





Integrative data analysis to identify persistent post-concussion deficits and subsequent musculoskeletal injury risk: project structure and methods

Melissa Anderson ¹, Claudio Cesar Claros,² Wei Qian ³,
Austin Brockmeier ^{2,4}, Thomas A Buckley ⁵

To cite: Anderson M, Claros CC, Qian W, *et al*. Integrative data analysis to identify persistent post-concussion deficits and subsequent musculoskeletal injury risk: project structure and methods. *BMJ Open Sport & Exercise Medicine* 2024;**10**:e001859. doi:10.1136/bmjsem-2023-001859

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjsem-2023-001859>).

Accepted 9 January 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹School of Health Sciences and Professions, Ohio University, Athens, Ohio, USA

²Department of Electrical and Computer Engineering, University of Delaware, Newark, Delaware, USA

³Department of Applied Economics and Statistics, University of Delaware, Newark, Delaware, USA

⁴Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, USA

⁵Department of Kinesiology & Applied Physiology, University of Delaware, Newark, Delaware, USA

Correspondence to
Dr Thomas A Buckley;
TBuckley@udel.edu

ABSTRACT

Concussions are a serious public health problem, with significant healthcare costs and risks. One of the most serious complications of concussions is an increased risk of subsequent musculoskeletal injuries (MSKI). However, there is currently no reliable way to identify which individuals are at highest risk for post-concussion MSKIs. This study proposes a novel data analysis strategy for developing a clinically feasible risk score for post-concussion MSKIs in student-athletes. The data set consists of one-time tests (eg, mental health questionnaires), relevant information on demographics, health history (including details regarding the concussion such as day of the year and time lost) and athletic participation (current sport and contact level) that were collected at a single time point as well as multiple time points (baseline and follow-up time points after the concussion) of the clinical assessments (ie, cognitive, postural stability, reaction time and vestibular and ocular motor testing). The follow-up time point measurements were treated as individual variables and as differences from the baseline. Our approach used a weight-of-evidence (WoE) transformation to handle missing data and variable heterogeneity and machine learning methods for variable selection and model fitting. We applied a training-testing sample splitting scheme and performed variable preprocessing with the WoE transformation. Then, machine learning methods were applied to predict the MSKI indicator prediction, thereby constructing a composite risk score for the training-testing sample. This methodology demonstrates the potential of using machine learning methods to improve the accuracy and interpretability of risk scores for MSKI.

INTRODUCTION

Concussions have been identified by both the US National Institutes of Health (NIH) and the US Centers for Disease Control and Prevention (CDC) as a serious public health problem, with an annual incidence of up to 3.8 million and associated costs of approximately US\$22 billion.¹ Healthcare professionals that manage concussions are guided by consensus and position statements

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Concussions are associated with an increased risk of musculoskeletal injuries (MSKIs), but there is currently no effective way to identify which individuals are at the highest risk.

WHAT THIS STUDY ADDS

⇒ This study provides preliminary evidence that the proposed risk score could be a valuable tool for clinicians to identify student-athletes at high risk for post-concussion MSKIs.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study provides a novel and effective data analysis strategy for developing risk scores for post-concussion MSKIs.

that make recommendations for a multifaceted approach to clinical concussion care. However, these clinical assessments may lack sensitivity to identify recovery as deficits in numerous sophisticated assessments (eg, neuroimaging, blood-based biomarkers and other instrumented measures) persist beyond clinical recovery, suggesting athletes may return to participation (RTP) before complete neurological recovery.² This premature RTP may result in the ~2× elevated musculoskeletal injury (MSKI) risk in the year following a concussion, which has been identified across diverse sports settings, ages and sexes.³ These MSKIs carry enormous societal and economic consequences affecting up to 12 million people annually, leading to 20 million lost school days, lost sports time and costing ~US\$33 billion annually in healthcare costs.⁴ Further, an MSKI also increases the risk of chronic physical complications, leading to reduced physical activity and may increase the risk of chronic health conditions such as diabetes and cardiovascular disease.⁵ Thus, there is a need to identify those at the



greatest risk for subsequent MSKI to implement injury risk reduction techniques.

