manchester United Football Club, Manchester, UK*
[3]*Centre for Biostatistics, University of Manchester, Manchester, UK*
[4]*Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK*
[5]*Department of Health Professions, Manchester Metropolitan University, Manchester, UK*
[6]*Centre for Statistics in Medicine, University of Oxford, Oxford, UK*

Correspondence
Garrett S. Bullock, Nuffield Department of Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
Email: garrett.bullock@wolfson.ox.ac.uk

ORCID

Garrett S. Bullock https://orcid.org/0000-0003-0236-9015
Gary S. Collins https://orcid.org/0000-0002-2772-2316

REFERENCES

1. Seow D, Graham I, Massey A. Prediction models for musculoskeletal injuries in professional sporting activities: a systematic review. *Transl Sports Med*. 2020;3(6):505-517.
2. Wolff RF, Moons KGM, 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.
3. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Medicine*. 2014;11(10):e1001744.
4. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.
5. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1-7.
6. 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):128-138.
7. Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer; 2019.
8. Riley RD, van der Windt D, Croft P, Moons KG. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford, UK: Oxford University Press; 2019.
9. Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res*. 2019;28(9):2768-2786.
10. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Int Med*. 2015;162(1):W1-W73.
11. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artificial Intelligence Res*. 2002;16:321-357.
13. 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 Method*. 2014;14(1):40.
14. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
15. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.

16. Riley RD, Snell KIE, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol*. 2020;132:88-96.
17. van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474.