



Sensitivity and Specificity of Computer-Based Neurocognitive Tests in Sport-Related Concussion: Findings from the NCAA-DoD CARE Consortium

Lauren L. Czerniak¹ · Spencer W. Liebel² · Gian-Gabriel P. Garcia³ · Mariel S. Lavieri¹ · Michael A. McCrea^{4,5} · Thomas W. McAllister⁶ · Steven P. Broglio⁷ · CARE Consortium Investigators

Accepted: 19 November 2020 / Published online: 14 December 2020
© Springer Nature Switzerland AG 2020

Abstract

Background To optimally care for concussed individuals, a multi-dimensional approach is critical and a key component of this assessment in the athletic environment is computer-based neurocognitive testing. However, there continues to be concerns about the reliability and validity of these testing tools. The purpose of this study was to determine the sensitivity and specificity of three common computer-based neurocognitive tests (Immediate Post-Concussion Assessment and Cognitive Testing [ImPACT], CNS Vital Signs, and CogState Computerized Assessment Tool [CCAT]), to provide guidance on their clinical utility.

Methods This study analyzed assessments from a cohort of collegiate athletes and non-varsity cadets from the NCAA-DoD CARE Consortium. The data were collected from 2014–2018. Study participants were divided into two testing groups [concussed, $n = 1414$ (baseline/24–48 h) and healthy, $n = 8305$ (baseline/baseline)]. For each test type, change scores were calculated for the components of interest. Then, the Normative Change method, which used normative data published in a similar cohort, and the Reliable Change Index (RCI) method were used to determine if the change scores were significant.

Results Using the Normative Change method, ImPACT performed best with an 87.5%-confidence interval and 1 number of components failed (NCF; sensitivity = 0.583, specificity = 0.625, $F1 = 0.308$). CNS Vital Signs performed best with a 90%-confidence interval and 1 NCF (sensitivity = 0.587, specificity = 0.532, $F1 = 0.314$). CCAT performed best when using a 75%-confidence interval and 2 NCF (sensitivity = 0.513, specificity = 0.715, $F1 = 0.290$). When using the RCI method, ImPACT performed best with an 87.5%-confidence interval and 1 NCF (sensitivity = 0.626, specificity = 0.559, $F1 = 0.297$).

Conclusion When considering all three computer-based neurocognitive tests, the overall low sensitivity and specificity results provide additional evidence for the use of a multi-dimensional assessment for concussion diagnosis, including symptom evaluation, postural control assessment, neuropsychological status, and other functional assessments.

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

The members of the CARE Consortium Investigators are listed in acknowledgements.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40279-020-01393-7>.

✉ Lauren L. Czerniak
czernL@umich.edu

Extended author information available on the last page of the article

Key points

No computer-based neurocognitive test outperforms the others

No computer-based neurocognitive test has a sensitivity and specificity necessary for clinical utility as a standalone measure

Caring for patients using a multi-dimensional concussion assessment is recommended

1 Background

Mild traumatic brain injury (mTBI), or concussion, is the most common type of mTBI and has become a significant epidemiologic phenomenon [12]. It is estimated that 1.6–3.8 million concussions occur in sports and recreational activities annually [29]. Concussions are typically associated with increased symptom reporting and declines in neurocognitive functioning and balance [35, 41, 44, 55]. A multi-dimensional approach to concussion assessment that measures change in each of these domains is critical to concussion management protocols [36].

Neuropsychological testing has been identified as a key component of the concussion assessment protocol, and it plays a crucial role in concussion management programs at all levels of sports [3, 19]. In the athletic environment, computer-based testing is commonly implemented to establish a pre-injury baseline of neurocognitive functioning and to measure potential neurocognitive change post-injury. In comparison to traditional neuropsychological testing, computer-based testing may be advantageous for multiple reasons, including administrations using auditory and visual modalities, the ability to be given individually or in large groups simultaneously, and results are immediately available for review [42]. Despite these advantages, however, computer-based neurocognitive testing has several drawbacks. For example, there continue to be concerns about their psychometric reliability and validity [4, 48, 53]. However, the psychometric reliability and validity of pencil-and-paper neurocognitive tests in concussion assessments have also been questioned [43].

There exist many computer-based tests for the assessment of sport-related concussion (SRC), and the most popular include Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) [32], CogState Computerized Assessment Tool (CCAT) [1], Automated Neuropsychological Assessment Metrics (ANAM[®]) [45], and CNS Vital Signs [23]. ImPACT is by far the most widely implemented as it is used by 83.5% of athletic trainers [30]. Yet, CCAT, ANAM, and CNS Vital Signs remain prevalent [38].

The sensitivity and specificity of these tools have been evaluated previously with mixed findings (see [42] for review). In one study, ImPACT was reported to possess sufficient sensitivity (91.4%) to detect post-SRC neurocognitive impairment but lower specificity (69.1%) [51]. A separate group, however, reported much lower sensitivity for ImPACT (62.5%) [7]. For CCAT, it has been reported that it has both high sensitivity and specificity for the detection of SRC neurocognitive impairment [31]. Conversely, Maruff et al. [34] reported that CCAT was sufficiently sensitive to distinguish healthy adults and patient

samples, but it was not sufficiently specific to distinguish groups of patients with mTBI. In a study investigating the sensitivity and specificity of ANAM modules in detecting SRC, sensitivity was low (< 1–6.6% at 95% CI) and specificity was high (94–100% at 95% CI) [46]. A study comparing all three tests, ImPACT, CCAT, and ANAM, concluded sensitivities of 67.8%, 60.3%, and 47.6%, respectively [39]. And finally, CNS Vital Signs has been shown to adequately discriminate between various non-SRC clinical groups, but its ability to do this in SRC samples is unknown (e.g. Ref. [9]).

As computer-based neurocognitive tests are wide-spread throughout SRC management programs, it is essential for medical teams to employ the most clinically useful measures to ensure appropriate patient care following SRC. However, it is currently unknown which computer-based neurocognitive test battery is optimal for the clinical care of SRC. Extant research on the sensitivity and specificity of these assessment tools are often underpowered and/or do not examine these constructs in SRC samples specifically. Thus, the aim of this investigation was to evaluate the accuracy of computer-based neurocognitive tests commonly implemented for SRC evaluations on a large scale and provide guidance on their clinical utility.

2 Methods

2.1 Study Participants

Study participants consisted of individuals from 30 National Collegiate Athletic Association (NCAA) military service academies and civilian universities who served as a cadet or participated in a NCAA sport during the 2014–2018 academic years (29 of the institutions currently provide data). Data for this study were provided by the Concussion Assessment, Research, and Education (CARE) Consortium [8]. Specifically, the CARE dataset contained 47,397 pre-season baseline (baseline) examinations and 2752 examinations performed 24–48 h (24–48 h) post-concussion. The 24–48 h examination was completed if a study participant was diagnosed as concussed by the local medical team at their institution using a standardized injury definition [11]. Individuals self-reporting a diagnosis of attention-deficit/hyperactivity disorder (ADD/ADHD) or a learning disability (LD) were excluded from the analysis ($n = 1060$). Excluding these individuals is supported by Elbin et al. [20] who found patients self-reporting ADD/ADHD and/or LD performed significantly worse on the components of the baseline ImPACT neurocognitive test. However, future research can extend the analysis presented in this manuscript to the subgroup of athletes in the CARE dataset with ADD/ADHD given the high prevalence of these individuals within the collegiate

athlete/cadet population. All individuals provided written informed consent which was approved by the local institution and the US Army Human Research Protection Office. Also, all computations were completed using the software R, Version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

2.2 Measurements

To evaluate both sensitivity (proportion of sample correctly identified as having sustained a SRC) and specificity (proportion of sample correctly identified as not having sustained a SRC), two groups were identified from the data. The first group consisted of individuals completing a baseline test and were later diagnosed with concussion, resulting in a 24–48 h test within the same academic year (i.e., baseline/24–48 h). The second group consisted of individuals who completed a baseline test in two consecutive academic years (i.e., baseline/baseline). In this group, individuals were removed if they experienced a concussion during the 1-year gap separating baseline tests. From here, each study group was stratified by test type: ImPACT, CNS Vital Signs, and CCAT. Baseline tests were screened for validity using embedded metrics and were excluded from the analysis if declared invalid ($n = 340$).