Multiple attempts have been made to identify those athletes at the highest risk for post-concussion MSKI but with limited success. Impairments in dual-task gait have been found in high school and collegiate athletes who experienced post-concussion MSKI, but these were weak associations without prognostic capabilities.⁶ Multifaceted clinical examinations, widely used by healthcare providers, are cost-effective and clinically feasible tests⁷; however, individual assessments (eg, symptoms, cognitive testing) were not effective in identifying elevated MSKI risk.⁸ Similarly, clinical mental health measures (eg, Brief Symptom Inventory (BSI-18), Hospital Anxiety and Depression Scale (HADS)) were also not predictive of subsequent injury, although satisfaction with life had a limited association.⁹ Others have posited broader risk factors such as persistent neurocognitive deficits¹⁰; however, while plausible, both lack empirical evidence. Independent of concussion, injury prediction is notoriously difficult, and standard interventional studies have largely been unsuccessful,¹¹ with the notable exception of ACL screening protocols.¹² Thus, developing a post-concussion MSKI prediction model requires innovative approaches.

An ideal risk-scoring model of post-concussion subsequent MSKI would determine a minimal set of predictive clinically feasible variables (eg, demographics, health history, concussion characteristics and recovery) that can identify individuals at high risk for a subsequent MSKI. Thus, an integrative statistical model is needed to combine these disparate test measurements with demographic information and health history to create a composite risk score model for subsequent MSKI. Model fitting is further complicated by missing data, prevalent in sports medicine research and clinical practice, often due to time constraints during assessments and patient non-compliance, which may lead to biased or incomplete risk inferences and ineffective interventions.¹³ To overcome this systemic issue, suitable statistical methodologies such as data imputation techniques are crucial for generating reliable risk models while considering missing data patterns.¹⁴

Herein, we propose to generate a composite risk score based on clinically feasible information for post-concussion MSKI risk through a two-step process. First, we propose a weight-of-evidence (WoE) transformation,^{15–17} which naturally handles missing data and heterogeneous variables by replacing the values with univariate risk scores. Second, we propose using a variable selection algorithm and logistic regression to form the multivariate composite risk score (the details will be described in the subsequent methodology section). Our approach overcomes the challenges stemming from numerous irrelevant covariates and prevalent missing values. Overall, this general and versatile data analysis and strategy is a step towards addressing the pressing need to understand post-concussion recovery and MSKI risk.

Research aims and approach

With the proposed novel analysis strategies, this study aims to identify post-concussion MSKI risk categories. We also aim to develop a clinical risk score similar to Zemek's prediction of persistent concussion symptoms approach¹⁸ for post-concussion MSKI. These clinically feasible approaches could allow clinicians to apply targeted interventions with known injury risk reduction approaches if successful.

IMPLEMENTED METHODOLOGY

We have developed an extensive longitudinal concussion data set (2015–2022), which includes data on 211 student-athlete concussions, including demographic information, medical history, concussion injury and recovery information, and common data elements (CDEs) across clinical milestones. Data collected between 2015 and 2021 were part of the Concussion Assessment, Research and Education (CARE) Consortium.¹⁹ All data collection occurred at the University of Delaware, which is in National Collegiate Athletic Association's (NCAA) Division I and the Mid-Atlantic region of the USA.

The time to complete all tests was 50–60 min at baseline and 30–40 min at three follow-up time points following concussion: (1) Acute (<48 hours post-concussion), (2) Asymptomatic (when no concussion symptoms are reported) and (3) Return to Play (RTP) (when the student-athlete returns to full participation without restriction). Data were extracted and compiled by the research team starting in February 2022 and were updated through January 2023 as new concussions occurred. Further, MSKI were updated until March 2023.

Patient and public involvement

Former NCAA athletes provided site-level feedback regarding study procedures, which was incorporated into the CARE Consortium study design.

Clinical assessments

The selected CDEs were collected following standard procedures established in the literature.¹⁹ Relevant confounding variables (eg, age, sex, injury mechanism and presentation, prior concussion and MSKI history) were collected as described below. All participants provided written and oral informed consent, and some participants consented to only a subset of access, as approved by the University of Delaware institutional review board (IRB). Each assessment has been thoroughly described. (online supplemental table 2). Briefly, neurocognitive functioning was evaluated through the computerised test Immediate Post Concussion Assessment Tool¹⁹ with composite scores representing verbal memory, visual memory, motor speed and reaction time. The Standardised Assessment of Concussion assesses mental status,²⁰ and the Balance Error Scoring System evaluates postural stability.²¹ We used two measures of symptom reporting, the Sport Concussion Assessment Tool 5 (SCAT5) symptom list,¹⁹ which lists 22 common

