Journal of Athletic Training 2015;50(12):1319–1322
doi: 10.4085/1062-6050-51.1.09

www.natajournals.org

short report

False-Positive Rates of Reliable Change Indices for Concussion Test Batteries: A Monte Carlo Simulation

Lindsay D. Nelson, PhD

Departments of Neurosurgery and Neurology, Medical College of Wisconsin, Milwaukee

Background: Neurocognitive testing is widely performed for the assessment of concussion. Athletic trainers can use preseason baselines with reliable change indices (RCIs) to ascertain whether concussed athletes' cognitive abilities are below preinjury levels. Although the percentage of healthy individuals who show decline on any individual test is determined by its RCI's confidence level (eg, 10% false-positive rate using an RCI with an 80% confidence interval), the expected rate of 1 or more significant RCIs across multiple indices is unclear.

Objective: To use a Monte Carlo simulation procedure to estimate the normal rate (ie, base rate) of significant decline on 1 or more RCIs in multitest batteries.

Results & Conclusion: For batteries producing 7 or more uncorrelated RCIs (80% confidence intervals), the majority of normal individuals would show significant declines on at least 1 RCI. Expected rates are lower for tests with fewer indices, higher inter-RCI correlations, and more stringent impairment criteria. These reference points can help testers interpret RCI output for multitest batteries.

Key Words: neurocognitive assessment, base rates, impairment

Key Points

- Clinicians evaluating concussed athletes often rely on the results of multiple tests or subtests to determine whether the athletes remain impaired.
- Base rates (ie, false-positive rates) of impairment are higher across multiple tests than for single tests, yet joint base rates of impairment are often not published.
- This simulation study illustrates how the properties of a test battery affect the expected base rates of impairment on 1 or more indices within the test battery.
- These data can be referenced when making decisions about how to set cut scores for determining impairment in the context of postconcussive evaluations.

Formal assessment of concussed athletes is commonplace in sports medicine, and a number of assessment tools are available to quantify symptoms, cognitive impairments, and other injury sequelae.[1–4] Computerized neurocognitive tests (CNTs) are especially popular for assessing neuropsychological abilities,[5,6] with nearly 40% of athletic trainers reporting use of a CNT in their concussion-management protocols in 2009–2010.[5] The vast majority (85.9% to 94.7%) of athletic trainers who use CNTs perform preseason assessments so that concussed athletes' postinjury scores can be compared with their individual premorbid estimates of ability.[5,7] The CNT programs facilitate the comparison of postinjury scores with baseline scores using output about the significance of reliable change indices (RCIs), which estimate the extent to which changes in athletes' performance are statistically unusual after taking into account measurement error inherent to a test. In addition to CNTs, RCI cutoffs have been published for other concussion tests, including the Sport Concussion Assessment Tool 3 (SCAT3)[8,9] and paper-and-pencil tests of psychomotor speed.[10]

Yet there is little published guidance about how to interpret RCI output for batteries with multiple indices. For any RCI, the expected false-positive rate is determined by the confidence interval (CI) applied to that RCI. For example, an RCI with a 90% confidence level should classify 5% of normal, healthy individuals as significantly declined from baseline (and, likewise, 5% as significantly improved). Similarly, 80% and 95% CIs should classify 10% and 2.5%, respectively, of normal individuals as significantly declined on average. Clinicians should select thresholds for significance according to their preferences for balancing sensitivity and specificity, with more lenient criteria expected to identify more impairment in concussed athletes (ie, increasing sensitivity) while inevitably also falsely identifying more healthy individuals as impaired (ie, diminishing specificity).