Change scores were then calculated for each component within each test type. A change score in the baseline/24–48 h group was calculated by subtracting the baseline score from the 24–48 h score. Similarly, a change score in the baseline/baseline group was calculated by subtracting the first baseline score from the second baseline score. Therefore, a negative change score showed an individual scored lower on the second test whereas a positive change score showed an individual scored higher on the second test. Further, the aim of this study was to analyze the neurocognitive components of the tests, so symptom scores were not considered in this analysis.

2.2.1 Immediate Post-concussion Assessment and Cognitive Test (ImPACT)

The ImPACT test is a “computerized neurocognitive test battery that is used to assess Sequencing/Attention, Word Memory, Visual Memory, and Reaction” [27]. The four components of interest for this study were the Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time Composites. Change scores were calculated for each. In general, higher scores indicated “better” performance with the exception of the Reaction Time Composite score, where lower/faster scores were better. Therefore, a negative change score represented improved performance for the Reaction Time Composite.

2.2.2 CNS Vital Signs (CNS)

The CNS Vital Signs computer test is a “clinical testing procedure used by clinicians to evaluate and manage the neurocognitive state of a patient” [14]. The eleven components of interest for this study were the Simple Attention Percentile, Composite Memory, Verbal Memory, Visual Memory, Psychomotor Speed, Reaction Time, Complex Attention, Cognitive Flexibility, Processing Speed, Executive Function, and Motor Speed Standard Scores. Change scores were calculated for each of these components and a positive change score represented improved performance for all measurements of interest.

2.2.3 Cogstate Computerized Cognitive Assessment Tool (CCAT)

The CCAT is a computer test that uses “psychological techniques to record learning, memory, processing speed and accuracy” [1]. Change scores for the Composite Processing Speed, Composite Attention, Composite Learning, and Working Memory Speed components were calculated and a positive change score represented improved performance for all measurements of interest.

2.2.4 Missing Data

Within the study, each participant in the ImPACT, CNS Vital Signs, and CCAT study groups required four, eleven, and four change scores, respectively. If any of these change scores were missing, the study participant was removed from the appropriate study group [ImPACT: baseline/24–48 h = 3, baseline/baseline = 5; CNS Vital Signs: baseline/24–48 h = 7, baseline/baseline = 56; CCAT: baseline/24–48 h = 1, baseline/baseline = 3]. In addition, the initial study sought to analyze the ANAM test, but insufficient data precluded the ability to do so (baseline/24–48 h = 62, baseline/baseline = 3).

2.3 Data Analysis

Demographic variables (age, gender, race, height, weight, number of previous concussions) were compared between the baseline/24–48 h and baseline/baseline groups using the non-parametric Mann–Whitney U test for continuous/ordinal variables and the non-parametric Chi-Squared test for categorical variables. A significance level of $\alpha = 0.05$ was considered significant. If significant differences were determined, a one-sided Mann–Whitney U test determined the directionality for continuous and ordinal variables. For categorical variables, the directionality was determined using the contingency tables from the Chi-Squared test. Sensitivity and specificity for the ImPACT, CNS Vital Signs, and CCAT

were evaluated using the Normative Change method (Normative) and Reliable Change Index method (RCI).

2.3.1 Normative Change Method

To evaluate change on single or multiple components of the ImPACT, CNS Vital Signs, and CCAT tests, the Normative method implemented change scores developed utilizing CARE Consortium data captured from 2014 to 2017 [6]. With this method, a study participant's change score for a specific component was considered "failed" if the change score fell outside of a given normative change confidence interval. Consistent with previous work, we evaluated the 75, 87.5, 90, 92.5, 95, 97.5, and 99 percent one-sided confidence intervals [6]. The overall classification of the study participant was determined by the Number of Components Failed (NCF) for a specific assessment. NCF represents how many neurocognitive test components a study participant would need to fail (with respect to the normative change confidence interval) for that specific study participant to be classified as concussed. For example, a NCF of two would mean the study participant's change scores would need to exceed the confidence interval for two or more components to be classified as concussed.

2.3.2 Reliable Change Index

The RCI method is defined for the ImPACT, but not the CNS Vital Signs nor CCAT tests. Therefore, this method was only used to analyze the ImPACT baseline/24–48 h and ImPACT baseline/baseline groups. With the RCI method, a meaningful change on a given component was noted if the study participant's change score fell outside of a given reliable change confidence interval. RCI calculations provided by the *ImPACT Administration and Interpretation Manual* [27] were completed. To be consistent with the normative method, 75, 87.5, 90, 92.5, 97.5, and 99 one-sided confidence intervals were used. In addition, the NCF determined if a study participant was classified as concussed.

2.3.3 Four Models Studied

Using the Normative method and RCI method, this study analyzed the performance of four models: ImPACT Normative, CNS Vital Signs Normative, CCAT Normative, and ImPACT RCI. Each of these models were analyzed with different confidence intervals and NCF values.

2.4 Test Performance Measures

The purpose of this study was two-fold. First, to characterize the ability of each neurocognitive test to discriminate between concussed and healthy participants by varying (1)

the cut-off point (as determined by one-sided confidence intervals) used to differentiate between normal and abnormal results and (2) the number of "failed" components (as determined by NCF) used to indicate an abnormal test result. Second, this study aimed to compare the accuracy between the ImPACT, CNS Vital Signs, and CCAT tests to provide clinical care guidance.

To achieve the aims of this study, the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were determined. These values determined the sensitivity and specificity in addition to other measurements discussed later. To clarify, these four values (i.e., TP, TN, FP, FN) were calculated for each model, each confidence interval, and each NCF value. For example, consider ImPACT Normative with a 75% confidence interval and $NCF = 2$. The change scores representing significant change are -5 , -6 , -2.1 , and 0.04 for verbal memory, visual memory, visual motor speed, and reaction time, respectively [6]. If a patient in the baseline/24–48 h group had two or more change scores exceeding these values, then the patient was classified as concussed and represented a TP. Otherwise, the baseline/24–48 h patient was classified as not concussed and represented a FN. If a patient in the baseline/baseline group had two or more change scores exceeding these values, then the patient was classified as concussed and represented a FP. Otherwise, the baseline/baseline patient was classified as not concussed and represented a TN.

When determining the confidence interval and NCF that maximized a neurocognitive test's performance, the objective of this study was to maximize the 2:1 weighted sum of sensitivity and specificity (i.e., $\frac{2}{3} \times \text{Sensitivity} + \frac{1}{3} \times \text{Specificity}$) while having a sensitivity and specificity of at least 0.5. This study maximized the 2:1 weighted sum of sensitivity and specificity because this study focused on the 24–48 h time point which is a time where medical professionals emphasize sensitivity over specificity. The 2:1 weight is consistent with other concussion literature such as Broglio et al. [6]. However, the study required both sensitivity and specificity to be at least 0.5 because if either of these values are less than 0.5 then random decision making would be more accurate than test administration. Overall, this study wanted to provide medical professionals with a comprehensive view of the test performance.

To provide additional measurements to characterize and compare each model's accuracy, the positive predictive value (PPV) [28], negative predictive value (NPV) [28], and F1 score [2, 33] were also computed. PPV (i.e., $TP/[TP + FP]$) and NPV (i.e., $TN/[TN + FN]$) are included to complement the sensitivity and specificity results. PPV and NPV depend on the prevalence of concussed patients in each model, so the reader is cautioned when comparing the PPV and NPV values between test types (e.g., comparing CNS Vital Signs to CCAT). The F1 score balances sensitivity and the PPV.

Further, the F1 scores are presented, but not used in the analysis since the F1 score fails to have a good intuitive explanation [2]. Finally, to assess the overall performance of each test type based on NCF, receiver operating characteristic (ROC) curves were plotted and the area under the ROC curve (AUC) were calculated using a linear approximation. AUC represents how well the model performs in terms of sensitivity and specificity. A higher AUC corresponds to overall higher sensitivity and specificity results. This study classifies the AUC values as follows: 0.5–0.59 (bad), 0.6–0.69 (poor), 0.7–0.79 (fair), 0.8–0.89 (good), 0.9–1.0 (excellent).

Finally, a bootstrap-based hypothesis test for paired samples was used to compare the AUC between two NCF variations of a test type. A bootstrap-based hypothesis test for unpaired samples was used to compare the AUC between two test types. The null hypothesis in this test is that both variations have an equal AUC. We rejected this null hypothesis at a significance level of $\alpha = 0.05$. This hypothesis testing was performed using the pROC package in the software R [49].

