ORIGINAL RESEARCH ARTICLE

Predicting Risk of Sport-Related Concussion in Collegiate Athletes and Military Cadets: A Machine Learning Approach Using Baseline Data from the CARE Consortium Study

Joel Castellanos[1,2] · Cheng Perng Phoo[3,4] · James T. Eckner[1] · Lea Franco[1] · Steven P. Broglio[5] · Mike McCrea[6] · Thomas McAllister[7] · Jenna Wiens[3] · The CARE Consortium Investigators

Abstract

Objective To develop a predictive model for sport-related concussion in collegiate athletes and military service academy cadets using baseline data collecting during the pre-participation examination.

Methods Baseline assessments were performed in 15,682 participants from 21 US academic institutions and military service academies participating in the CARE Consortium Study during the 2015–2016 academic year. Participants were monitored for sport-related concussion during the subsequent season. 176 baseline covariates mapped to 957 binary features were used as input into a support vector machine model with the goal of learning to stratify participants according to their risk for sport-related concussion. Performance was evaluated in terms of area under the receiver operating characteristic curve (AUROC) on a held-out test set. Model inputs significantly associated with either increased or decreased risk were identified.

Results 595 participants (3.79%) sustained a concussion during the study period. The predictive model achieved an AUROC of 0.73 (95% confidence interval 0.70–0.76), with variable performance across sports. Features with significant positive and negative associations with subsequent sport-related concussion were identified.

Conclusion(s) This predictive model using only baseline data identified athletes and cadets who would go on to sustain sport-related concussion with comparable accuracy to many existing concussion assessment tools for identifying concussion. Furthermore, this study provides insight into potential concussion risk and protective factors.

Joel Castellanos and Cheng Perng Phoo are co-first authors.

The members of The CARE Consortium Investigators are mentioned in "Acknowledgements" section.

This article is part of a collection on The NCAA-DoD Concussion Assessment, Research and Education (CARE) Consortium.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40279-020-01390-w) contains supplementary material, which is available to authorized users.

✉ James T. Eckner
jeckner@med.umich.edu

1   Department of Physical Medicine and Rehabilitation, Michigan Medicine, University of Michigan, 325 E. Eisenhower Parkway, Ann Arbor, MI 48108, USA

2   Present Address: Anestheshiology, School of Medicine, University of California San Diego, San Diego, CA, USA

3   Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA

4   Present Address: Computer Science, Cornell University, New York, USA

5   Kinesiology, University of Michigan, Ann Arbor, MI, USA

6   Neurosurgery, Medical College of Wisconsin, Milwaukee, WI, USA

7   Psychiatry, Indiana University, Indianapolis, IN, USA

Key Points

Using a data-driven approach, we used baseline data representing 18 variable categories to build a predictive model for sport-related concussion in collegiate athletes and military cadets with AUROC = 0.73.

Significant features in the predictive model can provide insight into risk and protective factors for sport-related concussion and be used to generate hypotheses for future research.

This is clinically important because a predictive model capable of identifying athletes at elevated risk for sustaining sport-related concussion can facilitate more targeted injury prevention, education, and surveillance strategies.

1 Introduction

Concussion, or mild traumatic brain injury (mTBI), is a common and serious injury faced by athletes and military personnel alike. The US Centers for Disease Control and Prevention estimates as many as 1.6–3.8 million sport and recreation-related concussions occur annually in the US across all age groups and levels of play [1]. Of these, approximately 10,500 concussions occur annually in the approximately 495,000 collegiate student-athletes competing in NCAA championship sports [2]. TBI is also a serious concern for US military personnel, and has been identified by the US Department of Defense as "one of the signature injuries of troops wounded in Afghanistan and Iraq" [3]. Approximately 82.3% of traumatic brain injuries sustained by Active-Duty US Service Members are mTBI/concussions, often occurring outside of combat with a mechanism similar to that in athletes [4]. In the US military service academies, non-varsity-athlete cadets frequently sustain concussions during club/intramural sport participation as well as military training.

In the short term, concussion causes a constellation of physical, cognitive, affective, and sleep-related signs and symptoms that may limit an individual's ability to participate and perform in school, at work, on the field of play, or on the battlefield. Additionally, there is growing concern that concussions and repetitive sport-associated head trauma may increase an individual's long-term risk for developing depression [5], dementia [6], and neurodegenerative diseases such as chronic traumatic encephalopathy later in life [7–10]. While concussions typically occur in the general population as a result of unpredictable traumatic events, athletes and military cadets are placed at risk for concussion during participation in sport activities. Given the myriad of potential short- and long-term consequences of concussion, efforts should be put forth to identify those at greatest risk for sport-related concussion so that targeted injury surveillance, prevention, and education strategies can be optimized.

We are aware of only one small study attempting to use a limited number of baseline factors to identify athletes' future concussion risk, without success [11]. Broadly speaking, injury prediction models can utilize factors which are either positively or negatively associated with an outcome of interest. In the case of sport-related concussion, other prior work has identified numerous intrinsic (i.e., individual athlete characteristics), and extrinsic (i.e., sport-associated conditions and mechanisms) features as potential risk or protective factors for injury [12–16]. However, many existing studies have limited their study population to a single sex [17, 18], or sport [17–25], and those studies assessing large, diverse populations have still considered only a relatively small set of specific risk factors [26–29]. As such, many potential concussion risk and protective factors remain unexplored.

We, therefore, sought to leverage the Concussion Assessment Research and Education (CARE) Consortium database [30] to develop a predictive model for sport-related concussion using baseline characteristics in NCAA collegiate athletes and US military cadets. We hypothesized the information commonly collected in athletes and cadets during their baseline pre-participation examinations could be used to predict subsequent concussions in this cohort. Using machine learning techniques, we developed an interpretable risk stratification model based on a large number of baseline covariates. By inspecting the model, we also sought insight into potential risk and protective factors for sport-related concussion which can be utilized to generate novel hypotheses for future study.

