The development of an online neuropsychological test battery

The Amsterdam Cognition Scan

Feenstra, H.E.M.

Chapter 5

Binary cut-off scores give rise to artefactual disagreement between neuropsychological tests in cancer patients

Heleen E. M. Feenstra
Ivar E. Vermeulen
Jaap M. J. Murre
Sanne B. Schagen

ABSTRACT

Objective

To evaluate current criteria for detecting cognitive impaired patients and advocate possible improvements towards valid classification, we applied standard guidelines for impairment classification on results from two highly comparable neuropsychological batteries (one traditional and one online) in a random sample of 200 adult non-CNS cancer patients.

Method and Results

Using the ICCTF guidelines (1.5SD below the normative mean on two tests and/or 2SDs below the normative mean), the proportion of cognitively impaired patients (56% female; median age 53 years) was 40% (80/200) based on the traditional test battery, and 38% (76/200) based on the online test battery. No difference between proportions was found ($\chi^2(1)=.17$, $p=.68$). Within-person agreement in impairment classification was “fair” (K=.35). Likewise, simulation-based analyses indicated limited average agreement for a 1.5SD criterion (K=.36 (SD=.10)), and average agreement dropped for a 2SDs criterion (K=.29 (SD=.15)).

Conclusions

The current evaluation of a standardized cognitive impairment detection method shows that using cut-off scores to classify patients can lead to a situation where two (neuropsychological) test batteries that are highly similar have only limited agreement on who is classified as impaired. Additional simulation results indicate that, with the use of binary cut-off scores, agreement between two classification methods is inherently low. Consequently, caution should be taken when applying cut-off scores for impairment detection in research and clinical practice. Initiatives to apply modern statistical tools are required to improve the validity of impairment detection.

INTRODUCTION

A subset of cancer patients without central nervous system disease develops cognitive impairment following chemotherapy (Ahles & Root, 2018; Janelsins et al., 2017; Vardy et al., 2015). This impairment is likely to affect daily functioning and quality of life. Typical detection of cognitive impairment includes assessment with standardized neuropsychological tests. Nevertheless, neuropsychological literature reports widely varying frequencies of detected cognitive impairment, e.g., ranging from 13% to 75% in patients following treatment (Janelsins et al., 2014; Wefel et al., 2011). Apart from differences in patient population and disease- and treatment characteristics, this variance has been attributed to studies' differences in employed neuropsychological tests, reference populations, and cut-off points for defining cognitive impairment (Clapp et al., 2018; Shilling, Jenkins, & Trapala, 2006; Wefel et al., 2011). Hence, for accurate estimates of incidence, as well as severity and risk factors, more standardized impairment detection criteria are needed. This need was addressed by the International Cancer and Cognition Task Force (ICCTF), who proposed to apply core neuropsychological tests as well as standard cut-off points to differentiate between impaired and unimpaired cancer patients.

In this paper, we reflect on such impairment detection criteria as several like these are used in neuropsychological research and clinical practice (Boringa et al., 2001; Carey et al., 2004; Litvan et al., 2012). We demonstrate their use in a sample of non-CNS cancer patients tested twice with two highly comparable test batteries. Participants completed one face-to-face and one online neuropsychological test battery—the Amsterdam Cognition Scan (ACS)—, which was developed as a mirror image of the traditional battery. The ACS showed satisfactory reliability (intraclass correlation coefficient for the battery =.78) and validity scores (Pearson's *r* between both batteries =.78). Our expectation was that, after applying the standardized procedure to assess cognitive impairment, both batteries would not only (1) classify a similar proportion of participants as impaired, but also (2) show relatively high agreement on which participants are classified as impaired.

METHOD

A total of 200 cancer patients (56% female; median age, 53 y [range, 21-76 y]; 41% breast, 19% testis/prostate, 40% other; initial treatment between 1 and 5 years prior to testing: 77% chemotherapy, 45.5% endocrine therapy, 14% immunotherapy), recruited through the Netherlands Cancer Institute to assess validity of the ACS, completed traditional neuropsychological tests conducted by a trained research assistant as well as the ACS online (unmonitored) in a counterbalanced design. Table 1 of Chapter 3 describes test domains and main outcome measures of the traditional tests and their ACS equivalents.

To assess cognitive impairment, we applied the ICCTF-proposed fixed cut-off points, combined with available norms for the tests in both batteries. Traditional norms were derived from test manuals or publications (see Table 1, Chapter 3). ACS norms were collected previously, and corrected for demographical factors (gender, age, education) (Feenstra et al., 2018a). All second assessments were corrected for order effects (Feenstra et al., 2018b). Following ICCTF recommendations, a patient who scored 1.5 standard deviations (SD) below the normative mean on two tests and/or 2SDs below the normative mean on one test was classified as cognitively impaired (Wefel et al., 2011). Agreement of classification between traditional and ACS assessments was evaluated using Cohen's Kappa, both for the overall batteries and for the individual measures (using a 1.5SD criterion). Binominal tests were used to compare observed proportions of impairment with expected proportions (Ingraham & Aiken, 1996). Analyses were conducted using IBM SPSS Statistics for Windows, Version 22.0 (Armonk, NY: IBM corp.).