2.5 Change Scores vs. Raw Scores

The objective of this study was to determine the sensitivity and specificity of common neurocognitive tests and in particular, this study aimed to achieve this through change scores. However, there has been much discussion in the concussion literature regarding the need for baseline testing. Some argue baseline testing is significant [50] whereas others indicate there is a small or negligible benefit in using athlete-specific baseline values over normative values [13, 17, 18, 21, 22, 40, 52]. For thoroughness, this study also analyzed the three neurocognitive tests (i.e., ImPACT, CNS Vital Signs, and CCAT) with raw scores to determine the significance of baseline testing.

In particular, the Year 1 normative mean and standard deviation raw scores published in Ref. [6] were used to construct the 75, 87.5, 90, 92.5, 97.5, and 99 one-sided confidence intervals for each test component of interest. For clarification, it was appropriate to construct these confidence intervals with the normality assumption because all values published in Ref. [6] had sufficiently large sample sizes (> 50) supporting the application of the Central Limit Theorem [54]. Then, the 24–48 h raw scores from the baseline/24–48 h group were used to calculate sensitivity and the first baseline raw scores from the baseline/baseline group were used to calculate specificity. The first baseline raw scores from the baseline/baseline group were chosen to analyze specificity over the baseline raw scores from the baseline/24–48 h group to keep consistency with the change score sample sizes. Similar to the change score method, the sensitivity, specificity, and other diagnostic measurements

were calculated for each test type, each confidence interval, and each NCF. Also, the best-performing model for each test type (i.e., confidence interval and NCF combination) was determined by maximizing the 2:1 weighted sum of sensitivity and specificity while having a sensitivity and specificity of at least 0.5. Finally, a bootstrap-based hypothesis test for paired samples was used to compare the AUC between the change score best-performing model and raw score best-performing model for each test type.

3 Results

Demographic information of the study cohort is presented in Table 1. For the baseline/24–48 h group, the mean \pm standard deviation (SD) time between tests was 143.01 ± 93.75 days (ImPACT), 148.56 ± 131.26 days (CNS Vital Signs), and 121.71 ± 80.08 days (CCAT). For the baseline/baseline group, the mean \pm SD time between tests was 347.84 ± 73.63 days (ImPACT), 321.95 ± 85.03 days (CNS Vital Signs), and 357.39 ± 49.86 days (CCAT).

The ImPACT baseline/24–48 h and baseline/baseline groups differed in gender ($p < 0.001$), race ($p < 0.001$), and number of previous concussions ($p < 0.001$), but not age ($p = 0.26$). The CNS Vital Signs baseline/24–48 h and baseline/baseline groups showed differences in age ($p < 0.01$), race ($p < 0.05$), and number of previous concussions ($p < 0.05$), but no difference in gender ($p = 0.23$). The CCAT baseline/24–48 h and baseline/baseline groups differed in age ($p < 0.01$) and number of previous concussions ($p < 0.001$), but not gender ($p = 0.34$) nor race ($p = 0.67$). Additional hypothesis test results for the demographic information including the directionality of significant differences can be found in Table 1. For clarification, these demographic differences were not accounted for in the analysis because the aim of this study was to analyze the overall performance of patients. Change score statistics for the ImPACT, CNS Vital Signs, and CCAT tests for the baseline/24–48 h and baseline/baseline groups are displayed in Table 2, along with the percentage of study participants who improved, declined, or did not change with respect to a specific test type and component.

3.1 Sensitivity, Specificity, PPV, NPV, and F1 Score

The sensitivity, specificity, PPV, NPV, and F1 score were calculated for each model (ImPACT Normative, CNS Vital Signs Normative, CCAT Normative, and ImPACT RCI), for each one-sided confidence interval (75, 87.5, 90, 92.5, 95, 97.5, and 99), and for each NCF (4 for ImPACT/CCAT and 11 for CNS Vital Signs). Further, the 2:1 weighted sum of sensitivity and specificity was calculated to determine the best-performing model (i.e., maximum 2:1 weighted sum

Table 1 Characteristics of study data by test group

Group	<i>n</i>	Age	Height	Weight	Prev. concussion	Percent female	Percent male	Percent white	Percent other race
ImPACT baseline/24–48 h	1263	19.2 (1.4)	69.5 (4.4)	<u>171.6 (41.7)*</u>	<u>0.5 (0.8)***</u>	<u>43.2***</u>	56.8***	69.2***	<u>30.8***</u>
ImPACT baseline/baseline	7421	19.1 (1.2)	69.6 (3.9)	166.2 (33.0)*	0.3 (0.6)***	34.7***	<u>65.3***</u>	<u>75.2***</u>	24.8***
CNS baseline/24–48 h	75	19.0 (1.2)**	71.0 (4.9)	<u>180.9 (40.5)*</u>	<u>0.5 (0.7)*</u>	38.7	61.3	68.0*	<u>32.0*</u>
CNS baseline/baseline	344	<u>19.3 (1.1)**</u>	70.1 (4.5)	168.9 (40.0)*	0.3 (0.6)*	47.1	52.9	<u>75.3*</u>	24.7*
CCAT baseline/24–48 h	76	<u>19.3 (1.3)**</u>	69.5 (3.7)	167.7 (32.6)	<u>0.9 (1.2)***</u>	53.9	46.1	89.5	3.9
CCAT baseline/baseline	540	18.8 (1.0)**	69.9 (4.2)	166.6 (37.0)	0.4 (0.8)***	47.4	52.6	81.3	10.0

Values in table reported: mean (SD) where appropriate

n, number of individuals

SD, standard deviation

Prev. Concussion, number of previous concussions

Mann-Whitney *U* test used for continuous and ordinal variables

Chi-Squared test used for categorical variables

****p* < 0.001 when comparing baseline/24–48 h to baseline/baseline

***p* < 0.01 when comparing baseline/24–48 h to baseline/baseline **p* < 0.05 when comparing baseline/24–48 h to baseline/baseline

When significant differences between groups exist (denoted by asterisk[s]), underlined values denote the group that is significantly greater in magnitude or proportion (e.g., the ImPACT baseline/24–48 h group has significantly more previous concussions than the ImPACT baseline/baseline group)

of sensitivity and specificity while having a sensitivity and specificity of at least 0.5).

ImPACT performed best when using the Normative method with an 87.5%-confidence interval and 1 NCF (sensitivity = 0.583, specificity = 0.625, *F1* = 0.308). CNS Vital Signs performed best when using the Normative method with a 90%-confidence interval and 1 NCF (sensitivity = 0.587, specificity = 0.532, *F1* = 0.314). CCAT performed best using the Normative method when using a 75%-confidence interval and 2 NCF (sensitivity = 0.513, specificity = 0.715, *F1* = 0.290). Finally, the ImPACT RCI method performed best with an 87.5%-confidence interval and 1 NCF (sensitivity = 0.626, specificity = 0.559, *F1* = 0.297). Table 3 provides a collective summary of the best-performing models. Also, results for all sensitivity, specificity, PPV, NPV, and *F1* for ImPACT Normative, CNS Vital Signs Normative, CCAT Normative, and ImPACT RCI models can be found in ESM1, ESM2, ESM3, and ESM4 of the Online Resources, respectively. This study defined the best-performing model as the one that maximized the 2:1 weighted sum of sensitivity and specificity while having a sensitivity and specificity of at least 0.5. To this end, we present the full spectrum of performance measures across all neurocognitive tests in the Online Resources so medical professionals can choose the performance measures which best suit their needs. Also, this study focused on the 24–48 h

time point, but future work can consider which measurement values should be considered at each time point in the concussion recovery process.