2 Methods

2.1 General Study Design

This observational study utilized CARE Consortium data collected at the 21 US academic institutions and military academies participating in the study during the 2015–2016 academic year. Participating student-athletes and military service academy cadets at each institution completed a preseason baseline assessment as part of the pre-participation examination process and were then prospectively monitored for concussion. Baseline assessments included a combination of "Level A" assessment measures common across all CARE Consortium Sites (demographics; medical, sport, academic, and family history; Brief Sensation Seeking

Scale [BSSS]; Sport Concussion Assessment Tool [SCAT] Symptom Checklist; Brief Symptom Inventory [BSI]; Standardized Assessment of Concussion [SAC]; Balance Error Scoring System [BESS]; and computerized neurocognitive assessment) as well as optional "Level B" assessment measures collected at each site's discretion (clinical reaction time, advanced measures of postural stability, oculomotor/oculovestibular assessments, and/or quality of life). One of four computerized concussion tests (ImPACT, Axon/CogState, CNS Vital Signs, or Automated Neuropsychological Assessment Metrics [ANAM]) was administered at each CARE Consortium site. Additional CARE Consortium Study details, including a description of Level A and B measures, are available from Broglio et al. [30]. Institutional Review Board (IRB) approval was obtained at the lead study site, with US Department of Defense Human Research Protection Office approval as well as local IRB approval at each participating site. This study was performed in accordance with the standards of ethics outlined in the Declaration of Helsinki.

2.2 Participants

All CARE Consortium Study participants enrolled during the 2015–2016 academic year were eligible for inclusion in this analysis. All cadets are eligible to consent for CARE participation at the US military service academies, regardless of varsity athlete status, and all varsity student-athletes are eligible to consent at the traditional colleges and universities. The only criterion for exclusion from this analysis was failure to complete the 2015–2016 baseline CARE assessment prior to sustaining a concussion during the study period. A total of 15,682 participants were included in the final dataset. All participants provided informed written consent.

2.3 Primary Outcome

The primary study outcome was sustaining a clinician-diagnosed concussion, as defined by consensus definition produced through evidence-based guidelines and adopted by the CARE Consortium Study group [30, 31], as a consequence of sport participation between August 1st, 2015 and July 31st, 2016. Participants sustaining a concussion during any form of sport participation (i.e., game or competition at varsity or non-varsity levels of play) were considered positive for the primary study outcome and negative otherwise, including participants who sustained concussions unrelated to sport.

2.4 Covariates

We considered 176 baseline covariates for each subject, grouped into the following categories: academic or military

institution, demographic variables, anthropometric variables, academic variables, primary and secondary sports, primary sport position, primary sport equipment details, concussion history, personal medical history, medications, family medical history, social history, self-reported concussion symptoms (SCAT symptom checklist), psychological and quality of life assessment results, neuro-cognitive assessment results, computerized concussion test results, balance assessment results, vision/vestibular–ocular test results (Table 1). Prior to analysis, the data were cleaned to remove obviously incorrect responses. We then mapped each variable to a set of binary features. Categorical variables were mapped to one binary feature per category. Continuous variables were first discretized based on quintile ranges, or were alternatively classified into discrete bins using clinically relevant cutoffs (e.g., normal vs. abnormal), when available. Clinically relevant cut-offs were applied for the Brief Symptom Inventory, Hospital Anxiety and Depression Scale, and Vestibular Ocular Motor Screening examination. Continuous data were then mapped to one binary feature for each discrete range/bin of each variable. This discretization procedure allowed the incorporation of domain knowledge (i.e., clinically meaningful thresholds) and extended the linear model's capacity to capture nonlinear relationships that could not be identified in the original continuous feature space. All baseline variables and their associated binary features are listed in Supplementary Appendices A and B.

2.5 Data Analysis

2.5.1 Learning Algorithm

We used a linear support vector machine (SVM) model [32] to stratify subjects' risk for sport-related concussion based on their baseline data. Nonlinear models were also considered, but performed similar to the SVM model (please refer to Supplementary Appendix C for additional detail regarding nonlinear model performance). As such, only the SVM results are reported given the greater interpretability of the SVM model, i.e., the potential that learned model coefficients may provide insight into potential risk or protective factors [33]. We assumed risk and protective factors were shared across sports, thus framing the problem as a single task and we therefore used a single task learning (STL) model. Sport-specific models were also considered, but performed similar to the STL model, likely in part because very few concussions occurred in some sports (please refer to Supplementary Appendix D for additional detail regarding sport-specific models). As such, only the STL model results are reported. We implemented our linear model using the popular Python Machine Learning package—scikit-learn (https://scikit-learn.org/stable/about.html).

Table 1 Baseline covariates

Variable category	Variable description	Number of variables prior to preprocessing	Number of features after discretization*
Institution	Individual Academic Institution or Military Service Academy[b]	1	21
Demographics	Age[a]; Sex[b]; Race[b]; Ethnicity[b]; Handedness[b]	5	29
Anthropometrics	Height[a]; Weight[a]; BMI[ac]	3	18
Academic	Current Academic Year[b]; High-School and Collegiate GPAs[a]; SAT[d] and ACT[e] Scores[a]; History of Skipped or Repeated Grade[b]; Prior Special Education/Academic Assistance[b]	17	90
Sport	Primary Collegiate Sport[b]; Secondary Sports[b]	2	77
Sport-position	Primary Sport Position[b], Starter Status[b], Depth Chart Position[b]	3	120
Sport-equipment	Sport-Specific Equipment (e.g., Helmet, Headgear, Mouth Guard)[b]	7	42
Concussion history	History[b] and Number[a] of Prior Concussions; History of Moderate/ Severe TBI[bf]; Age at First/Most Recent Concussion[a]; Duration of LOC[ag], PTA[ah], Concussion Symptoms[a]	43	60
Medical history	Personal History of ADD/ADHD[bi], ASD[bj], Balance Disorder[b], Bipolar Disorder[b], Brain Surgery[b], Depression[b], Diabetes[b], Hearing Problems[b], Learning Disorder[b], Memory Disorder[b], Meningitis[b], Headaches (Migraine, Non-Migraine)[b], Schizophrenia[b], Seizure Disorder[b], Stroke[b], Vision Problems[b]	17	50
Medications	Personal Use of Over-the-Counter and Prescription Medications[b]	15	32
Family medical history	Family History of Memory Disorder[b], Headaches (Migraine, Non-Migraine)[b], Parkinson's Disease[b] in a Parent, Grandparent, or Sibling	17	32
Social history	Use of Alcohol[b], Tobacco[b], Marijuana[b], Illicit Drugs[b] in the Past Month; Household Income[b]; Parental Education[b] and Occupation[b]; Average Hours of Sleep[a]	11	85
Symptoms	SCAT[k] Symptom Number[a] and Severity Score[a]	2	10
Psychological/quality of life	BSI-18[bl], BSSS[am], HADS[bn], SF-12[ao], SWLS[bp] Scores	20	46
Neurocognitive assessments	Total SAC[q] Score[a]; Mean RT_{clin}[ar]	9	12
Computerized concussion tests	ImPACT[as], CogState/AXON[a], CNS-Vital Signs[a], ANAM[at] Results	27	161
Balance tests	BESS[au] and SOT[av] Results	7	42
Vision/vestibular–ocular tests	King-Devick[a] and VOMS[bw] Results	8	30