Although the specificity for 1 RCI is predictable, the base rates (ie, rates of abnormal scores in the normal population) of significant decline across sets of RCIs have not been well documented. This is a problem because clinical decisions usually involve interpreting the results of multiple indices simultaneously, and the base rates of impairment for 1 or more RCIs considered together should be higher than the rate for individual RCIs. Knowing the joint probability of producing various numbers of significant indices in a set is critical to making informed clinical decisions for any test battery, as the neuropsy-

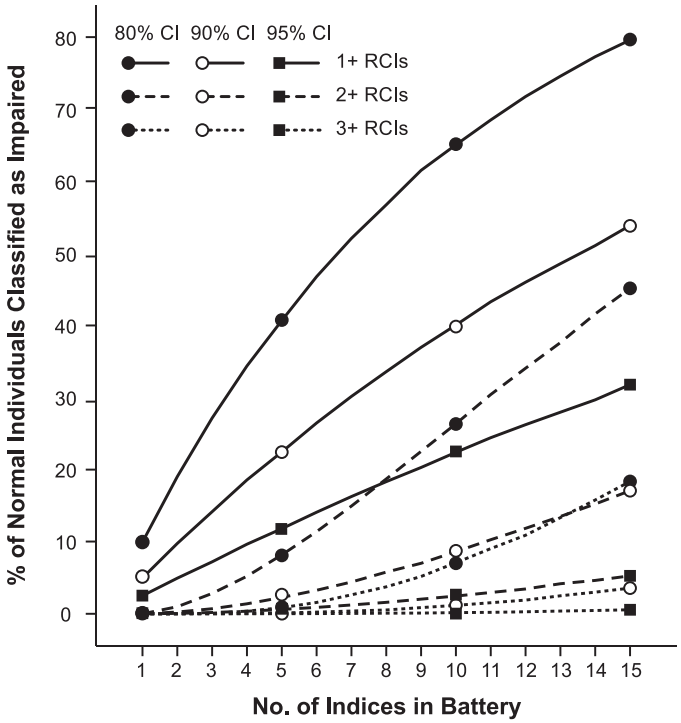

Figure 1. Expected percentage of normal, healthy individuals who would be classified as impaired on uncorrelated reliable change indices (RCIs), stratified by number of indices in a test battery and number of RCIs required to classify an individual as impaired. Abbreviation: CI, confidence interval.

chology literature has documented. For example, among community participants who completed a comprehensive neuropsychological assessment (comprising an average of 24 tests producing 43 scores), 71.9% of individuals showed at least 1 impaired score using a 1-standard deviation cutoff for impairment.[11] For a shorter neuropsychological battery (Wechsler Adult Intelligence Scale–III) with only 4 composite (index) scores, 24% of healthy individuals produced at least 1 abnormal index score (below the 10th percentile).[12] These rates depend on a number of factors, including the number of indices in the battery, their intercorrelations, and the impairment threshold for each measure. Supporting this principle in the context of concussion testing, 2 published studies on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) battery demonstrated that 22.2% to 46% of healthy individuals produced at least 1 significantly declined RCI out of 5.[13,14]

The aim of our study was to estimate the base rates of significant decline in 1 or more of a set of RCIs simulating a range of conditions that match most concussion-assessment batteries. This was achieved using a Monte Carlo simulation method that has been found to accurately estimate overall impairment base rates in other neuropsychological batteries.[11,12,15] By varying the number of indices per battery, the correlations among indices, and the criteria for significance, the relationship between these factors and test specificity was illustrated. I discuss these data in the context of the advantages and costs of applying different impairment criteria for concussion-management decisions.

METHODS

Data were simulated using a modified version of the Monte Carlo procedure described by Crawford et al.[12] Broadly, Monte Carlo simulation involves repeated random sampling from 1 or more data sets to estimate the probability of an event of interest. For the current study, this required simulating data sets to match important aspects of potential concussion-assessment batteries (eg, number of indices) and observing how often certain outcomes occurred across simulations. Specifically, the aim was to identify how varying the length of a test battery (conceptualized as the number of indices being interpreted) and the impairment criteria (eg, 80% versus 90% CI) influences the proportion of normal individuals deemed impaired on 1 or more indices in the battery. In order to best estimate these base rates for a particular set of measures, one needs to know the correlations among the measures (in this case, RCIs), which have not been published for the available concussion-assessment batteries. Thus, the primary aim of this analysis was to demonstrate the relationship among test battery length, impairment criteria, and base rates of impairment rather than to definitively estimate the true base rates for any particular assessment tool.