3.2 ROC Curves and AUC

The ROC Curves for each model can be found in Fig. 1. Specifically, the 1 NCF curve for CNS Vital Signs Normative dominates all other CNS Vital Signs Normative curves (AUC = 0.610). Similarly, the 2 NCF curve for CCAT Normative dominates all other CCAT Normative curves (AUC = 0.640). For the other two models, ImPACT Normative and ImPACT RCI, there are trade-offs between the NCF = 1 and NCF = 2 curves dependent on the confidence interval employed. Overall, ImPACT Normative performed the best with 2 NCF (AUC = 0.638) and ImPACT RCI performed the best with 2 NCF (AUC = 0.632). It should be noted the best-performing ImPACT NCF in terms of AUC is not consistent with the best-performing confidence interval and NCF defined earlier for neither Normative nor RCI. However, AUC equally weights sensitivity and specificity whereas the best-performing model gives more weight to sensitivity. When comparing the best-performing models to one another, the results support similar AUC values (*p* > 0.05 for all using bootstrap test). Details regarding this

Table 2 Characteristics of individuals who improved, declined, or no change for different groups and test types

Group	Test type	Component	Performance of all patients	% improved	% declined	% no change	Performance of improved patients	Improved min	Improved max	Performance of declined patients	Declined Min	Declined Max
Baseline/24–48 h	ImPACT Test	Verbal memory	- 3 (13)	40	53	7	8 (7)	1	38	- 12 (10)	- 63	- 1
Baseline/24–48 h	ImPACT Test	Visual memory	- 5 (15)	34	62	4	10 (8)	1	51	- 14 (11)	- 58	- 1
Baseline/24–48 h	ImPACT Test	Visual motor speed	- 1.57 (6.58)	42.52	57.12	0.40	3.97 (3.12)	0.02	15.90	- 5.70 (5.35)	- 30.20	- 0.02
Baseline/24–48 h	ImPACT Test	Reaction time	0.04 (0.14)	35.47	58.99	5.54	- 0.07 (0.07)	- 0.58	- 0.01	0.11 (0.13)	0.01	1
Baseline/Baseline	ImPACT Test	Verbal memory	1 (10)	47	44	9	8 (7)	1	42	- 8 (7)	- 45	- 1
Baseline/Baseline	ImPACT Test	Visual memory	1 (12)	52	43	5	10 (7)	1	51	- 9 (7)	- 50	- 1
Baseline/Baseline	ImPACT Test	Visual motor speed	0.75 (4.62)	57.81	41.84	0.35	3.76 (3.01)	0.02	30.40	- 3.40 (2.97)	- 31.07	- 0.02
Baseline/Baseline	ImPACT Test	Reaction time	0.00 (0.08)	46.87	46.33	6.81	- 0.06 (0.06)	- 0.97	- 0.01	0.06 (0.06)	0.01	0.88
Baseline/24–48 h	CNS Test	Composite memory	- 3 (19)	40	52	8	15 (8)	2	31	- 17 (15)	- 65	- 1
Baseline/24–48 h	CNS Test	Verbal memory	- 4 (28)	44	49	7	20 (14)	3	64	- 26 (20)	- 86	- 3
Baseline/24–48 h	CNS Test	Visual memory	- 1 (14)	40	55	5	13 (9)	3	32	- 11 (9)	- 31	- 1
Baseline/24–48 h	CNS Test	Psychomotor speed	- 5 (16)	43	51	7	7 (6)	1	23	- 16 (14)	- 63	- 2
Baseline/24–48 h	CNS Test	Reaction time	- 6 (27)	44	51	5	12 (11)	1	47	- 23 (27)	- 126	- 1
Baseline/24–48 h	CNS Test	Complex attention	- 14 (74)	47	49	4	18 (17)	3	91	- 45 (95)	- 551	- 3
Baseline/24–48 h	CNS Test	Cognitive flexibility	- 2 (22)	60	37	3	11 (10)	1	46	- 23 (20)	- 86	- 1
Baseline/24–48 h	CNS Test	Processing speed	- 1 (15)	41	52	7	13 (9)	1	33	- 12 (10)	- 40	- 2
Baseline/24–48 h	CNS Test	Executive function	- 1 (21)	57	41	1	12 (9)	1	37	- 20 (19)	- 84	- 1
Baseline/24–48 h	CNS Test	Simple attention	- 8 (42)	33	47	20	37 (23)	3	78	- 45 (23)	- 78	- 5
Baseline/24–48 h	CNS Test	Motor speed	- 7 (19)	33	63	4	7 (7)	1	31	- 15 (19)	- 105	- 1

Table 2 (continued)

Group	Test type	Component	Performance of all patients	% improved	% declined	% no change	Performance of improved patients	Improved min	Improved max	Performance of declined patients	Declined Min	Declined Max
Baseline/Base-line	CNS Test	Composite memory	4 (61)	49	45	6	20 (84)	1	1102	- 13 (9)	- 43	- 1
Baseline/Base-line	CNS Test	Verbal memory	1 (19)	49	43	8	16 (13)	1	76	- 16 (11)	- 54	- 2
Baseline/Base-line	CNS Test	Visual memory	3 (57)	48	45	7	19 (79)	1	1020	- 14 (11)	- 53	- 1
Baseline/Base-line	CNS Test	Psychomotor speed	0 (10)	49	43	8	8 (6)	1	36	- 8 (6)	- 32	- 1
Baseline/Base-line	CNS Test	Reaction time	3 (55)	48	49	3	17 (77)	1	988	- 10 (8)	- 38	- 1
Baseline/Base-line	CNS Test	Complex attention	2 (40)	45	45	10	16 (56)	1	696	- 11 (8)	- 51	- 1
Baseline/Base-line	CNS Test	Cognitive flexibility	2 (12)	53	41	6	11 (7)	1	32	- 9 (7)	- 58	- 1
Baseline/Base-line	CNS Test	Processing speed	2 (13)	56	40	4	10 (8)	1	44	- 10 (8)	- 41	- 1
Baseline/Base-line	CNS Test	Executive function	2 (11)	53	42	4	11 (7)	1	33	- 8 (6)	- 25	- 1
Baseline/Base-line	CNS Test	Simple attention	- 4 (34)	29	44	27	36 (17)	3	78	- 34 (20)	- 78	- 3
Baseline/Base-line	CNS Test	Motor speed	- 1 (11)	39	53	8	8 (11)	1	91	- 7 (8)	- 66	- 1
Baseline/Base-line/24-48 h	AXON Test	Processing speed	- 3.1 (10.2)	42.1	56.6	1.3	3.8 (3.9)	0.2	20.8	- 8.3 (10.6)	- 49.8	- 0.1
Baseline/Base-line/24-48 h	AXON Test	Composite attention	- 4.0 (9.4)	32.9	67.1	0.0	3.2 (3.0)	0.2	11.6	- 7.5 (9.5)	- 49.2	- 0.4
Baseline/Base-line/24-48 h	AXON Test	Composite learning	2.3 (9.3)	60.5	34.2	5.3	7.9 (6.4)	0.1	25.7	- 7.1 (5.9)	- 24.0	- 0.1
Baseline/Base-line/24-48 h	AXON Test	Working Memory speed	- 1.4 (7.3)	36.8	63.2	0.0	5.2 (4.3)	0.1	14.1	- 5.2 (5.7)	- 26.1	- 0.1
Baseline/base-line	AXON Test	Processing speed	0.4 (5.4)	55.6	43.1	1.3	3.9 (3.5)	0.1	27.7	- 4.2 (3.9)	- 26.0	- 0.1
Baseline/base-line	AXON Test	Composite attention	0.0 (4.5)	48.1	51.1	0.7	3.5 (3.1)	0.1	23.2	- 3.3 (2.8)	- 16.6	- 0.1
Baseline/base-line	AXON Test	Composite learning	1.2 (9.1)	58.9	40.0	1.1	8.1 (5.8)	0.1	27.8	- 6.6 (5.4)	- 25.8	- 0.1

Table 2 (continued)

Group	Test type	Component	Performance of all patients	% improved	% declined	% no change	Performance of improved patients	Improved min	Improved max	Performance of declined patients	Declined Min	Declined Max
Baseline/base-line	AXON Test	Working memory speed	1.2 (5.6)	58.9	40.4	0.7	4.7 (3.7)	0.1	20.7	-4.0 (3.3)	-15.2	-0.1

Change score values in table reported: mean (SD) where appropriate

Number of significant figures in table correspond to numerical values in original dataset
SD standard deviation, *Min* minimum value, *Max* maximum value

AUC hypothesis test analysis can be found in ESM8 of the Online Resources.

3.3 Raw Score Analysis

Table 4 displays a side-by-side comparison of the best-performing change score and raw score models for all three test types. Recall, this study defined the best-performing model as the one that maximized the 2:1 weighted sum of sensitivity and specificity while having a sensitivity and specificity of at least 0.5. First, the ImPACT change score models (Normative and RCI method) outperformed the ImPACT raw score model when considering the 2:1 weighted sum. Further, the bootstrap-based hypothesis test showed the AUC values between the change score model and raw score model for both the Normative and RCI method are significantly different ($p < 0.05$). In particular, a one-sided bootstrap-based hypothesis test showed the ImPACT change score best-performing model has a significantly higher AUC for both the Normative and RCI method. Second, the CCAT change score model outperformed the CCAT raw score model in terms of the 2:1 weighted sum of sensitivity and specificity, but the bootstrap-based hypothesis test showed similar AUC values ($p > 0.05$). Third, the CNS Vital Signs raw score model outperformed the change score model in terms of the 2:1 weighted sum, but the bootstrap-based hypothesis test showed similar AUC values ($p > 0.05$). All values for the ImPACT, CNS Vital Signs, and CCAT raw score analysis can be found in ESM5, ESM6, and ESM7 of the Online Resources, respectively. Further, details regarding the AUC hypothesis test can be found in ESM8 of the Online Resources.