*Refer to Supplementary Appendix B for a complete feature list by variable category; a: continuous feature; b: binary or categorical feature; c: body mass index; d: Scholastic Aptitude Test; e: American College Test; f: Traumatic Brain Injury; g: loss of consciousness; h: post-traumatic amnesia; i: attention deficit disorder/attention deficit-hyperactivity disorder; j: autism spectrum disorder; k: Sport Concussion Assessment Tool; l: Brief Symptom Inventory-18 (normal/abnormal classification based on BSI reference values); m: Brief Sensation Seeking Scale; n: Hospital Anxiety and Depression Scale (normal/abnormal classification based on HADS reference values); o: Short Form-12; p: Satisfaction With Life Scale (classified based on SWLS reference values); q: Standardized Assessment of Concussion; r: Clinical Reaction Time; s: Immediate Post-Concussion Assessment and Cognitive Testing; t: Automated Neuropsychological Assessment Metrics; u: Balance Error Scoring System; v: Sensory Organization Test; w: vestibular/ocular-motor screening (normal/abnormal classification based on convergence > 5 cm and any symptoms reported during convergence, saccade, pursuit, vestibular–ocular reflex, visual-motion sensitivity testing)

2.5.2 Model Selection

We trained the model by repeating the following process 20,000 times (note: 20,000 repetitions were performed so we could assess the statistical significance of the predictive factors, as described in Sect. 2.5.3):

1. Split. Data were split into a training set and a held-out test set such that approximately 80% of participants contributed to the training set and approximately 20% to the test set. For sports with less than 10 participants, all

participants were included in the training set. For sports with more than 10 participants, we employed stratified splits to ensure equal proportions of positive and negative examples (i.e., participants with and without sport-related concussions) between the training and test sets, except when less than 5 positive examples were present (i.e., less than 5 participants sustained a concussion), in which case we used random splits.

2. Train. A model was trained (i.e., learned the model parameters) using the training data set with a pre-selected hyperparameter to control for the tradeoff

between regularization and loss (i.e., to optimize model performance without overfitting). Such a setup is essential in high-dimensional settings to avoid simply memorizing the data.

We pre-selected the hyperparameter C by repeated nested cross-validation with a grid search (i.e., $C = [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$) and five-fold cross validation [34] on the training data-set for 100 splits. For computational efficiency, we used the mode of the resultant 100 hyperparameters ($C = 10^{-4}$) as the preselected hyperparameter for all 20,000 models (please refer to Supplementary Appendix E for additional detail regarding hyperparameter selection).

3. Test. The trained model was applied to the test dataset to evaluate its performance, as quantified using the area under the receiver operating characteristics curve (AUROC). AUROC is a common measure of a model's ability to discriminate positive vs. negative examples (i.e., participants with vs. without sport-related concussion) [35] and intuitively represents the probability of the model correctly ranking two randomly chosen examples. A model with AUROC of 1 would be perfect, while a model with AUROC of 0.5 would perform no better than chance.

2.5.3 Model Performance and Predictive Factor Analysis

By repeating the above process 20,000 times, we produced an empirical distribution of model performance allowing calculation of the mean AUROC with an associated 95% confidence interval. The confidence interval was calculated using the 2.5th percentile and 97.5th percentile values of the empirical distribution of AUROCs. Using a linear model allowed us to investigate the learned model parameters, thus identifying potential risk/protective factors. We considered those features with the same sign over the 20,000 repeated runs as statistically significant since features with the same sign over k runs would have a p value of at most $1/k$ when testing a null hypothesis of zero effect size for that feature. Therefore, to claim statistical significance at a p value $< 5.22 \times 10^{-5}$ (based on a Bonferroni Correction to account for comparisons over 957 features, $0.05/957 = 5.22 \times 10^{-5}$), k must be at least 19,140 (equivalent to $1/k < 0.05/957$) which we elected to round up to a total of 20,000 repetitions of the experiment.

3 Results

Of the 15,682 study participants included in this analysis, 595 (3.79%) sustained a subsequent sport-related concussion during the study period. Characteristics of the study population are presented in Table 2.

Table 2 Study population characteristics

Characteristic	Study population ($N = 15,682$)
Age [median (interquartile range) in years]	19 (18, 21)
Biological sex (%)	
Male	65.8
Female	34.2
Height [median (interquartile range) in cm]	177.80 (170.18, 185.42)
Weight [median (interquartile range) in kg]	74.84 (65.77, 86.18)
BMI [median (interquartile range)]	23.8 (22.1, 25.8)
Race (%)	
African-American	10.7
American Indian/Alaska Native	0.5
Asian	4.2
Native Hawaiian/Pacific Islander	0.8
White	73.8
Multiple races	7.3
Unknown/skipped/missing	2.8
Ethnicity (%)	
Hispanic	9.3
Non-Hispanic	80.5
Unknown/skipped/missing	10.2
Academic year (%)	
Freshman	31.0
Sophomore	25.1
Junior	22.2
Senior	18.2
5th year senior	1.7
Graduate	0. 7
Skipped/missing	1.1
Concussion history (%)	
Yes	23.3
No	76.7
Other medical history (%)	
Migraine headache	10.4
ADD/ADHD	4.8
Learning disorder	2.0
Depression	2.5

After preprocessing, our model included 957 binary features (Supplementary Appendices A and B). After splitting the data, our training sets contained 12,539 baseline records (476 positive for sport-related concussion) and our test sets contained 3143 baselines (119 positive for concussion), on average. Applied to the test sets, our model achieved a mean AUROC of 0.73 [95% CI 0.70–0.76]. Figure 1 presents the mean ROC curve. Model performance varied across sports, as did the number of participants available for both the training and test data sets (Fig. 2). Despite American football having the greatest number of
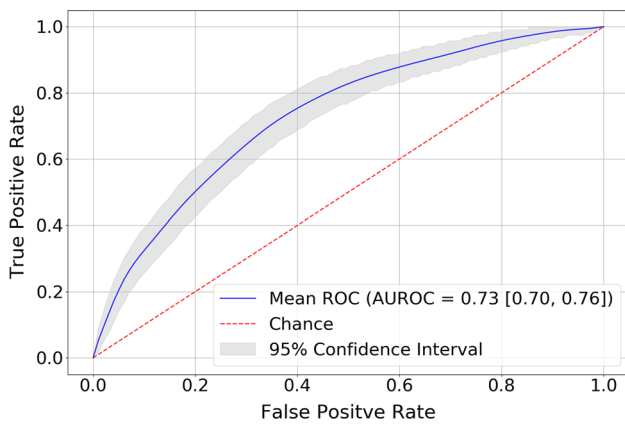
Fig. 1 Mean receiver operating characteristics (ROC) curve on the test set. The mean ROC curve is generated by first evaluating all the models on the test set and then macro-averaging all the true-positive rates interpolated at 100 evenly spaced false-positive rates starting from 0 to 1 inclusively

concussions, model performance was strongest for Swimming (AUROC = 0.86 [95% CI 0.61–1.00]) and weakest for Cheerleading (AUROC = 0.41 [95% CI 0–1.00]), where the model performed no better than chance.

