Binary classification threatens the validity of cognitive impairment detection

Luijendijk, Maryse J.; Feenstra, Heleen E. M.; Vermeulen, Ivar E.; Murre, Jaap M. J.; Schagen, Sanne B.

published in
Neuropsychology
2023

DOI (link to publisher)
10.1037/neu0000831

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Luijendijk, M. J., Feenstra, H. E. M., Vermeulen, I. E., Murre, J. M. J., & Schagen, S. B. (2023). Binary classification threatens the validity of cognitive impairment detection. *Neuropsychology*, *37*(3), 344-350. https://doi.org/10.1037/neu0000831

AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology

ISSN: 0894-4105

2023, Vol. 37, No. 3, 344–350
https://doi.org/10.1037/neu0000831

Binary Classification Threatens the Validity of Cognitive Impairment Detection

Maryse J. Luijendijk[1], Heleen E. M. Feenstra[1, 2], Ivar E. Vermeulen[3], Jaap M. J. Murre[2], and Sanne B. Schagen[1, 2]

[1] Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[2] Brain and Cognition Group, Psychology, University of Amsterdam
[3] Department of Communication Science, VU University Amsterdam

Objective: Neuropsychological literature reports varying prevalence of cognitive impairment within patient populations, despite assessment with standardized neuropsychological tests. Within the domain of oncology, the International Cognition and Cancer Task Force (ICCTF) proposed standard cutoff points to harmonize the operationalization of cognitive impairment. We evaluated how this binary classification affects agreement between two highly comparable test batteries. ***Method:*** Two hundred non-central nervous system (non-CNS) cancer patients who finished treatment (56% females; median age 53 yrs) completed traditional tests and their online equivalents in a counterbalanced design. Following ICCTF standards, impairment was defined as a score of ≥1.5 standard deviations (*SD*s) below normative means on two tests and/or ≥2 *SD*s below normative means on one test. Agreement of classification between traditional and online assessment was evaluated using Cohen's κ. Additional Monte Carlo simulations were conducted to demonstrate how different cutoff points and test characteristics affect agreement. ***Results:*** The correlation between total scores of traditional and online assessment was .78. Proportions of impaired patients did not differ between assessment methods: 40% using traditional tests and 38% using online equivalents, $\chi^2(1) = .17$, $p < .68$. Nevertheless, *within-person* agreement in impairment classification between traditional and online assessment was merely fair ($K = .35$). Monte Carlo simulations showed similarly low agreement scores ($K = .41$ for 1.5 *SD*; $K = .33$ for 2 *SD* criterion). ***Conclusions:*** Our results show that binary classification can lead to a situation where two highly similar batteries fail to identify the same individuals as impaired. Additional simulations suggest that within-person agreement between assessment methods using binary classification is *inherently* low. Modern statistical tools may help to improve validity of impairment detection.

Key Points

Question: How does the use of binary cutoff points affect agreement between neuropsychological assessments in the classification of cognitively impaired versus intact? ***Findings:*** Comparable test batteries (traditional tests and their online equivalents) show limited agreement on who is classified as cognitively impaired; simulations show this limited agreement is inherent to the use of binary cutoff points. ***Importance:*** Great caution should be taken when applying binary cutoff points to differentiate between cognitively intact and impaired individuals in scientific research and clinical practice. ***Next Steps:*** Future research should investigate the use of modern statistical tools to improve the validity of cognitive impairment detection.

Keywords: binary classification, cognitive impairment detection, cutoff point, dichotomization

This article was published Online First July 4, 2022.