4 Discussion

Sport-related concussion (SRC) is an ever-increasing public health concern and accurate assessment of neurocognitive functioning has long been included as part of the post-concussion multi-faceted assessment. However, it remains unknown which commonly implemented computer-based neurocognitive tests are optimal for this aspect of injury management. Thus, the current study evaluated the sensitivity and specificity of three computer-based neurocognitive assessments in a large and diverse sample to provide athletic trainers, physicians, neuropsychologists, and other health-care providers. guidance on their clinical utility.

For ImPACT, change score performance was best (sensitivity = 0.583, specificity = 0.625, $F1 = 0.308$) with an 87.5%-confidence interval and when participants failed at least one neurocognitive test component (NCF = 1) using previously developed normative data [6]. The ImPACT RCI method performed best with an 87.5%-confidence

Table 3 Best-performing models for change score analysis

Test	Confidence interval	NCF	Sensitivity	Specificity	1-Specificity	2:1 Sensitivity + specificity	PPV	NPV	F1
ImPACT normative	87.5	1	0.583	0.625	0.375	0.597	0.209	0.898	0.308
CNS normative	90	1	0.587	0.532	0.468	0.568	0.215	0.855	0.314
CCAT normative	75	2	0.513	0.715	0.285	0.580	0.202	0.913	0.290
ImPACT RCI	87.5	1	0.626	0.559	0.440	0.604	0.195	0.898	0.297

NCF number of components failed, *PPV* positive predictive value, *NPV* negative predictive value

interval and $NCF=1$ (sensitivity = 0.626, specificity = 0.559, $F1=0.297$). These results are generally consistent with previous research [4, 39, 47] and the embedded ImPACT algorithms employing 80% two-sided confidence intervals (90% one-sided) [27]. Also, neurocognitive testing is used to provide increased sensitivity to detect deficits not apparent on routine clinical examinations, but the low sensitivity results suggest their ability to do so is very poor.

The change score performance of the CCAT and CNS Vital signs were similar to ImPACT. CCAT's best-performing model had a sensitivity (0.513) and specificity (0.715) exceeding chance with a 75%-confidence interval when participants failed two neurocognitive test components ($NCF=2$). CNS Vital Signs performed best (sensitivity = 0.587, specificity = 0.532, $F1=0.314$) using normative data with a 90%-confidence interval and when participants failed one assessment component ($NCF=1$).

As previously mentioned, there has been much discussion regarding baseline testing in the concussion literature. The findings from the raw score analysis showed ImPACT and CCAT change score models performed better than raw score models when looking at the 2:1 weighted sum of sensitivity and specificity. Further, ImPACT showed the change score models (for both Normative and RCI) have significantly higher AUC values, but CCAT showed the change score model and raw score model have similar AUC values. The CNS Vital Signs raw score model performed better than the change score model when looking at the 2:1 measurement, but they exhibited similar AUC values. The CCAT and CNS Vital Signs results support baseline testing has a negligible impact on neurocognitive test performance. The ImPACT results support baseline testing aids neurocognitive test performance but considering the difference in magnitude between the 2:1 weighted sum for the change score and raw score models (i.e., 0.01 for Normative and 0.017 for RCI), the impact is small. Overall, these results support current literature which states there is a small or negligible benefit in using athlete-specific baseline values over normative values [13, 17, 18, 21, 22, 40, 52].

Extant research comparing the sensitivity and specificity of traditional pencil-and-paper neurocognitive tests and computer-based tests in accurately classifying patients

with SRC and non-injured patients report mixed findings. For instance, a prior study showed that a battery of neuropsychological tests including measures of verbal learning and memory, processing speed, executive functioning, and working memory demonstrated 87.5% sensitivity and 90% specificity [15]. Yet, Randolph and colleagues [43] argue that these and other studies providing evidence that traditional neurocognitive tests are sensitive to the effects of SRC suffer from methodologic flaws that limit their comparability and generalizability. Still other research reported that computer-based neurocognitive tests do not fare any better with respect to sensitivity and specificity than traditional neurocognitive tests [42]. Along these lines, Resch et al. [48] summarized extant research on the sensitivity (79.2–94.6%) and specificity (89.4–97.3%) of ImPACT and the reported sensitivity (70.8%) of CogState. Houck et al. [26] recently reported baseline to baseline testing in non-concussed athletes commonly shows failure on one testing component. When considered together, there is little evidence suggestive that one neurocognitive test measure is superior or better than another, leaving such decision making in the hands of the medical provider.

When considering all of the approaches to test accuracy, no test or interpretative approach evaluated here appeared substantially better than the other, suggesting equivalence between the measures. However, the overall low sensitivity and specificity estimates solidifies the clinical examination as the gold standard for concussion diagnosis, supported by a multi-dimensional objective assessment protocol. In most instances, this will include a symptom evaluation, postural control assessment, neuropsychological status, and other functional assessments. Indeed, consensus statements support the use of neurocognitive tests [36, 37] and other studies support that neurocognitive tests, when included in a battery, increase the clinical utility over symptoms alone [5, 13, 17, 21]. With this, we argue that computer-based testing should not be abandoned, but rather be used in a multi-dimensional assessment protocol or at the discretion of the appropriate clinician when circumstances dictate (e.g., when athletes are slow to recover).

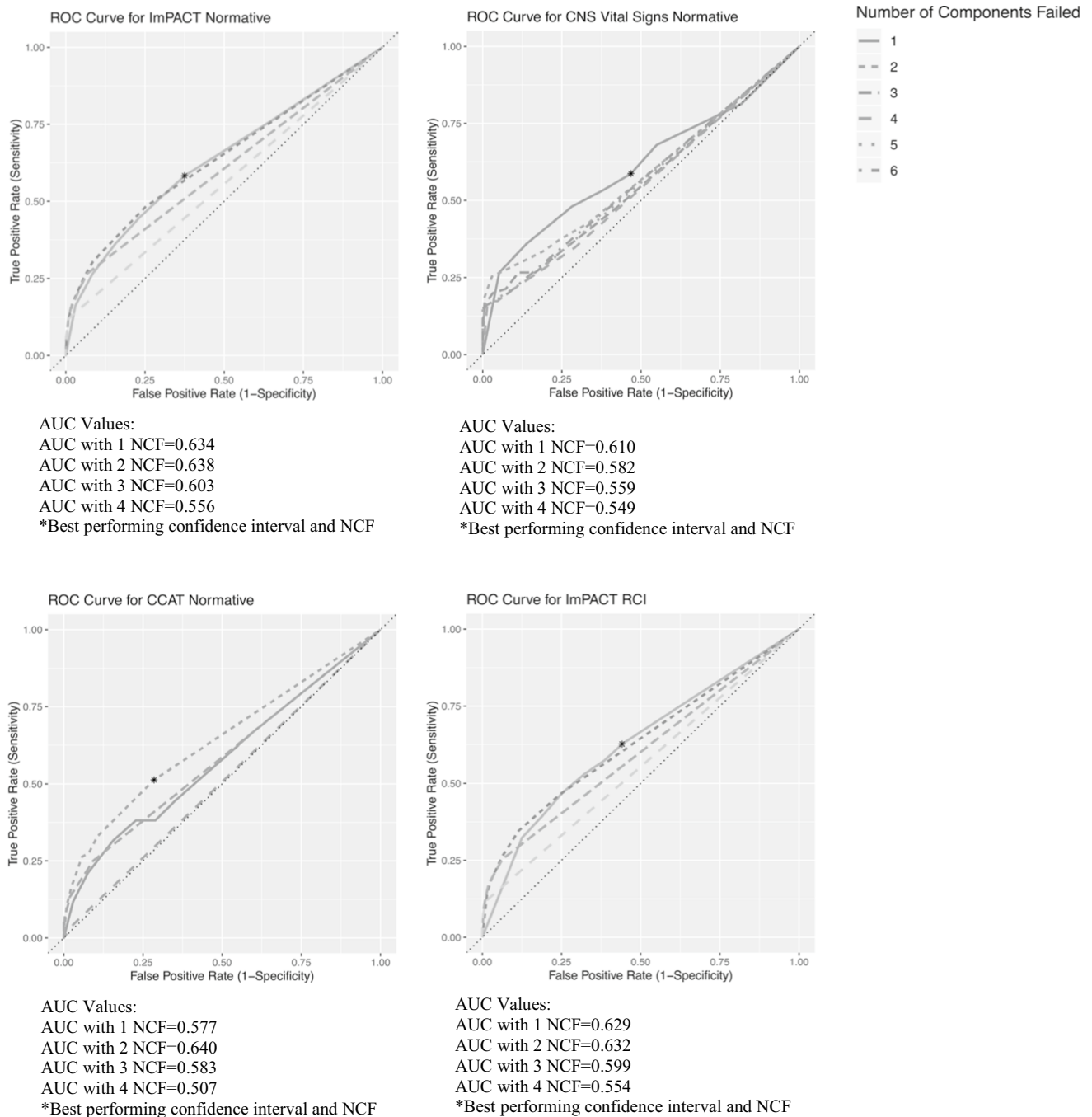


Fig. 1 Each neurocognitive test's receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) for change score models