A routine was developed using the *mvrnorm* function in R[16] (package MASS, R Foundation for Statistical Computing, Vienna, Austria) to produce a series of multivariate random normal data sets of N = 100 000 individuals each for a range of conditions. In particular, for each iteration of the procedure, the number of tests/RCIs (from 1 to 15), correlations among RCIs (from 0 to 0.9 in 0.1 increments), and criteria for significance (z = 1.282, 1.645, and 1.960) were varied until all combinations of these factors were simulated. The z-score cutoffs were selected to match the most commonly used cutoffs for significant RCIs in the concussion-assessment literature (corresponding to 80%, 90%, and 95% confidence intervals, respectively). For each data set produced, the percentage of individuals with 1 or more, 2 or more, or 3 or more RCIs meeting the criteria for significant decline was computed. Because the base rates of interest are a result of the number of indices considered, the impairment cutoff for each index, and the correlations among the indices, it is not necessary to produce data with the same properties (eg, subtest means) to obtain valid estimates of the joint probabilities of impairment, nor is it important to actually compute RCIs from the data to estimate the joint probabilities of significance for different numbers of indices.

RESULTS

Figure 1 illustrates the relationship among test battery length, impairment criteria, and the expected false-positive rate for RCIs that are uncorrelated. As expected, the estimated rate of impairment increased substantially as the number of subtests/RCIs in the test battery increased and as the threshold for classifying scores as impaired became more lenient. For example, in a battery with 5 uncorrelated RCIs and at a confidence level of 80%, 41% of healthy

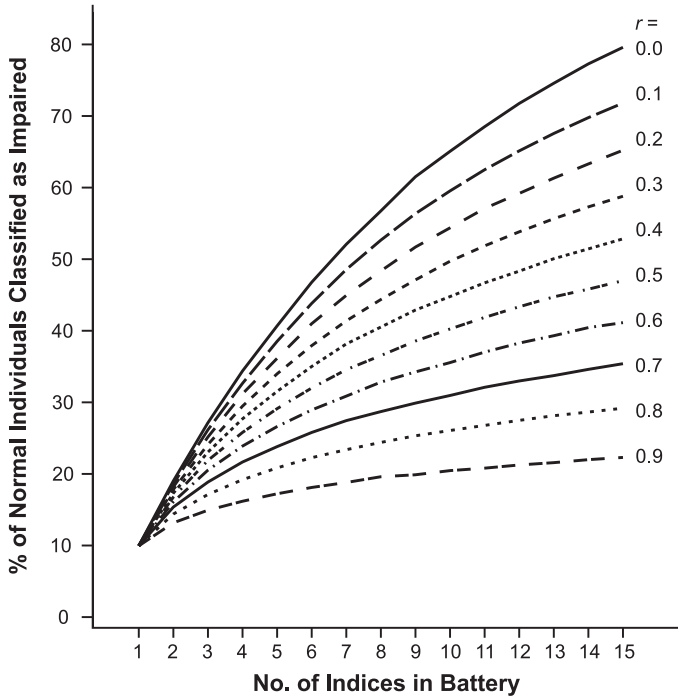

Figure 2. Expected percentage of normal, healthy individuals with 1 or more impaired indices by number of indices and their intercorrelations (using a $z = -1.282$ cutoff for impairment, akin to an 80% confidence interval). The *r* value presented on the right side of each line denotes the correlation between the indices (reliable change indices) in each simulated data set.

individuals would be expected to produce at least 1 RCI indicating a decline from baseline, with only 8% of individuals declining significantly on 2 or more tests and 1% declining on 3 or more tests. Once the number of indices increases to 7, the majority of normal individuals (52%) would be expected to show 1 or more significant declines when using RCIs with 80% CIs. For any given test battery, these rates decrease as one tightens the threshold for significance of each index in the battery (eg, for the given example of a test that produces 5 RCIs, the percentage of healthy individuals showing significant decline on 1 or more subtests falls to 23% and 12% when using a 90% or 95% CI for each RCI, respectively). Note that the percentages reported also reflect the expected base rates of significant improvement on retesting.