Maryse J. Luijendijk https://orcid.org/0000-0002-0435-8566
Heleen E. M. Feenstra https://orcid.org/0000-0002-7038-7875
Ivar E. Vermeulen https://orcid.org/0000-0003-3589-8773
Jaap M. J. Murre https://orcid.org/0000-0003-4447-9482
Sanne B. Schagen https://orcid.org/0000-0003-0153-8059

R code of the simulations is openly available at the project's Open Science Framework page (https://osf.io/t4yc2/?view_only=31ace180ecc14351906eb905e93ff806). The authors have no known conflict of interest to disclose. This work was supported by the Dutch Cancer Society (Grant NKI 2015-7937) to Sanne B. Schagen. This study was not preregistered. The study was approved by the institutional review board of the Netherlands Cancer Institute and is in accordance with the 1964 Helsinki Declaration and its later amendments. Informed consent was obtained from all individual participants included in the study.

Maryse J. Luijendijk and Heleen E. M. Feenstra contributed equally to this work and are both first authors.

Maryse J. Luijendijk played equal role in conceptualization, formal analysis, and writing of original draft. Heleen E. M. Feenstra played equal role in conceptualization, formal analysis, and writing of original draft. Ivar E. Vermeulen played equal role in conceptualization and writing of review and editing. Jaap M. J. Murre played equal role in conceptualization and writing of review and editing. Sanne B. Schagen played equal role in conceptualization and writing of review and editing.

Correspondence concerning this article should be addressed to Maryse J. Luijendijk, Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. Email: m.luijendijk@nki.nl

Many neuropsychological studies use (prevalence of) cognitive impairment as an endpoint, for example, in studying the effects of human immunodeficiency virus, traumatic brain injury, neurodegenerative diseases, schizophrenia, or cancer. Within such research areas, between-study inconsistencies in the prevalence of cognitive impairment are often observed, despite assessment with standardized neuropsychological tests. Throughout this article, we will address this problem using cancer-related cognitive impairment as a case in point.

Neuropsychological studies investigating cognitive effects of cancer treatment report widely varying prevalence of cognitive impairment in non-central nervous system (non-CNS) cancer patients, ranging from 13% to up to 75% (Ahles & Root, 2018; Mayo et al., 2020). Apart from differences in patient populations, disease and treatment characteristics, employed neuropsychological tests, and reference populations, this variance has been attributed to studies' differences in cutoff points for defining cognitive impairment (Bernstein et al., 2017; Clapp et al., 2018; Wefel et al., 2011). Hence, for accurate estimates of prevalence, as well as severity and risk factors, more standardized impairment detection criteria are needed. This need was addressed by the International Cognition and Cancer Task Force (ICCTF), who proposed to harmonize the operationalization of cognitive impairment within the domain of oncology by applying standard cutoff points to differentiate between impaired and unimpaired cancer patients (Wefel et al., 2011). In this article, we reflect on such impairment detection criteria as several like these are used in neuropsychological research and clinical practice (Artemiadis et al., 2021; Goldman et al., 2018; Nightingale et al., 2021; Strassnig et al., 2018).

The goal of this article is to evaluate the reliability of binary cutoff points for identifying cognitive impairment in terms of their agreement in classification between assessment methods. First, 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; correlations for individual tests vary between .29 and .71) and validity scores (Pearson's r between both batteries = .78; correlations for individual tests vary between .36 and .70; Feenstra, Murre, et al., 2018). Using these data, we test the expectation that, after applying the standardized procedure to assess cognitive impairment, both batteries will not only (a) classify a similar proportion of participants as impaired but also (b) show relatively high agreement on which participants are classified as impaired.

Second, we conducted a simulation study to further illustrate how the choice of the cutoff point used as impairment detection criterion affects agreement in classification. We manipulated test characteristics such as the correlations between tests to demonstrate what expected levels of agreement would be under varying conditions.