4.1 Limitations

The current study should be considered in light of several limitations. First, the number of sustained concussions of each participant prior to their participation in the current study was uncontrolled for in analyses. As has been reported previously [16, 24], multiple concussions have been associated with prolonged symptoms, recovery time, and risk

for future concussions and may have impacted participants' performance on neurocognitive assessment. Future research incorporating the number of previously sustained SRC in analyses would help elucidate these potential neurocognitive performance differences. Second, gender, race, and socioeconomic status differences in pre- and post-SRC neurocognitive performance in addition to demographic differences between the baseline/baseline and baseline/24–48 h groups

Table 4 Side-by-side comparison of best-performing models for change score and raw score analysis

Test	Change score					Raw score										
	CI	NCF	Sensitivity	Specificity	2:1 S&S	PPV	NPV	F1	CI	NCF	Sensitivity	Specificity	2:1 S&S	PPV	NPV	F1
ImPACT normative	87.5	1	0.583	0.625	0.597	0.209	0.898	0.308	75	2	0.625	0.513	0.587	0.179	0.889	0.278
CNS normative	90	1	0.587	0.532	0.568	0.215	0.855	0.314	87.5	5	0.653	0.526	0.611	0.231	0.874	0.341
CCAT normative	75	2	0.513	0.715	0.580	0.202	0.913	0.290	75	3	0.500	0.672	0.557	0.177	0.905	0.261
ImPACT RCI	87.5	1	0.626	0.559	0.604	0.195	0.898	0.297	75	2	0.625	0.513	0.587	0.179	0.889	0.278

Italics represents model with greatest 2:1 S&S when comparing change score and raw score models

CI confidence interval, NCF number of components failed, 2:1 S&S 2:1 weighted sum of sensitivity and specificity, PPV positive predictive value, NPV negative predictive value

were not controlled for in the analyses. Third, we evaluated cadets and athletes collectively, consistent with the CARE Consortium aims [8]. Fourth, this study analyzed the sensitivity and specificity of the embedded ImPACT algorithm (i.e., RCI 80% two-sided confidence interval [27]) but the study did not analyze the embedded algorithms of CNS Vital Signs nor CCAT. The embedded algorithms for these two tests are proprietary algorithms and the information was not captured by the CARE dataset. Fifth, when analyzing the embedded ImPACT algorithm, the RCI values published in the *ImPACT Administration and Interpretation Manual* [27] were employed because using the manual is the standard. However, future research can analyze how different RCI measurements [25] and more recent RCI calculations [10] impact the sensitivity and specificity results. Sixth, athletes and cadets with invalid baseline tests were removed from the analysis, yet Table 2 demonstrates that approximately 30–60% of study participants performed better on the second test regardless of concussion status. Such improvement from one test administration to another suggests that multiple factors, including effort, motivation, and physical and mental fatigue, affect test performance and warrant consideration when determining the validity of neurocognitive testing results. Thus, future research that describes methods that more accurately account for these various performance factors to identify invalid baseline tests to improve the diagnostic utility of computer-based neurocognitive tests is needed. Finally, the measures that comprise the computer-based neurocognitive tests utilized in the current study are not equivalent to the original paper and pencil measures from which they were derived. Typically, traditional paper and pencil neuropsychological tests were designed to evaluate gross changes in neurocognitive functioning, not the subtle deficits associated with SRC. Future research that takes a more granular, task-level approach, rather than the component level approach used here, would assist in identifying those measures that exhibit better or worse sensitivity and specificity in SRC assessment.

5 Conclusion

This investigation sought to examine the sensitivity and specificity of commonly used computer-based neurocognitive tests in SRC management to provide relevant clinicians additional guidance for appropriate patient care. Our findings indicate that no assessment or interpretative approach is substantially better than the other. Also, the overall low sensitivity and specificity results provide additional evidence for contemporary multi-dimensional concussion assessment approaches and indicate the need for improved sensitivity of neurocognitive assessment tools used in concussion assessment.

Acknowledgements Alphabetically by last name, the CARE Consortium investigators include: Scott A. Anderson, ATC (University of Oklahoma, USA); Holly J. Benjamin, MD, FAAP, FACSM (University of Chicago, USA); Alison Brooks, MD, MPH (University of Wisconsin-Madison, USA); Thomas Buckley, EdD, ATC (University of Delaware, USA); Kenneth L. Cameron, PhD, MPH, ATC, FNATA (United States Military Academy, USA); Sara PD Chrisman, MD, MPH (University of Washington, USA); James R. Clugston, MD, MS, CAQSM (University of Florida, USA); Stefan M. Duma, PhD (Virginia Tech, USA); James Eckner, MD, MS (University of Michigan, USA); Luis A. Feigenbaum, PT, DPT, ATC/L (University of Miami, USA); Joshua T. Goldman, MD, MBA (University of California, Los Angeles, USA); Joseph B. Hazzard Jr., ATC (Bloomsburg University, USA); Megan N. Houston, PhD, ATC (United States Military Academy West Point, USA); April Hoy, MS, ATC, CSCS (Azusa Pacific University, USA); Thomas W. Kaminski, PhD, ATC, FNAK, FNATA, FACSM, RFSA (University of Delaware, USA); Louise A. Kelly, PhD (California Lutheran University, USA); Anthony P. Kontos, PhD (University of Pittsburgh, USA); Laura Lintner, DO (Wake Forest University School of Medicine, USA); Christina L. Master, MD, FAAP, CAQSM, FACSM (University of Pennsylvania, USA); Jane McDevitt, PhD, LAT, ATC, CSCS (Temple University, USA); Gerald McGinty, PT, DPT (United States Air Force Academy, USA); Jessica Dysart Miles, LAT, ATC (University of North Georgia, USA); Chris Miles, MD (Wake Forest University, USA); Justus Ortega, PhD (Humboldt State University, USA); Nicholas Port, PhD (Indiana University, USA); Margot Putukian, MD, FACSM, FAMSSM (Princeton University, USA); Steve Rowson, PhD (Virginia Tech, USA); Julianne Schmidt, PhD, ATC (University of Georgia, USA); Joel D. Stitzel, PhD (Wake Forest University, USA); Adam James Susmarski, DO, CDR, MC, USN (United States Naval Academy, USA); Christopher T. Whitlow, MD, PhD, MHA (Wake Forest University, USA). We would also like to thank Jody Harland, Janetta Matesan, and Larry Riggen (Indiana University), Ashley Rettmann, Nicole L'Heureux (University of Michigan), Melissa Koschnitzke (Medical College of Wisconsin), Michael Jarrett, Vibeke Brinck, and Bianca Byrne (Quesgen), Thomas Dompier, Christy Collins, Melissa Niceley Baker, and Sara Dalton (Datalys Center for Sports Injury Research and Prevention), and the research and medical staff at each of the participating sites.

Declarations

Funding This publication was made possible, in part, by support from the Grand Alliance CARE Consortium, funded by the NCAA and the Department of Defense. The US Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick, MD 21702-5014, USA, is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Program under Award no. W81XWH-14-2-0151. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense (Defense Health Program funds).

Conflict of interest Lauren L. Czerniak, Spencer W. Liebel, Gian-Gabriel P. Garcia, and Mariel S. Lavieri have no competing interests directly relevant to the content of this study. Michael A. McCrea, Thomas W. McAllister, and Steven P. Broglio received research support from the NCAA and the U.S. Department of Defense.