Figure 2 illustrates the role of inter-RCI correlations on the expected false-positive rates across a set of RCIs (using 80% CIs for each). To the extent that RCIs within a test battery are more correlated with each other, the base rates would be reduced. For example, in a set of RCIs correlated with each other at $r = 0.2$, the majority of healthy individuals would show a decline on at least 1 RCI in a set of 9 or more indices, whereas at a correlation of 0.4, a battery would need to produce at least 13 indices before the base rate of decline (on at least 1 RCI) rises over 50%.

CONCLUSIONS

This study illustrates an important and perhaps overlooked fact in the interpretation of multiple RCIs within a concussion test battery: that the percentage of normal, healthy individuals who are expected to show significant decline on at least 1 RCI in a multiple-test battery is higher than that percentage for individual RCIs. In particular, for a test battery that produces 7 or more uncorrelated RCIs, most normal individuals would demonstrate significant impairment on at least 1 RCI with 80% CIs for each. This concern is not unique to concussion tests or neuropsychological tests and rather is a statistical truism whenever one aggregates findings across a set of tests. The principle is the same as the inflated type I error rate resulting from multiple statistical comparisons but is easily forgotten in the context of clinical decisions. Although this language emphasizes the interpretation of RCIs given their widespread use for concussion assessment, the data also apply to the interpretation of single (eg, postinjury) scores.

As shown in Figure 2, the impairment rates should be lower when the indices are more highly correlated. For illustrative purposes, a wide range of correlations was modeled, but in real data sets, these correlations are likely relatively modest (ie, <0.3 on average). This is because indices with high correlations are unlikely to contribute uniquely to concussion assessment (and therefore to warrant separate inclusion in a battery) and because the difference scores that make up RCIs likely have lower maximum correlations with each other given that difference scores are often less reliable than single scores.[17] Authors[13,14] of the limited number of published studies on this topic have reported that 22.2% to 46% of healthy participants produce at least 1 (out of 5) significantly declined RCIs on ImPACT.

Sports medicine professionals should be aware of this principle and the data provided to make informed decisions about the most appropriate cutoff for impairment given the number of indices produced by their concussion-testing protocol and their goals for balancing sensitivity and specificity. Especially for longer test batteries, this may mean (1) selecting wider CIs around each RCI or (2) requiring more than 1 significant RCI before declaring an athlete below baseline. For users of commercialized CNTs who cannot select different CIs, option 2 may be their only choice. Interpretive guidelines have been published for ImPACT,[18] but it will be valuable for researchers to more routinely publish these base rates for a variety of common test batteries, including multimodal batteries that include cognitive, balance, and other measures.

These findings highlight the importance of reporting the rates of impairment across sets of indices and including uninjured control athletes, as both the sensitivity and specificity of a test determine its potential to contribute useful clinical information. However, few researchers have reported the sensitivity of CNTs across a set of RCIs; variable estimates ranged from values near the estimated rates of impairment reported here (42.9%)[19] to significantly greater values (78.6% to 90%).[20,21] It will be valuable for future investigators to identify the sources of variability in sensitivity rates (eg, differences in sample demographics, injury definition, or injury severity) to best determine the predictors of a poor neurocognitive outcome after concussion and the conditions under which neurocognitive testing is more or less informative for making concussion-management decisions. Building the evidence base around neurocognitive tests with an emphasis on clinically relevant metrics will undoubtedly further advance their clinical

utility and appropriate use in concussion-management programs.

ACKNOWLEDGMENTS

I thank Christopher Randolph, PhD, and Ashley LaRoche, CCRC, for reviewing a draft of this manuscript before its submission.