The mean effect sizes of those features reaching statistical significance is illustrated in Fig. 3. Features with a negative effect sizes can be interpreted as protective (i.e., reduce estimated risk of concussion), while features with a positive effect size are risk factors (i.e., increase estimated risk of concussion). Of the 259 statistically significant features in the model, 179 are protective and 80 are risk factors. The most heavily weighted risk and protective factors reaching statistical significance in each variable category are listed in Table 3.
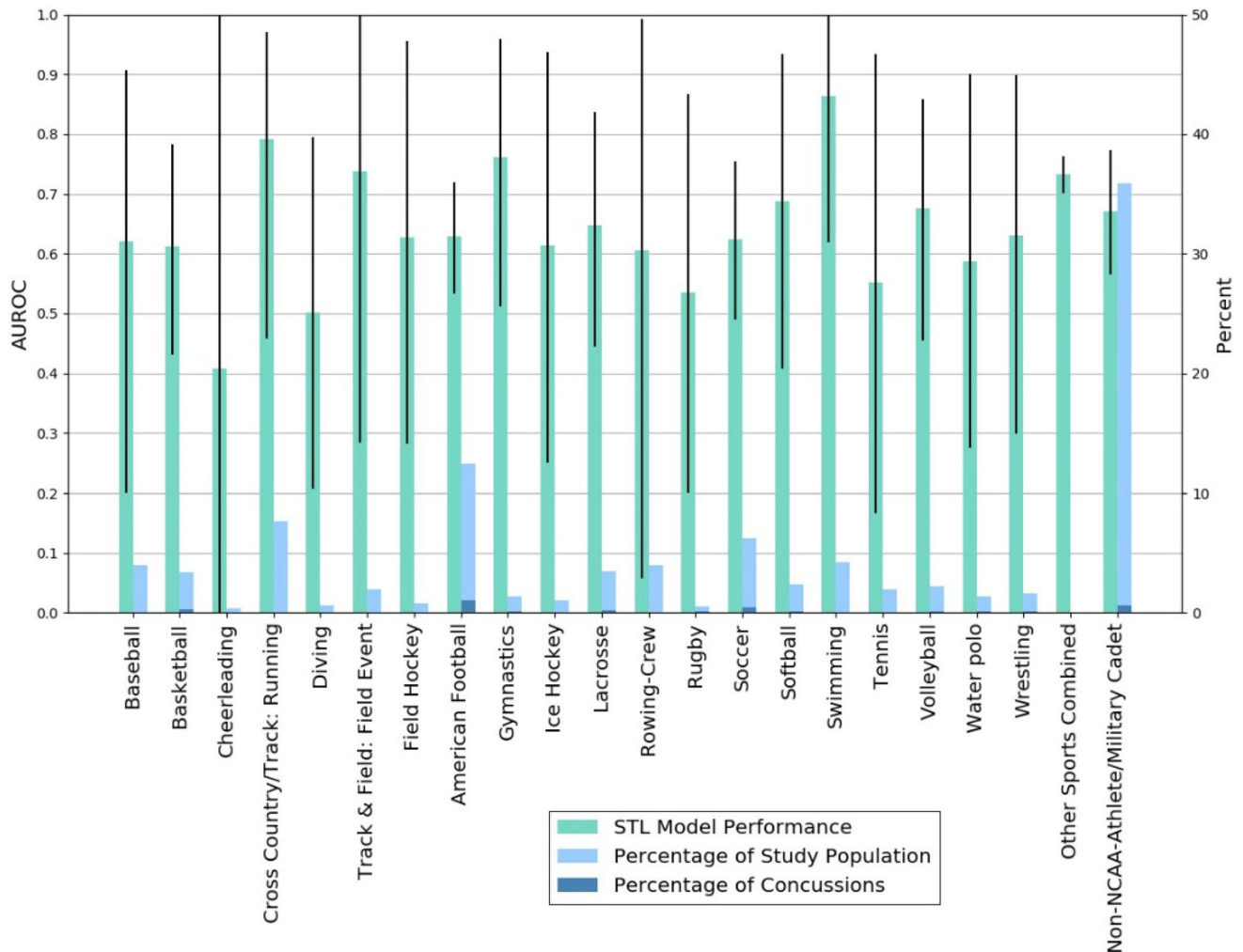
Fig. 2 Sport-specific model performance and fractions of the entire study population and concussions observed across sports. AUROC error bars represent 95% confidence intervals
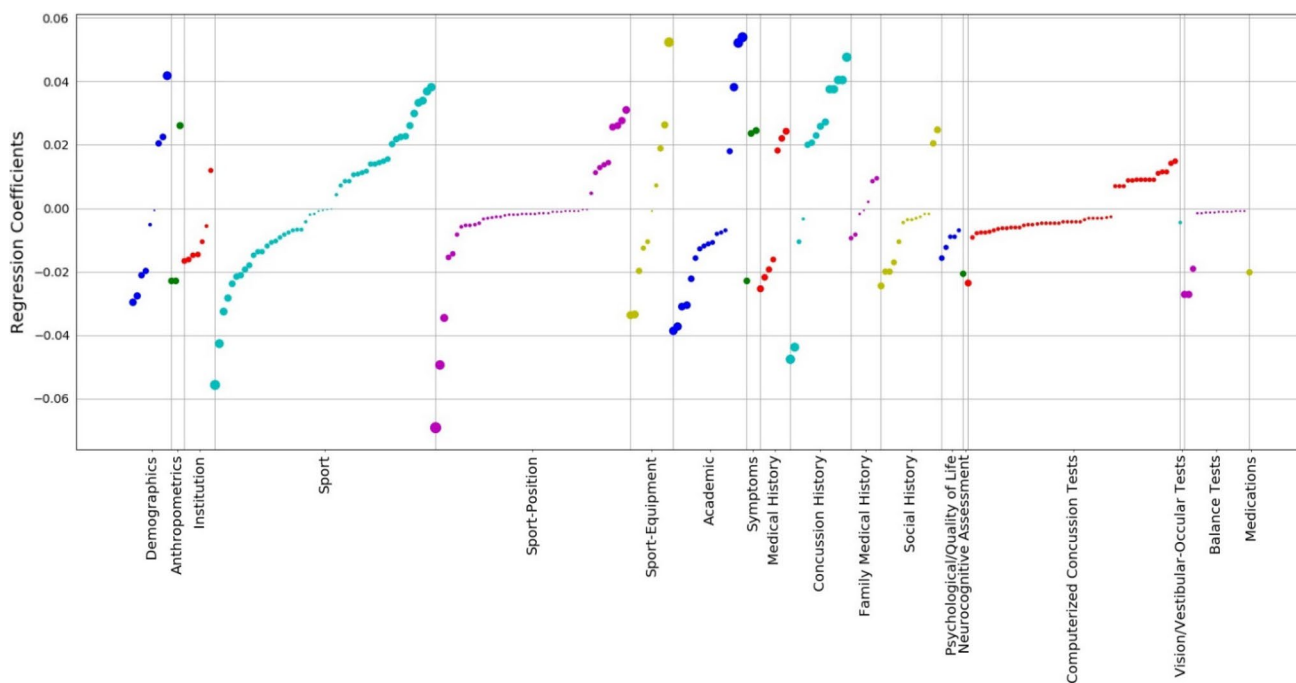
Fig. 3 Relative weights (regression coefficients/mean effect sizes, mES) of all statistically significant features ($p = 0.00005$) identified by the STL model, illustrated by variable category. Higher values are associated with greater risk of sport-related concussion, while lower values are associated with lower risk. Dot radius is proportional to mean effect size absolute value. See Table 3 for features with the smallest and largest mES in each variable category, and Supplementary Appendices A and B for all individual feature identifications