**Table 1** Categorisation schema of musculoskeletal injuries

Region	Side	Severity
1. Head	R. Right	A. Acute/non-surgical:
2. Neck	L. Left	time lost from sport
3. Shoulder	B. Bilateral	<21 days
4. Upper arm		B. Chronic/non-surgical:
5. Forearm/ elbow		time lost ≥21 days
6. Wrist		C. Chronic/surgical: re-
7. Hand		quired surgery to re-
8. Torso		pair, lost ≥30 days
9. Pelvis		
10. Thigh		
11. Knee		
12. Shank		
13. Ankle		
14. Foot		

concussion symptoms weighted from 0 (symptom not present) to 6 (severe symptom). We used the total number of symptoms and symptom severity from the SCAT5. The BSI-18¹⁹ is a self-report questionnaire that evaluates psychological distress and psychological disorders like depression, anxiety and somatisation. The King-Devick test was used to evaluate saccadic eye movements,¹⁹ and the Vestibular Ocular Motor Screen¹⁹ was used to evaluate vestibular and oculomotor function and symptoms. Tandem gait was used to evaluate gait and balance control under single and dual-task conditions, which involves performing a secondary task while walking.²² Lastly, both the reliable and valid Satisfaction with Life Scale and the HADS evaluated participants' quality of life.¹⁹

Electronic medical records

The participants' MSKI history was obtained by accessing the University of Delaware SportsWareOnline (Computer Sports Medicine, Stoughton, Massachusetts, USA) electronic health record through IRB-approved approaches and with the participant's informed consent. The MSKI was categorised by region, side, severity and time loss (table 1).²³ Time from each injury in relation to a concussion was calculated in days, with a negative value indicating the MSKI occurred before the concussion and a positive value indicating that the MSKI occurred after the concussive injury. Finally, the total number of unique MSKI was calculated for each participant (range=0–13 injuries). For this study, we only examined MSKI that occurred after a concussion.

Challenges and justification for data analysis strategy

The preliminary analysis efforts are met with four issues that make effective MSKI risk modelling challenging. First, incomplete and missing data is a substantial challenge as prospectively assessing intercollegiate athletes in-season has inherent limitations; however, simply ignoring the missing data or using imputation may result in biased estimation and inaccurate inference.²⁴ Second, our initial

data exploration revealed non-linear and non-monotone relationships between the covariates and the MSKI risk, making it difficult to justify using a linear model such as logistic regression that assumes a monotone variable association with the risk. Third, our electronic health records contain a set of variables that are measured at four time points (baseline, acute, asymptomatic, RTP), which may hold strong potential for building clinically informative risk scores; however, they also pose a technical challenge to identify a (reasonably modest-sized and explainable) set of important variables from all possible pairs of time point and measurement. There are also difficulties in comparing the relative importance of categorical and continuous variables for interpretation purposes. Finally, dimensionality increases when we attempt to categorise continuous variables or encode categorical variables (see online supplemental table 1 for the complete list of categorical and continuous variables in our study).

In response to these challenges for effective MSKI risk modelling, we propose to perform a variable transformation method called WoE, which deals with missing data and variable heterogeneity by replacing variable values with their estimated univariate risk-related scores.^{15–17}

These scores operate on the same scale, so it simplifies comparison across diverse data types. Additionally, it helps resolve the possible non-linear relationships between differing values and risk and avoids the need to increase the number of variables excessively. Subsequent modelling using variable selection methods such as Recursive Feature Elimination²⁵ and Least Absolute Shrinkage and Selection Operator (LASSO)^{26 27} are then applied to the transformed variables to select a minimal set of transformed patient variables for logistic regression analysis, which combines variables into a composite score that quantifies the risk for subsequent MSKI.

MSKI data analysis

The data set consists of one-time tests (eg, mental health questionnaires), relevant information on demographics, health history (including details regarding the concussion, such as day of the year and time lost) and athletic participation (current sport and contact level) that are collected at a single time point as well as multiple time points of the clinical assessments (baseline and follow-up time points after the concussion). The follow-up time point measurements are treated as individual variables and as differences from the baseline.

The statistical analysis can be described in four steps following the discussion of the MSKI study challenges and our proposed general methodology above.

Step 1. Apply a training-testing sample splitting scheme with a 2-to-1 ratio of their sample sizes.