Observational Study

Method

A total of 200 cancer patients (56% females; median age 53 years [range 21–76 years]; 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 (for a more detailed description of the data acquisition, see Feenstra, Murre, et al., 2018). Table 1 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 cutoff points, combined with available norms for the tests in both batteries. Traditional norms were derived from test manuals or publications (see Table 1). ACS norms were collected previously and corrected for demographical factors (gender, age, education; Feenstra, Vermeulen, et al., 2018). All second assessments were corrected for order effects (Feenstra, Murre, et al., 2018). Following ICCTF recommendations, a patient who scored 1.5 standard deviations (*SD*s) below the normative mean on two tests and/or 2 *SD*s

Table 1
The Online Tests of the Amsterdam Cognition Scan and Their Equivalent Traditional Tests

Online tests	Test domains	Main outcome measure	Traditional equivalent
Connect the dots I; connect the dots II	Visuomotor tracking, planning, cognitive flexibility, divided attention	Completion time (I and II)	Trail Making Test A Trail Making Test B (Reitan, 1958)
Wordlist learning	Verbal learning	Total number of correct words (Trial 1–5)	15 words test (Dutch version of Rey Auditory Verbal Learning Test; van den Burg et al., 1985)
Wordlist delayed recall and recognition	Retention of information: free recall and recognition	Total number of correct words; free recall and recognition	15 words test
Reaction speed	Information processing speed and attention	Mean reaction time	Visual reaction time (subtest FePsy; Alpherts & Aldenkamp, 1995)
Place the beads	Planning, response inhibition, visuospatial memory	Total number of extra moves	Tower of London, Drexel University (TOL-dx) (Culbertson & Zillmer, 2001)
Box tapping	Visuospatial short-term memory	Total number of correctly repeated sequences	Corsi block-tapping test (Kessels et al., 2000)
Fill the grid	Fine motor skills	Completion time	Grooved Pegboard (Kløve, 1963)
Digit sequences I; digit sequences II	I: Attention II: Working memory	Total number of correctly repeated sequences (I and II)	WAIS-III Digit Span (forward and backward; Tulsky et al., 1997)

Note. WAIS-III = Wechsler Adult Intelligence Scale-III.

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 κ, both for the overall batteries and for the individual measures (using a 1.5 *SD* criterion). Binomial tests were used to compare observed proportions of impairment with base rate proportions (Brooks & Iverson, 2010; Ingraham & Aiken, 1996). Analyses were conducted using IBM SPSS Statistics for Windows, Version 22.0 (Armonk, New York: 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 2). 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 18% base rate in a healthy population (using a 2 *SD* criterion), a base rate commonly observed (Binder et al., 2009), both test batteries showed an elevated degree of cognitive impairment in our patient sample (binomial tests: 35.5% [71/200], $p < .001$ for the traditional battery; 32% [64/200], $p < .001$ for the ACS).

Analysis of agreement in impairment classification between the traditional battery and the ACS yielded a *K* of .35, indicating only a "fair" agreement (Altman, 1991). For individual ACS subtests and their traditional equivalents, *K* ranged from .11 (Wechsler Adult Intelligence Scale-III [WAIS-III] Digit Span/digit sequences) to .24 (Grooved Pegboard/fill the grid), indicating poor to fair agreement, despite moderate-to-strong correlations between the raw measures (see Table 3).

Discussion

Our findings show that the use of impairment detection criteria to binary classify test scores can lead to a situation where two comparable (neuropsychological) test batteries classify the same percentage of patients as impaired, but nonetheless fail to agree on who is cognitively impaired. This lack of reliability in classification may question the external validity of the tests used: Apparently, the test batteries fail to unambiguously identify patients who suffer from impaired cognitive functioning.

At first, the relatively low-to-moderate test correlations at the individual level seem a plausible explanation for the limited agreement. Yet, agreement is not necessarily higher for tests that correlate more strongly. In the study below, we will test whether the limited agreement is due to these low correlations and/or specific to the tests we employed, or an inherent problem of the use of binary cutoff points. Using simulations, we will demonstrate what levels of agreement can be expected if test scores correlate to different degrees. We will show that the use of such cutoff points will generally lead to limited categorization agreement between assessments, even when these assessments correlate strongly.

Table 2
ICCTF Impairment Numbers Based on Traditional (Vertical) and ACS (Horizontal) Assessments

	Traditional		
ACS	Unimpaired % (*n*)	Impaired % (*n*)	Total % (*n*)
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)

Note. ICCTF = International Cognition and Cancer Task Force; ACS = Amsterdam Cognition Scan.