4 Discussion

Concussion is a serious injury affecting athletes and military service members that can be associated with significant short- and long-term morbidity [36, 37]. Since many concussions occur during sport participation, the ability to identify athletes at elevated risk for sustaining sport-related concussion would be of significant clinical value. Early identification of at-risk athletes could allow sports medicine providers to apply more targeted preventative measures, educational intervention, and injury surveillance strategies, and has the potential to influence injury management to prevent re-injury. To this end, we sought to develop a risk-stratification tool capable of identifying athletes and cadets at elevated risk for sport-related concussion. To our knowledge, this is the first successful attempt to develop a model for predicting concussions from prospectively collected baseline data. Using only data elements collected during athletes' and cadets' baseline pre-participation examinations, we developed a risk-stratification model that was able to predict those who would go on to sustain a sport-related concussion during the same academic year with an overall mean AUROC of 0.73. This is remarkable considering this model's ability to predict future concussions falls within the range of sensitivities and specificities demonstrated by existing concussion assessment tools for distinguishing already concussed from non-concussed athletes [38–41]. The variability in model performance across sports is not surprising given the large differences in the number of athletes participating in each sport as well as the number of observed concussions.

While the primary goal of this study was to develop a predictive algorithm for sport-related concussion, evaluation of heavily weighted factors in the model can provide insight and generate hypotheses regarding potential concussion risk and protective factors. In a recent evidence-based systematic review of risk factors for sport concussion, Abrahams et al. identified only two high-certainty concussion risk factors: having sustained two or more previous concussions and match vs. practice play [12]. Of these, only concussion history is a baseline characteristic. Not surprisingly, the current investigation also identified concussion history as a significant predictor of subsequent sport-related concussion, and not having a concussion history as protective. With respect to concussion history, recent work has demonstrated that sustaining a first concussion at an earlier age is associated with a higher number of subsequent concussions [42]. Similarly, while more specific to repetitive sport-associated head impact exposure than diagnosed concussions, other evidence in retired professional American football athletes has implicated earlier age of first exposure to tackle football as a potential risk factor for adverse long-term neuroanatomical, neurocognitive, and neuropsychiatric outcomes [43–46]. It

Table 3 Statistically significant risk and protective factors of greatest mean effect size (mES) by variable category

Variable category	Largest mES/greatest risk: feature name (mES)	Smallest mES/lowest risk: feature name (mES)
Demographics	Age in 1st quintile (mES = 0.042)	Right-handed (mES = − 0.029)
Anthropometrics	Weight in 5th quintile (mES = 0.026)	Height in 3rd quintile (mES = − 0.023)
Institution	Consortium site S (mES = 0.012)	Consortium site A (mES = − 0.016)
Sport	Primary sport is soccer (mES = 0.038)	Primary sport is track and field running event/cross-country runner (mES = − 0.056)
Sport-position	Expected first string position on depth chart (mES = 0.031)	Expected depth chart position missing (mES = − 0.069)
Sport-equipment	Wears a mouth guard (mES = 0.052)	Football helmet details not applicable, i.e., non-football athlete (mES = − 0.034)
Academic	Collegiate GPA[a] not-applicable, i.e., new freshman (mES = 0.054)	Current collegiate GPA in 4th quintile (mES = − 0.038)
Symptoms	SCAT[b] 3 Symptom Severity Score in 4th quartile (mES = 0.025)	SCAT[b] 3 Symptom Severity Score in 2nd quintile (mES = − 0.023)
Medical history	Positive for ADD/ADHD[c] (mES = 0.024)	Negative for ADD/ADHD[c] (mES = − 0.025)
Concussion history	Yes/positive (mES = 0.048)	Number of concussions = 0 (mES = − 0.047)
Family medical history	Parental history of non-migraine headache disorder missing (mES = 0.010)	Positive parental history of non-migraine headache disorder (mES = − 0.009)
Social history	No alcohol use in the past month (mES = 0.025)	Estimated family income, i.e., parents/guardians, between $60,001 and $90,000 (mES = − 0.024)
Psychological/quality of life	*N/A**	Satisfaction with Life Scale Score missing (mES = − 0.016)
Neurocognitive Assessment	*N/A**	SAC[d] total score missing (mES = − 0.021)
Computerized concussion tests	ANAM[e] Mathematical Processing Score missing (mES = 0.015)	ImPACT[f] Verbal Memory Composite Score in 5th quintile (mES = − 0.023)
Vision/vestibular–Ocular Tests	*N/A**	Abnormal VOMS[g] near point convergence (mES = − 0.004)
Balance tests	*N/A**	BESS[h] total score missing (mES = − 0.027)
Medications	*N/A**	Prescription birth control medication not applicable, i.e., males (mES = − 0.020)

*Not-applicable due to no feature reaching statistical significance in this category; a: grade point average; b: Sport Concussion Assessment Tool; c: attention deficit disorder/attention deficit-hyperactivity disorder; d: Standardized Assessment of Concussion; e: Automated Neurological Assessment Metrics; f: Immediate Post-Concussion Assessment and Testing; g: Vestibular Ocular Motor Screening (abnormal classification based on convergence > 5 cm); h: Balance Error Scoring System

is, therefore, interesting to note in the present investigation that earlier age of first concussion did not receive a significant weight in the model. Other results from the CARE Consortium and elsewhere have similarly failed to identify an association between age of first exposure and outcomes during adolescence/young adulthood [47–49].