Step 2. Perform variable preprocessing with the WoE transformation. Technically, the WoE transformation is directly applicable to a discrete-valued random variable $X \in \{x_1, x_2, \dots, x_D\}$ and is a function $WoE(x) \mapsto \log \frac{\Pr(X=x|Y=1)}{\Pr(X=x|Y=0)}$ that replaces the discrete value



of a variable with the ratio of the logarithms of the conditional probability of the variable value given an MSKI, $p_{X|1}(x) = \Pr(X = x|Y = 1)$, to the conditional probability of the variable given no MSKI, $p_{X|0}(x) = \Pr(X = x|Y = 0)$. As a difference of the log-probabilities, the WoE value is large and positive if the variable value occurs more frequently with an MSKI than without an MSKI; conversely, if WoE is large and negative, then the variable value is more frequently given no MSKI than MSKI. In practice, the true probabilities are replaced by empirical estimates. Consequently, with limited data, too many discrete values lead to poor estimates for values with few occurrences.

To summarise the WoE of a variable across all variable values, the information value (IV) is computed, $IV(X) = \sum_{d=1}^D (p_{X|1}(x_d) - p_{X|0}(x_d)) \cdot WoE(x_d)$, which is also known as Jeffrey's divergence between the conditional distribution functions of the variable given the MSKI outcome (any injury regardless of severity or time loss).^{28 29} Larger divergence values between the variable's conditional distributions correspond to more informative variables (higher IV values).

For a continuous variable $\tilde{X} \in R$, defining the conditional distribution and the computation of the WoE transform requires the discretisation/binning of the variable values into discrete ranges. This discretisation is achieved by searching for the optimal binning (number of bins and bin edges) that maximises the IV. Ordinal variables can be grouped in the same manner. For categorical variables, maximising the IV can also consist of grouping different categories. Once the optimal binning is determined, the WoE transformation is applied to the discretised variable described above.

Step 3. Apply machine learning methods to fit a combination of the WoE-transformed measures to predict the MSKI indicator prediction, thereby constructing a composite risk score. Specifically, the well-established high-dimensional regression methods, including the LASSO²⁷ and the high-dimensional sufficient dimension reduction,²⁶ can be applied for both variable selection and constructing the optimal linear combinations of the WoE measures. After modelling fitting, the linear coefficients can be examined to quantify the contribution of different variables.

Step 4. Apply the predictive model from Step 3 to the testing data set and evaluate the results' specificity and sensitivity compared with logistic regression with all original variables.

CONCLUSIONS

We present a novel and effective data analysis strategy for developing risk scores for post-concussion MSKIs. By replacing each variable with its estimated univariate risk score, we address the challenges of MSKI risk modeling, including missing data and variable heterogeneity. Our method also simplifies comparison across diverse data types and identifies and accounts for non-linear

relationships between different variables without adding too many variables to the model. If successful in a larger data set, these clinically feasible approaches could help clinicians develop and implement targeted interventions that reduce the risk of post-concussion MSKIs.

Twitter Melissa Anderson @MelissaAndrsn and Thomas A Buckley @ConcussionUD

Contributors Conceptualisation and methodology (TAB, AB, WQ), supervision (TAB, AB, WQ), data curation (MA, CCC), writing—original draft (MA), writing—review and editing (MA, CCC, AB, WQ, TAB).

Funding National Institute of Health: National Institute of Neurological Disorders and Stroke (NINDS). TAB (PI), AB, WQ. Integrative Data Analysis to Identify Persistent Post-Concussion Deficits and Subsequent Musculoskeletal Injury Risk. Award: 1R21NS122033-01A1. 3 September 2021 to 31 July 2023

Competing interests All authors have read and understood BMJ policies on declaration of interests. MA, AB, WQ and CCC declare that they have no competing interests. TAB has a research contract with StateSpace.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by University of Delaware IRB, (804454-14) Multifaceted Concussion Assessment and Management Protocol. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; internally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available upon reasonable request. Data related to baseline and post-injury concussion assessments are available on FITBUR. Data specific to subsequent injuries are maintained by the University of Delaware and can be made available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Melissa Anderson <http://orcid.org/0000-0001-7626-6997>

Wei Qian <http://orcid.org/0000-0003-1022-1141>

Austin Brockmeier <http://orcid.org/0000-0002-7293-8140>

Thomas A Buckley <http://orcid.org/0000-0002-0515-0150>

REFERENCES

- Langlois JA, Rutland-Brown W, Wald MM. The epidemiology and impact of traumatic brain injury: a brief overview. *J Head Trauma Rehabil* 2006;21:375–8.
- Patricios JS, Schneider KJ, Dvorak J, et al. Consensus statement on concussion in sport: the 6th international conference on concussion in sport—Amsterdam, October 2022. *Br J Sports Med* 2023;57:695–711.
- McPherson AL, Nagai T, Webster KE, et al. Musculoskeletal injury risk after sport-related concussion: a systematic review and meta-analysis. *Am J Sports Med* 2019;47:1754–62.
- Swenson DM, Collins CL, Best TM, et al. Epidemiology of knee injuries among us high school athletes, 2005/06–2010/11. *Med Sci Sports Exerc* 2013;45:462–9.