RESULTS

Using the ICCTF guidelines, the proportion of cognitively impaired patients was 40% (80/200) based on traditional tests, and 38% (76/200) based on the ACS (see Table 1). There were no differences in the percentage of patients classified as cognitively impaired between both test batteries ($\chi^2(1)=.17$, *p*=.68).Compared to an expected 18% in a healthy population (using a 2SD criterion), both test batteries showed an elevated degree of cognitive impairment in our patient sample (binominal tests: 35.5% (71/200), *p*<.001 for the traditional battery; 32% (64/200), *p*<.001 for the ACS).

Table 1. Cross table with ICCTF impairment numbers based on traditional (vertical) and ACS (horizontal) assessments.

		Traditional		
		Unimpaired % (n)	Impaired % (n)	Total % (n)
ACS	Unimpaired % (n)	45.5 (91)	16.5 (33)	62 (124)
	Impaired % (n)	14.5 (29)	23.5 (47)	38 (76)
	Total % (n)	60 (120)	40 (80)	100 (200)

Analysis of agreement in impairment classification between the traditional battery and the ACS yielded a K of .35, indicating a "fair" agreement (Altman, 1991). For individual measures, K ranged from .12 (Visual Reaction Time/ Reaction Speed) to .24 (Grooved Pegboard/ Fill the Grid), indicating poor to fair agreement (see Table 2).

There are several possible reasons why both test batteries, despite their significant correlation, show only limited agreement on who is classified as impaired. A likely reason relates to the crudeness of using binary cut-off points. Such cut-off points are susceptible to random variation; slight decreases or increases in raw scores may flip classification. To illustrate how this may affect agreement, we conducted a bootstrapping simulation. We randomly selected 200 cases from a normally distributed population of 2,000. Each case had two test scores, correlating at .7 (a typical correlation for two tests of the same cognitive function). Cases with a score of 1.5SD below the normative mean (operationalized as 1.2SD below the population mean)[1] were classified as impaired. We repeated this procedure 5,000 times, each time assessing impairment detection agreement between both "tests". Results showed an average agreement of K=.36 (SD=.10), similar to our observed K=.35. When running 5,000 bootstraps with 2SDs as impairment criterion (1.7SD below population mean), average agreement dropped to K=.29 (SD=.15)[2]. These results suggest that using binary cut-off points will lead to limited categorization agreement between assessments, even when such assessments correlate strongly.

Table 2. Agreement between traditional and ACS impairment numbers.

Traditional/ ACS (online) tests	Cohen's kappa
Trail Making Test A/ Connect the Dots I	.22
Trail Making Test B/ Connect the Dots II	.20
15 Words test/ Wordlist Learning	.21
15 Words test delay/ Wordlist Delayed Recall	.19
Visual Reaction Time/ Reaction Speed	.12
Tower of London/ Place the Beads	.21
Corsi Block-tapping/ Box Tapping[a]	.14
Grooved Pegboard/ Fill the grid	.24
WAIS III Digit Span/ Digit Sequences[b]	N.A.
ICCTF impaired	.35

a = Zero-scores (n=8) on the ACS test "Box Tapping" were not included in the analysis, because they resulted from a technical error. b = Unable to perform analyses because of one empty cross-table cell.

Additionally, in our specific case the use of battery-specific norms, a regression-to-the-mean effect well known in repeated testing (e.g., Bartnett et al., 2004), and an unequal distribution of "impaired" and "unimpaired" cases (Flight & Julious, 2015) may have contributed negatively the observed agreement scores.

DISCUSSION

The current evaluation of a standardized method for detection of cognitive impairment shows that using cutoff scores to classify patients can lead to a situation where two (neuropsychological) test batteries that are highly similar and that classify the same percentage of patients as impaired nonetheless fail to identify the same patients as cognitively impaired. This may question the external validity of the tests used: apparently, the test batteries fail to unambiguously identify patients who suffer from impaired cognitive functioning. However, a bootstrap simulation shows that expected agreement between two classification methods based on cut-off scores is low to begin with. From our analyses, we conclude that using binary cut-off scores for impairment detection may introduce large artefacts and tends to foster misclassification. This

[1] That is, for this analysis we set the (patient) population mean to be .3SD below the normative mean.
[2] Continued simulations show that agreement increases when (a) correlation between the tests increases, and (b) impairment cut-off points are closer to the normative mean.

means that great caution should be taken when applying this method in research and clinical practice. Fortunately, new initiatives are undertaken to apply modern statistical tools, such as the Multivariate Normative Comparison (MNC, (van Rentergem, Murre, & Huizenga, 2017)), to improve the validity of classifying individuals as impaired. In contrast to crude impairment classification methods such as the ICCTF criteria, MNC take into account whether the *entire* cognitive profile of a patient (as assessed through a test battery) is common in healthy persons using a single comparison.

A more nuanced interpretation of neuropsychological test scores will improve the diagnostic process of patients, including cancer patients, as well as research efforts into the cognitive effects of cancer and cancer therapies.