In addition to those associated with concussion history, some other heavily weighted factors in this study came as little surprise. Concussion rates are consistently found to differ across sports, with contact and collision sports associated with greater concussion risk compared to sports with limited or no contact [50–52]. In keeping with previous research, [26, 53], this investigation also found a number of contact/collision sports were associated with an increased concussion risk (e.g., soccer, basketball, American football, water polo, rugby, diving, lacrosse), while a number of non-contact sports were associated with decreased risk (e.g., cross-country/track, rowing, swimming, baseball, golf, field

events, tennis). Additionally, this study identified female sex as a concussion risk factor, with male sex being protective. This sex-based result is consistent with previous research comparing concussion rates between males and females in sex-comparable sports with similar rules and physicality between males and females [13, 53, 54], but the present study extends this finding by analyzing concussion risk across all sports.

It is interesting to note that reporting higher baseline symptoms (3–5 positive symptoms on the SCAT Symptom Checklist; SCAT Symptom Severity Score 8–86) tended to be associated with higher weight in the model, while lower baseline concussion symptoms (SCAT Symptom Severity Score of 2–3) tended to be associated with negative weight. In addition, several significant predictive factors identified in this study are associated with neurological comorbidities generally considered to represent concussion modifying factors. For example, the model identified the presence

of migraine headache at baseline, either by self-report of a prior medical diagnosis or a positive response to the ID Migraine Questionnaire [55] as a heavily weighted concussion risk factor, while the absence of migraine headaches was protective. These findings offer prospective evidence that is consistent with previous speculation for migraine headache as a potential concussion risk factor [56, 57]. In a similar manner, this model identified a history of ADHD [58], or learning disability at baseline as associated with higher risk, while the absence of both conditions at baseline was protective. While a comparable trend was present for history of depression, only the absence of depression at baseline reached statistical significance as a protective factor after Bonferroni correction. Also, it is interesting to note while a normal baseline Hospital Anxiety and Depression Scale (HADS)-Anxiety score at baseline was a significant protective factor, the risk associated with HADS-Anxiety scores falling in the abnormal or borderline-abnormal ranges failed to reach statistical significance as a risk factor.

It is important to recognize that a feature's significance as a risk or protective factor in this predictive statistical model does not imply causality. In fact, it is likely that a number of significant predictive factors in the model are correlated with concussion risk in the absence of any clinically interpretable causal relationship. For example, an abnormal Anxiety score on the BSI-I8 at baseline was identified a significant protective factor, in contrast to the decreased risk associated with normal HADS-Anxiety scores described above. Furthermore, the model identified being right-handed as a significant protective factor, not using marijuana or alcohol in the previous month as significant risk factors, and missing answers to the marijuana and alcohol questions as significantly protective. In addition, the model identified many academic features as significant predictors of sport-related concussion. For example, several factors corresponding to higher academic performance (e.g., having a high-school GPA in the 4th quintile or a collegiate GPA in the 4th or 5th quintile; having an ACT-Math score in the 4th or 5th quintile, an ACT-Science score in the 3rd or 5th quintile, or an ACT-Reading score in the 4th quintile) were associated with lower risk of concussion, while having a total ACT score in the 1st quintile, corresponding to lower academic performance, was associated with higher risk. The model also identified several significant factors associated with age. Specifically, freshman class status and being in the lowest age quintile (age ≤ 18 years) were significant concussion risk factors, while senior and "fifth-year-senior" class status as well as being in the fourth or fifth age quintiles (20 years $<$ age ≤ 21 years; 21 years $<$ age ≤ 30 years) were identified as being protective. One might hypothesize the age/class-status trend to be associated either with physical maturation over a collegiate career or a "weed-out" effect with some athletes who sustained early concussions not

continuing to compete through their entire academic careers. However, in many cases significant model observations fail to suggest clinically relevant hypotheses.

It is also noteworthy that a high-degree of collinearity was present among many of the baseline variables assessed. As such, the significance of some heavily -weighted factors in the model has a high likelihood of being attributable to an association with other more obvious risk or protective factors. For example, the observation that wearing protective gear and wearing a mouth guard are both concussion risk factors is likely attributable to protective equipment and mouth guards both being more frequently worn in higher risk contact sports. Furthermore, the observation of African-American race and weight falling into the 5th quintile are concussion risk factors may be attributed to disproportionate participation in American football by African-Americans (about one-third of African-Americans in the study population were football players) and athletes in the highest quintile for weight (about half of the athletes in this weight quintile were football players).

Ultimately, individual features, regardless of their weight, must be interpreted in the context of the entire model and not as independent risk or predictive factors. As such, it cannot be assumed based on the present results that intervening on a potentially modifiable risk or protective factor would necessarily influence an athlete's future concussion risk. Such a conclusion would require additional support from a prospective intervention trial. Identification of heavily weighted factors, especially when unanticipated, should prompt novel hypothesis generation and future concussion research. For example, based on the greater risk of sport-related concussion observed in younger freshman athletes and cadets, it would be reasonable for future research to investigate the more gradual incorporation of incoming freshman into varsity collegiate sport participation as a concussion risk reduction strategy.