- 5 Lynall RC, Pietrosimone B, Kerr ZY, *et al.* Osteoarthritis prevalence in retired national football league players with a history of concussion and lower extremity injury. *J Athl Train* 2017;52:518–25.
- 6 Oldham JR, Howell DR, Knight CA, *et al.* Gait performance is associated with subsequent lower extremity injury following concussion. *Med Sci Sports Exerc* 2020;52:2279–85.
- 7 Echemendia RJ, Burma JS, Bruce JM, *et al.* Acute evaluation of sport-related concussion and implications for the sport concussion assessment tool (SCAT6) for adults, adolescents and children: a systematic review. *Br J Sports Med* 2023;57:722–35.
- 8 Lynall RC, Mauntel TC, Pohlig RT, *et al.* Lower extremity musculoskeletal injury risk after concussion recovery in high school athletes. *J Athl Train* 2017;52:1028–34.
- 9 Buckley TA, Bryk KN, Enrique AL, *et al.* Clinical mental health measures and prediction of postconcussion musculoskeletal injury. *J Athl Train* 2023;58:401–7.
- 10 C Herman D, Zaremski JL, Vincent HK, *et al.* Effect of neurocognition and concussion on musculoskeletal injury risk. *Curr Sports Med Rep* 2015;14:194–9.
- 11 Bahr R. Why screening tests to predict injury do not work—and probably never will...: a critical review. *Br J Sports Med* 2016;50:776–80.
- 12 Hewett TE, Myer GD, Ford KR, *et al.* Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. *Am J Sports Med* 2005;33:492–501.
- 13 Khanna K. Missing medical information adversely affects care of patients. *BMJ* 2005;330:276.
- 14 van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007;16:219–42.
- 15 Breiman L, Friedman D, Olshen R, *et al.* *CART: Classification and Regression Trees*. Wadsworth Int. Group, 1984.
- 16 GoodlJ. Probability and the weighing of evidence [Internet]. 1950. Available: <https://philpapers.org/rec/GOOPAT-3>
- 17 Wod. Weight of evidence: a brief survey. *Bayesian Stat* 1985;2:249–70.
- 18 Zemek R, Osmond MH, Barrowman N, *et al.* Predicting and preventing postconcussive problems in paediatrics (5p) study: protocol for a prospective multicentre clinical prediction rule derivation study in children with concussion. *BMJ Open* 2013;3:e003550.
- 19 Broglio SP, McCrea M, McAllister T, *et al.* A national study on the effects of concussion in collegiate athletes and US military service academy members: the NCAA–DoD concussion assessment. *Sports Med* 2017;47:1437–51.
- 20 McCrea M, Kelly JP, Randolph C, *et al.* Standardized assessment of concussion (SAC): on-site mental status evaluation of the athlete. *J Head Trauma Rehabil* 1998;13:27–35.
- 21 Finnoff JT, Peterson VJ, Hollman JH, *et al.* Intrarater and Interrater reliability of the balance error scoring system (BESS). *PM&R* 2009;1:50–4.
- 22 Oldham JR, DiFabio MS, Kaminski TW, *et al.* Normative tandem gait in collegiate student-athletes: implications for clinical concussion assessment. *Sports Health* 2017;9:305–11.
- 23 Knight KL. More precise classification of orthopaedic injury types and treatment will improve patient care. *J Athl Train* 2008;43:117–8.
- 24 Little RJA, Rubin DB. *Statistical analysis with missing data*; 2002.
- 25 Guo A. Gene selection for cancer classification using support vector machines; 2002.
- 26 Qian W, Ding S, Cook RD. Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *J Am Stat Assoc* 2019;114:1277–90.
- 27 Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc: Series B (Methodol)* 1996;58:267–88. 10.1111/j.2517-6161.1996.tb02080.x Available: <https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x>
- 28 Siddiqi N. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, 2012.
- 29 Refaat M. Credit risk scorecard: development and implementation using SAS. Lulu.com; 2011. Available: <https://www.lulu.com/shop/mamdouh-refaat/credit-risk-scorecards-development-and-implementation-using-sas/hardcover/product-1qzweer5.html?page=1&pageSize=4>