Ethics approval This research study follows the standards of ethics stated in the Declaration of Helsinki.

Consent to participate All individuals provided written informed consent which was approved by the local institution and the US Army Human Research Protection Office.

Consent for publication CARE Consortium Review Board as well as the US Army Human Research Protection Office.

Availability of data and material The NCAA-DoD CARE dataset is available in the FITBIR database (<https://fitbir.nih.gov/>).

Code availability The R code is accessible by contacting the corresponding author.

Author contributions LLC performed all data analysis, created and formatted tables and figures, and was a major contributor in writing the manuscript. SWL was a major contributor in writing the manuscript. GPG contributed greatly in data acquisition, data analysis, and writing the manuscript. MSL contributed to the design of the statistical analysis and writing of the manuscript. MAM and TM helped with data collection, securing funding, and final approval of the manuscript. SPB helped with the study design, securing funding, data interpretation, drafting and final approval of the manuscript.

References

1. About the Cogstate CCAT. Retrieved September 8, 2019, from <http://axonsports.ca/index.cfm?pid=65>
2. Aoullay A. What's WRONG with Metrics? Retrieved February 2, 2020. (2018). <https://towardsdatascience.com/choosing-the-right-metric-is-a-huge-issue-99ccbe73de61>
3. Aubry M, Cantu R, Dvorak J, Graf-Baumann T, Johnston K, Kelly J, et al. Summary and agreement statement of the first international conference on concussion in sport, Vienna 2001. *Br J Sports Med.* 2002;36(1):6–10. <https://doi.org/10.1136/bjism.36.1.6>
4. Broglio SP, Ferrara MS, Macciocchi SN, Baumgartner TA, Elliott R. Test-retest reliability of computerized concussion assessment programs. *J Athl Train.* 2007;42(4):509–14.
5. Broglio SP, Harezlak J, Katz B, Zhao S, McAllister T, McCrea M, Investigators CC. Acute sport concussion assessment optimization: a prospective assessment from the CARE Consortium. *Am J Sports Med.* 2019;49(12):1977–87. <https://doi.org/10.1007/s40279-019-01155-0>.
6. Broglio SP, Katz BP, Zhao S, McCrea M, McAllister T, CARE Consortium Investigators. Test–retest reliability and interpretation of common concussion assessment tools: findings from the NCAA-DoD CARE consortium. *Am J Sports Med.* 2018;48(5):1255–68. <https://doi.org/10.1007/s40279-017-0813-0>.
7. Broglio SP, Macciocchi SN, Ferrara MS. Sensitivity of the concussion assessment battery. *J Neurosurg.* 2007;60(6):1050–8. <https://doi.org/10.1227/01.NEU.0000255479.90999.CO>.
8. Broglio SP, McCrea M, McAllister T, Harezlak J, Katz B, Hack D, 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. *Am J Sports Med.* 2017;47(7):1437–51. <https://doi.org/10.1007/s40279-017-0707-1>.
9. Brooks BL, Iverson GL, Sherman EMS, Roberge M. Identifying cognitive problems in children and adolescents with depression using computerized neuropsychological testing. *Appl Neuropsychol.* 2010;17(1):37–43. <https://doi.org/10.1080/09084280903526083>.
10. Bruce J, Echemendia R, Meeuwisse W, Comper P, Sisco A. 1 year test–retest reliability of ImPACT in professional ice hockey

- players. *Clin Neuropsychol.* 2014;28(1):14–25. <https://doi.org/10.1080/13854046.2013.866272>.
11. Carney N, Ghajar J, Jagoda A, Bedrick S, Davis-O'Reilly C, du Coudray H, Hack D, Helfand N, Huddleston A, Nettleton T, Riggio S. Concussion guidelines step 1: systematic review of prevalent indicators. *J Neurosurg.* 2014;75:S3–15. <https://doi.org/10.1227/NEU.0000000000000433>.
 12. Cassidy JD, Carroll LJ, Peloso PM, Borg J, von Holst H, Holm L, et al. Incidence, risk factors and prevention of mild traumatic brain injury: Results of the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *J Rehabil Med.* 2004;43(Suppl):28–60. <https://doi.org/10.1080/16501960410023732>.
 13. Chin EY, Nelson LD, Barr WB, McCrory P, McCrea MA. Reliability and validity of the sport concussion assessment tool-3 (SCAT3) in high school and collegiate athletes. *Am J Sports Med.* 2016;44(9):2276–85. <https://doi.org/10.1177/0363546516648141>.
 14. *CNS Vital Signs Interpretation Guide.* (2014). CNS Vital Signs.
 15. Collins MW, Grindel SH, Lovell MR, Dede DE, Moser DJ, Phalin BR, et al. Relationship between concussion and neuropsychological performance in college football players. *JAMA.* 1999;282(10):964–70. <https://doi.org/10.1001/jama.282.10.964>.
 16. Covassin T, Moran T, Wilhelm K. Concussion symptoms and neurocognitive performance of high school and college athletes who incur multiple concussions. *Am J Sports Med.* 2013;41(12):2885–9. <https://doi.org/10.1177/0363546513499230>.
 17. Downey RI, Hutchison MG, Comper P. Determining sensitivity and specificity of the Sport Concussion Assessment Tool 3 (SCAT3) components in university athletes. *Brain Inj.* 2018;32(11):1345–52. <https://doi.org/10.1080/02699052.2018.1484166>.
 18. Echemendia RJ, Bruce JM, Bailey CM, Sanders JF, Arnett P, Vargas G. The utility of post-concussion neuropsychological data in identifying cognitive change following sports-related MTBI in the absence of baseline data. *Clin Neuropsychol.* 2012;26(7):1077–91. <https://doi.org/10.1080/13854046.2012.721006>.
 19. Echemendia RJ, Iverson GL, McCrea M, Macciocchi SN, Gioia GA, Putukian M, Comper P. Advances in neuropsychological assessment of sport-related concussion. *Br J Sports Med.* 2013;47(5):294–8. <https://doi.org/10.1136/bjsports-2013-092186>.
 20. Elbin RJ, Kontos AP, Kegel N, Johnson E, Burkhart S, Schatz P. Individual and combined effects of LD and ADHD on computerized neurocognitive concussion test performance: evidence for separate norms. *Arch Clin Neuropsychol.* 2013;28(5):476–84. <https://doi.org/10.1093/arclin/ac024>.
 21. Garcia G-GP, Broglio SP, Lavieri MS, McCrea M, McAllister T, McAllister T. Quantifying the value of multidimensional assessment models for acute concussion: an analysis of data from the NCAA-DoD care consortium. *Am J Sports Med.* 2018;48(7):1739–49. <https://doi.org/10.1007/s40279-018-0880-x>.
 22. Garcia G-GP, Yang J, Lavieri MS, McAllister TW, McCrea MA, Broglio SP. Optimizing components of the sport concussion assessment tool for acute concussion assessment. *J Neurosurg.* 2020. <https://doi.org/10.1093/neuros/nyaa150>.
 23. Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Arch Clin Neuropsychol.* 2006;21:623–43. <https://doi.org/10.1016/j.acn.2006.05.007>.
 24. Guskiewicz KM, McCrea M, Marshall SW, Cantu RC, Randolph C, Barr W, et al. Cumulative effects associated with recurrent concussion in collegiate football players: The NCAA Concussion Study. *JAMA.* 2003;290(19):2549–55. <https://doi.org/10.1001/jama.290.19.2549>.
 25. Hinton-Bayre AD. Choice of reliable change model can alter decisions regarding neuropsychological impairment after sports-related concussion. *Clin J Sport Med.* 2012;22(2):105–8. <https://doi.org/10.1097/JSM.0b013e318248a526>.
 26. Houck ZM, Asken BM, Bauer RM, Kontos AP, McCrea MA, McAllister TW, Broglio SP, Clugston JR, Care Consortium Investigators. Multivariate base rates of low scores and reliable decline on ImPACT in healthy collegiate athletes using CARE consortium norms. *J Int Neuropsychol Soc.* 2019;25(09):961–71. <https://doi.org/10.1017/S1355617719000729>.
 27. *ImPACT Administration and Interpretation Manual.* ImPACT Applications, Inc. (2016).
 28. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning, vol. 103. New York : Springer; 2013. <https://doi.org/10.1007/978-1-4614-7138-7>.
 29. Langlois J, Rutland-Brown W, Wald M. The epidemiology and impact of traumatic brain injury: a brief overview. *J Head Trauma Rehabil.* 2006;21(5):375–8.
 30. Lempke LB, Schmidt JD, Lynall RC. Athletic trainers' concussion-assessment and concussion-management practices: an update. *J Athl Train.* 2020;55(1):17–26. <https://doi.org/10.4085/1062-6050-322-18>.
 31. Louey AG, Cromer JA, Schembri AJ, Darby DG, Maruff P, Maddisi M, McCrory P. Detecting cognitive impairment after concussion: sensitivity of change from baseline and normative data methods using the CogSport/Axon Cognitive Test Battery. *Arch Clin Neuropsychol.* 2014;29(5):432–41. <https://doi.org/10.1093/arclin/acu020>.
 32. Lovell MR, Collins MW, Podell K, Powell J, Maroon J. ImPACT: Immediate post-concussion assessment and cognitive testing. Pittsburgh: NeuroHealth Systems LLC; 2000.
 33. Lucini FR, Fogliatto FS, da Silveira GJC, Neyeloff JL, Anzanello MJ, Kuchenbecker RS, Schaan BD. Text mining approach to predict hospital admissions using early medical records from the emergency department. *Int J Med Inform.* 2017;100:1–8. <https://doi.org/10.1016/j.ijmedinf.2017.01.001>.
 34. Maruff P, Thomas E, Cysique L, Brew B, Collie A, Snyder P, Pietrzak RH. Validity of the CogState Brief Battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. *Arch Clin Neuropsychol.* 2009;24(2):165–78. <https://doi.org/10.1093/arclin/acp010>.
 35. McAllister T, McCrea M. Long-term cognitive and neuropsychiatric consequences of repetitive concussion and head-impact exposure. *J Athl Train.* 2017;52(3):309–17. <https://doi.org/10.4085/1062-6050-52.1.14>.
 36. McCrory P, Meeuwisse WH, Aubry M, Cantu RC, Dvorak J, Echemendia RJ, et al. Consensus statement on concussion in sport—the 4th international conference on concussion in sport held in Zurich, November 2012. *PM&R.* 2013;5(4):255–79. <https://doi.org/10.1016/j.pmrj.2013.02.012>.
 37. McCrory P, Meeuwisse W, Dvorak J, Aubry M, Bailes J, Broglio S, Cantu RC, Cassidy D, Echemendia RJ, Castellani RJ, Davis GA, Ellenbogen R, Emery C, Engebretsen L, Feddermann-Demont N, Giza CC, Guskiewicz KM, Herring S, Iverson GL, et al. Consensus statement on concussion in sport—The 5th international conference on concussion in sport held in Berlin, October 2016. *Br J Sports Med.* 2017. <https://doi.org/10.1136/bjsports-2017-097699>.
 38. Meehan WP, d'Hemecourt P, Collins CL, Taylor AM, Comstock RD. Computerized neurocognitive testing for the management of sport-related concussions. *J Pediatr.* 2012;129(1):38–44. <https://doi.org/10.1542/peds.2011-1972>.
 39. Nelson LD, LaRoche AA, Pfaller AY, Lerner EB, Hammeke TA, Randolph C, et al. Prospective, head-to-head study of three computerized neurocognitive assessment tools (CNTs): reliability and validity for the assessment of sport-related concussion. *J Int*