This study was not without limitations. Despite likely differences in the mechanisms of injury causing concussions in different sports, as well as differences between the varsity and non-varsity levels of play, we were not able to develop sport-specific models. This was largely due to variability in the number of athletes participating in different sports and a small number of positive examples (i.e., concussions) in some sports, leading to over-fitted models that did not generalize well to test data. Other factors were also likely at play, given that American football, which included more participants than any other single varsity sport as well as the greatest number of concussions, had an AUROC falling below that of the full model. Given challenges in developing sport-specific models, we employed a model trained using data from all sports at both the varsity and non-varsity levels of play. With additional data, future research could seek to develop sport- and level-of-play-specific models, as

well models specific to military cadets, which might have greater predictive ability. Another study limitation was the presence of "missing data" and the potential for non-values to be coded in more than one way for many variables. For example, non-values could potentially be coded as "skipped" (i.e., a "skip this question" response was selected/provided), "missing" (i.e., no response was selected/provided), "unknown" (i.e., a "don't know" response was selected/provided), or "N/A" (i.e., the question did not apply to the athlete/cadet; e.g., collegiate GPA for an incoming freshman, helmet type for an athlete participating in a non-helmeted sport, or results of a "Level B" measure or computerized concussion test not used at the athlete's/cadet's institution) for many variables. Given that missing data rates varied greatly across features (from a minimum of 0.6% to a maximum of 99.9%), we elected to analyze the dataset retaining all potential non-value options to account for the possibility they might contain predictive information. However, it is challenging to develop clinically relevant hypotheses for most significant non-value features identified in the model. In addition, it is also possible that some inaccurate self-reported information may have remained in the dataset despite our careful review and attempt to remove obviously incorrect data. It would be impractical to independently review records to verify all self-reported information collected during CARE baseline assessments so self-reporting errors are possible. Furthermore, while we attempted to discretize continuous variables using clinically relevant cut-offs when available, our procedure for mapping other continuous variables to discrete sets of binary features using quintile ranges may not have captured clinically important or statistically significant cut-offs in some cases. We elected to use quintiles because of their standardization and reproducibility, but other discretization strategies may have yielded different results. Next, while the effect size values reported in Supplementary Appendices A and B are meaningful for comparing the relative model weights between features in the present study, they cannot be interpreted outside the context of this study in the same way as standardized effect size values like Cohen's D. In addition, these results should not be extrapolated beyond a collegiate athlete/military cadet population. Even within a population of collegiate athletes and military cadets, it would be challenging to apply this model outside of the CARE Consortium study given the number of baseline variables used as model inputs, many of which are not routinely collected outside of the CARE study protocol. Future work should build on the present results to develop more streamlined models utilizing a subset of the most predictive features so that the models can more easily be applied in a routine clinical setting. In a similar vein, these results should not be extrapolated to non-sport concussions. In fact, since this study focused only on sport-related concussion, it is possible that some participants considered

negative for the primary study outcome could potentially have sustained concussions outside of sport participation during the study period. Lastly, given potential athlete under-reporting of concussion symptoms and because there is no objective confirmatory test for concussion, the concussion diagnosis relies both on accurate injury identification as well as the clinical impression of the evaluating medical provider. While issues of potential missed injuries and diagnostic uncertainty are common across all concussion research, the CARE Consortium's use of a standard concussion definition by all sites should at least mitigate the potential for diagnostic uncertainty. Nonetheless, any diagnostic inaccuracy is undesirable in the context of developing a data-driven risk prediction model where an accurate classification of "case-ness" is relied upon heavily for the analysis and will render injury prediction and evaluation of those predictions more challenging.

5 Conclusion

This collaborative data-driven study leverages powerful analytical techniques and a robust clinical dataset to develop a novel model for predicting collegiate athletes' and military cadets' risk of sustaining sport-related concussion. This is clinically important because, to our knowledge, it represents the first successful attempt to predict sport-related concussion using only baseline data. As such tools are developed and refined, clinicians will increasingly need to determine how to apply them in clinical practice. Might a future model perform so accurately that one day certain athletes would be restricted from participating in certain sports due to the amount of model-predicted risk? For now, this study suggests it is feasible to identify athletes at elevated risk of sport-related concussion in whom targeted prevention, education, and injury surveillance strategies may be employed. Furthermore, this study offers insight into potential risk and protective factors for sport-related concussion, generating novel hypotheses for future concussion research. Future work is needed to develop a predictive model that can be easily applied in routine clinical practice, as well as to identify modifiable concussion risk factors that can be intervened upon to modify an athlete's injury risk.

not necessarily endorsed by the Office of the Assistant Secretary of Defense for Health Affairs.

Declarations

References

1. Langlois JA, Rutland-Brown W, Wald MM. The epidemiology and impact of traumatic brain injury: a brief overview. J Head Trauma Rehabil. 2006;21(5):375–8.
2. National Collegiate Athletic Association. Number of NCAA college athletes reaches all-time high. 2018.
3. US Department of Defense. Special report—traumatic brain injury. 2015.
4. BrainLine, How Many Service Members Have Sustained a TBI? Defense Medical Surveillance System, 2018.

5. Guskiewicz KM, et al. Recurrent concussion and risk of depression in retired professional football players. Med Sci Sports Exerc. 2007;39(6):903–9.

6. Guskiewicz KM, et al. Association between recurrent concussion and late-life cognitive impairment in retired professional football players. Neurosurgery. 2005;57(4):719–26 **(discussion 719–2)**.

7. McKee AC, et al. Chronic traumatic encephalopathy in athletes: progressive tauopathy after repetitive head injury. J Neuropathol Exp Neurol. 2009;68(7):709–35.

8. McKee AC, et al. The spectrum of disease in chronic traumatic encephalopathy. Brain. 2013;136(Pt 1):43–64.

9. McKee AC, et al. The neuropathology of chronic traumatic encephalopathy. Brain Pathol. 2015;25(3):350–64.

10. Mez J, et al. Clinicopathological evaluation of chronic traumatic encephalopathy in players of American Football. JAMA. 2017;318(4):360–70.

11. Caccese JB et al. Does baseline concussion testing aid in identifying future concussion risk? Res Sports Med. 2019:1–6.

12. Abrahams S, et al. Risk factors for sports concussion: an evidence-based systematic review. Br J Sports Med. 2014;48(2):91–7.

13. Harmon KG, et al. American Medical Society for Sports Medicine position statement: concussion in sport. Br J Sports Med. 2013;47(1):15–26.

14. Kerr HA. Concussion risk factors and strategies for prevention. Pediatr Ann. 2014;43(12):e309–15.

15. Noble JM, Hesdorffer DC. Sport-related concussions: a review of epidemiology, challenges in diagnosis, and potential risk factors. Neuropsychol Rev. 2013;23(4):273–84.

16. Finnoff JT, Jelsing EJ, Smith J. Biomarkers, genetics, and risk factors for concussion. PM R. 2011;3(10 Suppl 2):S452–9.

17. Hollis SJ, et al. Incidence, risk, and protective factors of mild traumatic brain injury in a cohort of Australian nonprofessional male rugby players. Am J Sports Med. 2009;37(12):2328–33.