Table 3
Correlations and Agreement in Impairment Numbers Between Traditional and ACS Tests

Traditional/ACS (online) tests	Correlation[a]	Cohen's κ
Trail Making Test A/connect the dots I	.57	.22
Trail Making Test B/connect the dots II	.70	.20
15 Words test/wordlist learning	.64	.21
15 Words test delay/wordlist delayed recall	.59	.19
Visual reaction time/reaction speed	.49	.12
Tower of London/place the beads	.42	.21
Corsi block-tapping/box tapping[b]	.36	.14
Grooved Pegboard/fill the grid	.45	.24
WAIS-III Digit Span forward/digit sequences I[c]	.43	N.A.
WAIS-III Digit Span backward/digit sequences II[c]	.52	N.A.
Total score/ICCTF impaired	.78	.35

Note. ACS = Amsterdam Cognition Scan; ICCTF = International Cognition and Cancer Task Force; WAIS-III = Wechsler Adult Intelligence Scale-III.
[a] Adapted from "Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan" by Feenstra, Murre, et al. (2018), *Journal of Clinical and Experimental Neuropsychology*, *40*(3), p. 260 (https://doi.org/10.1080/13803395.2017.1339017). [b] Zero scores ($n = 8$) on the ACS test "box tapping" were not included in the analysis because they resulted from a technical error. [c] Unable to perform analyses because of one empty cross-table cell.

Simulation Study

Method

To illustrate how binary classification may affect agreement in general, we conducted Monte Carlo simulations. Each time, we simulated a sample of cases ($N = 200$) with two normally distributed sets of "test scores," correlating at a certain level. Cases with a score lower than a certain cutoff point below the normative mean (which we assume to be .3 *SD* higher than the patient mean, to account for generally lower scores in patient populations than in healthy populations) were classified as impaired. This was repeated 10,000 times, each time assessing impairment detection agreement between both "tests" by calculating Cohen's κ. To demonstrate how (a) the choice of the cutoff point used as impairment criterion and (b) the correlation between the two tests affects agreement scores, we conducted this simulation for cutoff points −.5, −1, −1.5, −2, and −2.5, correlations .5, .6, .7, .8, and .9, and all their 25 combinations.

Given that in research and clinical practice, cognitive impairment detection is generally based on a test battery with multiple tests rather than a single test, we repeated the simulation, this time simulating a sample of cases ($N = 200$) with two normally distributed sets of seven test scores. The simulation was performed for the same combinations of cutoff points and correlations. The intertest correlations (i.e., between two tests of different cognitive functions) were varied as well: .2, .3, .4, .5, and .6. Cases with two or more

scores below the cutoff point were classified as impaired. To assess impairment detection agreement between the simulated test batteries, Cohen's κ was calculated.

All simulations were performed using R. R code of the simulations is openly available at the project's Open Science Framework page (https://osf.io/t4yc2/?view_only=31ace180ecc14351906eb905e93ff806).

Results

Simulation results showed an average agreement of $K = .41$ ($SD = .10$) between two normally distributed sets of test scores correlating at .7 (a typical correlation for two tests of the same cognitive function; Agelink van Rentergem et al., 2020; Calamia et al., 2013) when applying 1.5 *SD*s below the normative mean as impairment criterion. This outcome of the simulation studies is quite similar to the $K = .35$ we observed in our participant population. When using 2 *SD*s below the normative mean as impairment criterion in the simulation, average agreement dropped to $K = .33$ ($SD = .15$). Continued simulations showed that agreement further decreases when the correlation between the two tests decreases, and when impairment cutoff points are further from the normative mean (see Figure 1A).