REFERENCES

1. McCrory P, Meeuwisse WH, Aubry M, et al. Consensus statement on concussion in sport: the 4th International Conference on Concussion in Sport held in Zurich, November 2012. *Br J Sports Med*. 2013; 47(5):250–258.
2. Covassin T, Elbin R III, Stiller-Ostrowski JL. Current sport-related concussion teaching and clinical practices of sports medicine professionals. *J Athl Train*. 2009;44(4):400–404.
3. Giza CC, Kutcher JS, Ashwal S, 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):2250–2257.
4. Harmon KG, Drezner JA, Gammons M, et al. American Medical Society for Sports Medicine position statement: concussion in sport. *Br J Sports Med*. 2013;47(1):15–26.
5. Meehan WP III, d'Hemecourt P, Collins CL, Taylor AM, Comstock RD. Computerized neurocognitive testing for the management of sport-related concussions. *Pediatrics*. 2012;129(1):38–44.
6. Resch JE, McCrea MA, Cullum CM. Computerized neurocognitive testing in the management of sport-related concussion: an update. *Neuropsychol Rev*. 2013;23(4):335–349.
7. Covassin T, Elbin RJ III, Stiller-Ostrowski JL, Kontos AP. Immediate post-concussion assessment and cognitive testing (ImPACT) practices of sports medicine professionals. *J Athl Train*. 2009;44(6):639–644.
8. Putukian M, Echemendia R, Dettwiler-Danspeckgruber A, et al. 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):36–42.
9. Barr WB, McCrea M. Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *J Int Neuropsychol Soc*. 2001;7(6):693–702.
10. Hinton-Bayre AD, Geffen GM, Geffen LB, McFarland KA, Friis P. Concussion in contact sports: reliable change indices of impairment and recovery. *J Clin Exp Neuropsychol*. 1999;21(1):70–86.
11. Schretlen DJ, Testa SM, Winicki JM, Pearlson GD, Gordon B. Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *J Int Neuropsychol Soc*. 2008;14(3):436–445.
12. Crawford JR, Garthwaite PH, Gault CB. Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: a generic method with applications. *Neuropsychology*. 2007;21(4): 419–430.
13. Resch J, Driscoll A, McCaffrey N, et al. ImPact test-retest reliability: reliably unreliable? *J Athl Train*. 2013;48(4):506–511.
14. Van Kampen DA, Lovell MR, Pardini JE, Collins MW, Fu FH. The "value added" of neurocognitive testing after sports-related concussion. *Am J Sports Med*. 2006;34(10):1630–1635.
15. Brooks BL, Iverson GL. Comparing actual to estimated base rates of "abnormal" scores on neuropsychological test batteries: implications for interpretation. *Arch Clin Neuropsychol*. 2010;25(1):14–21.
16. R Core Team. R: A language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2014.
17. Furr RM, Bacharach VR. *Psychometrics: An Introduction*. 2nd ed. Thousand Oaks, CA: SAGE Publications; 2014.
18. Iverson GL, Schatz P. Advanced topics in neuropsychological assessment following sport-related concussion. *Brain Inj*. 2015; 29(2):263–275.
19. Iverson GL, Lovell MR, Collins MW. Interpreting change on ImPACT following sport concussion. *Clin Neuropsychol*. 2003; 17(4):460–467.
20. Broglio SP, Macciocchi SN, Ferrara MS. Sensitivity of the concussion assessment battery. *Neurosurgery*. 2007;60(6):1050–1058.
21. Iverson GL, Brooks BL, Collins MW, Lovell MR. Tracking neuropsychological recovery following concussion in sport. *Brain Inj*. 2006;20(3):245–252.

Address correspondence to Lindsay D. Nelson, PhD, Departments of Neurosurgery and Neurology, Medical College of Wisconsin, 8701 West Watertown Plank Road, Milwaukee, WI 53226. Address e-mail to linelson@mcw.edu.