18. Schneider KJ, et al. Preseason reports of neck pain, dizziness, and headache as risk factors for concussion in male youth ice hockey players. Clin J Sport Med. 2013;23(4):267–72.

19. Curran-Sills G, Abedin T. Risk factors associated with injury and concussion in sanctioned amateur and professional mixed martial arts bouts in Calgary, Alberta. BMJ Open Sport Exerc Med. 2018;4(1):e000348.

20. Yeung A, Munjal V, Virji-Babul N. Development of the sports organization concussion risk assessment tool (SOCRAT). Brain Inj. 2017;31(4):542–9.

21. Baugh CM, et al. Frequency of head-impact-related outcomes by position in NCAA division I collegiate football players. J Neurotrauma. 2015;32(5):314–26.

22. Black AM, et al. Policy change eliminating body checking in non-elite ice hockey leads to a threefold reduction in injury and concussion risk in 11- and 12-year-old players. Br J Sports Med. 2016;50(1):55–61.

23. Blumenfeld RS, et al. The epidemiology of sports-related head injury and concussion in water polo. Front Neurol. 2016;7:98.

24. Hollis SJ, et al. Mild traumatic brain injury among a cohort of rugby union players: predictors of time to injury. Br J Sports Med. 2011;45(12):997–9.

25. Teramoto M, et al. Style of play and rate of concussions in the national football league. Orthop J Sports Med. 2015;3(12):2325967115620365.

26. Collins CL, et al. Neck strength: a protective factor reducing risk for concussion in high school sports. J Prim Prev. 2014;35(5):309–19.

27. Dretsch MN, et al. Evaluating the clinical utility of the Validity-10 for detecting amplified symptom reporting for patients with mild traumatic brain injury and comorbid psychological health conditions. Appl Neuropsychol Adult. 2017;24(4):376–80.

28. Van Pelt KL, et al. A cohort study to identify and evaluate concussion risk factors across multiple injury settings: findings from the CARE Consortium. Inj Epidemiol. 2019;6(1):1.

29. Schulz MR, et al. Incidence and risk factors for concussion in high school athletes, North Carolina, 1996–1999. Am J Epidemiol. 2004;160(10):937–44.

30. Broglio SP, et al. A national study on the effects of concussion in collegiate athletes and US military service academy members: the NCAA-DoD concussion assessment, research and education (CARE) consortium structure and methods. Sports Med. 2017;47(7):1437–51.

31. Carney N, et al. Concussion guidelines step 1: systematic review of prevalent indicators. Neurosurgery. 2014;75(Suppl 1):S3–15.

32. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.

33. Ustun B, Traca S, Rudin C. Cynthia supersparse linear integer models for interpretable classification. 2014. arXiv:1306.6677.

34. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international joint conference on Artificial intelligence—Volume 2. 1995, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada. p. 1137–1143.

35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.

36. Wasserman EB, et al. Academic dysfunction after a concussion among US High School and College Students. Am J Public Health. 2016;106(7):1247–53.

37. Manley G, et al. A systematic review of potential long-term effects of sport-related concussion. Br J Sports Med. 2017;51(12):969–77.

38. Barlow M, et al. Differences in change scores and the predictive validity of three commonly used measures following concussion in the middle school and high school aged population. Int J Sports Phys Ther. 2011;6(3):150–7.

39. Randolph C, et al. Concussion symptom inventory: an empirically derived scale for monitoring resolution of symptoms following sport-related concussion. Arch Clin Neuropsychol. 2009;24(3):219–29.

40. Barr WB, McCrea M. Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. J Int Neuropsychol Soc. 2001;7(6):693–702.

41. Mucha A, et al. A Brief Vestibular/Ocular Motor Screening (VOMS) assessment to evaluate concussions: preliminary findings. Am J Sports Med. 2014;42(10):2479–86.

42. Schmidt JD, et al. Age at first concussion influences the number of subsequent concussions. Pediatr Neurol. 2018;81:19–24.

43. Stamm JM, et al. Age at first exposure to football is associated with altered corpus callosum white matter microstructure in former professional football players. J Neurotrauma. 2015;32(22):1768–76.

44. Stamm JM, et al. Age of first exposure to football and later-life cognitive impairment in former NFL players. Neurology. 2015;84(11):1114–20.

45. Alosco ML, et al. Age of first exposure to American football and long-term neuropsychiatric and cognitive outcomes. Transl Psychiatry. 2017;7(9):e1236.

46. Schultz V, et al. Age at first exposure to repetitive head impacts is associated with smaller thalamic volumes in former professional american football players. J Neurotrauma. 2018;35(2):278–85.

47. Caccese JB, et al. Estimated age of first exposure to american football and neurocognitive performance amongst NCAA male student-athletes: a cohort study. Sports Med. 2019;49(3):477–87.

48. Caccese JB et al. Estimated age of first exposure to contact sports is not associated with greater symptoms or worse cognitive

functioning in male U.S. Service Academy Athletes. J Neurotrauma. 2020;37(2):334–9.

49. Brett BL, et al. Age of first exposure to american football and behavioral, cognitive, psychological, and physical outcomes in high school and collegiate football players. Sports Health. 2019;11(4):332–42.

50. Koh JO, Cassidy JD, Watkinson EJ. Incidence of concussion in contact sports: a systematic review of the evidence. Brain Inj. 2003;17(10):901–17.

51. 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):311–9.

52. Marar M, et al. Epidemiology of concussions among United States high school athletes in 20 sports. Am J Sports Med. 2012;40(4):747–55.

53. Giza CC, et al. Summary of evidence-based guideline update: evaluation and management of concussion in sports: Report of the Guideline Development Subcommittee of the American Academy of Neurology. Neurology. 2013;80(24):2250–7.

54. Dick RW. Is there a gender difference in concussion incidence and outcomes? Br J Sports Med. 2009;43(Suppl 1):i46–50.

55. Lipton RB, et al. A self-administered screener for migraine in primary care: the ID Migraine validation study. Neurology. 2003;61(3):375–82.

56. Eckner JT, et al. Is migraine headache associated with concussion in athletes? A case-control study. Clin J Sport Med. 2017;27(3):266–70.

57. Gordon KE, Dooley JM, Wood EP. Is migraine a risk factor for the development of concussion? Br J Sports Med. 2006;40(2):184–5.

58. Alosco ML, Fedor AF, Gunstad J. Attention deficit hyperactivity disorder as a risk factor for concussions in NCAA division-I athletes. Brain Inj. 2014;28(4):472–4.