- Neuropsychol Soc. 2016;22:24–37. <https://doi.org/10.1017/S1355617715001101>.
40. Putukian M, Echemendia R, Dettwiler-Danspeckgruber A, Duliba T, Bruce J, Furtado JL, Murugavel M. Prospective clinical assessment using sideline concussion assessment tool-2 testing in the evaluation of sport-related concussion in college athletes. *Clin J Sport Med*. 2015;25(1):7.
 41. Rabinowitz AR, Levin HS. Cognitive sequelae of traumatic brain injury. *Psychiatr Clin N Am*. 2014;37(1):1–11. <https://doi.org/10.1016/j.psc.2013.11.004>.
 42. Rahman-Filipiak AAM, Woodard JL. Administration and environment considerations in computer-based sports-concussion assessment. *Neuropsychol Rev*. 2013;23(4):314–34. <https://doi.org/10.1007/s11065-013-9241-6>.
 43. Randolph C, McCrea M, Barr WB. Is neuropsychological testing useful in the management of sport-related concussion? *J Athl Train*. 2005;40(3):139–52.
 44. Randolph C, Millis S, Barr WB, McCrea M, Guskiewicz KM, Hammeke TA, Kelly JP. Concussion Symptom Inventory: an empirically derived scale for monitoring resolution of symptoms following sport-related concussion. *Arch Clin Neuropsychol*. 2009;24(3):219–29. <https://doi.org/10.1093/arclin/acp025>.
 45. Reeves DL, Winter KP, Bleiberg J, Kane RL. ANAM[®] genogram: historical perspectives, description, and current endeavors. *Arch Clin Neuropsychol*. 2007;22(1):15–37. <https://doi.org/10.1016/j.acn.2006.10.013>.
 46. Register-Mihalik JK, Guskiewicz KM, Mihalik JP, Schmidt JD, Kerr ZY, McCrea MA. Reliable change, sensitivity, and specificity of a multidimensional concussion assessment battery: implication for caution in clinical practice. *J Head Trauma Rehabil*. 2013;28(4):274–83. <https://doi.org/10.1097/HTR.0b013e3182585d37>.
 47. Resch JE, Brown CN, Schmidt J, Macciocchi SN, Blueitt D, Cullum CM, Ferrara MS. The sensitivity and specificity of clinical measures of sport concussion: three tests are better than one. *Med BMJ Open Sport Exerc*. 2016. <https://doi.org/10.1136/bmjsem-2015-000012>.
 48. Resch JE, McCrea MA, Munro CC. Computerized neurocognitive testing in the management of sport-related concussion: an update. *Neuropsychol Rev*. 2013;23(4):335–49. <https://doi.org/10.1007/211065-013-9242-5>.
 49. Robin X. Package “pROC.” Retrieved from <https://cran.r-project.org/web/packages/pROC/pROC.pdf>. (2019).
 50. Schatz P, Robertshaw S. Comparing post-concussive neurocognitive test data to normative data presents risks for underclassifying “above average” athletes. *Arch Clin Neuropsychol*. 2014;29(7):625–32. <https://doi.org/10.1093/arclin/acu041>.
 51. Schatz P, Sandel N. Sensitivity and specificity of the online version of ImpACT in high school and collegiate athletes. *Am J Sports Med*. 2013;41(2):321–6. <https://doi.org/10.1177/0363546512466038>.
 52. Schmidt JD, Register-Mihalik JK, Mihalik JP, Kerr ZY, Guskiewicz KM. Identifying impairments after concussion: normative data versus individualized baselines. *Med Sci Sports Exerc*. 2012;44(9):1621–8. <https://doi.org/10.1249/MSS.0b013e318258a9fb>.
 53. Thoma RJ, Cook JA, McGrew C, King JH, Pulsipher DT, Yeo RA, et al. Convergent and discriminant validity of the ImpACT with traditional neuropsychological measures. *Cogent Psychol*. 2018;5(1):1430199. <https://doi.org/10.1080/23311908.2018.1430199>.
 54. Walpole RE, Myers RH, Myers SL, Ye K. Probability & statistics for engineers & scientists. 9th ed. Upper Saddle River: Prentice Hall; 2012.
 55. Wood TA, Hsieh KL, An R, Ballard RA, Sosnoff JJ. Balance and gait alterations observed more than 2 weeks after concussion: a systematic review and meta-analysis. *Am J Phys Med Rehabil*. 2019;98(7):566–76. <https://doi.org/10.1097/PHM.0000000000001152>.

Affiliations

Lauren L. Czerniak¹  · Spencer W. Liebel²  · Gian-Gabriel P. Garcia³  · Mariel S. Lavieri¹ · Michael A. McCrea^{4,5} · Thomas W. McAllister⁶ · Steven P. Broglio⁷  · CARE Consortium Investigators

¹ Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Ave, Ann Arbor, MI 48109, USA

² Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, USA

³ Massachusetts General Hospital Institute for Technology Assessment, Harvard Medical School, Boston, MA 02114, USA

⁴ Department of Neurosurgery, Medical College of Wisconsin, Milwaukee, WI, USA

⁵ Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA

⁶ Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN, USA

⁷ Michigan Concussion Center, University of Michigan, Ann Arbor, MI, USA