When classification is based not on a single test but on a test battery comprised of multiple tests (in which tests of the same cognitive function correlate at .7 while tests of different cognitive functions correlate at .4, an intertest correlation commonly observed in the field; Agelink van Rentergem et al., 2020), simulation results showed an average agreement of $K = .56$ ($SD = .08$) between the two test batteries using an impairment criterion of 1.5 *SD*s below the normative mean. For the criterion of 2 *SD*s below the normative mean, average agreement again dropped to $K = .48$ ($SD = .12$). Although somewhat higher than when classification is based on a single test, agreement is still only "moderate." Continued simulations again showed that agreement further deteriorates with decreasing correlations and stricter cutoff points (see Figure 1B). Similarly, manipulating the intertest correlations resulted in overall lower agreement as the correlations decreased.

Discussion

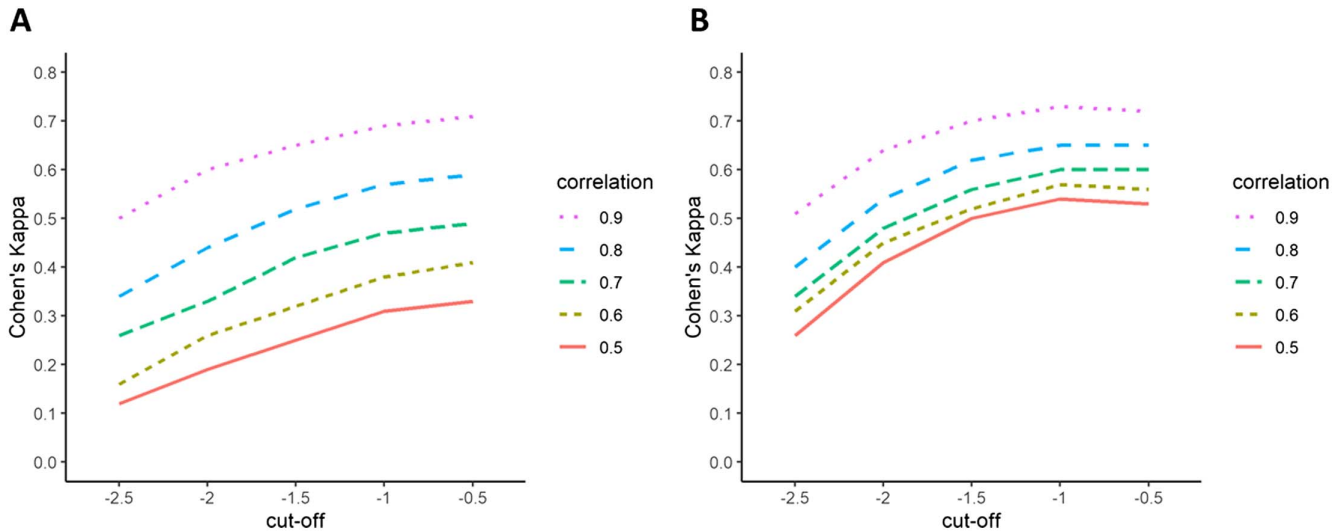
Using Monte Carlo simulations, we demonstrate that the expected within-person agreement between assessment methods (or between repeated assessments for that matter) using binary classification is low to begin with. Even in the ideal, simulated situation where data are perfectly normally distributed, the level of agreement remains limited; more so when reliability of the measurement instrument itself decreases. Our observed agreement appeared to be slightly lower than would be expected from the simulations, indicating that in practice, when circumstances are less optimal and several other factors influence the outcomes, these simulated agreement scores will not be reached.

General Discussion

The current evaluation of a standardized method for detection of cognitive impairment illustrates a problem associated with the use of binary classification: two comparable test batteries fail to identify the same patients as cognitively impaired. Our simulations show that for binary classification methods based on cutoff scores, agreement is *inherently* low. From our analyses, we conclude that using cutoff scores for impairment detection yields unreliable classification and thus tends to foster misclassification.

There are several possible reasons why two test batteries show only limited agreement on who is classified as impaired. Our simulation results indicate that this cannot be explained by low-to-moderate correlations between two batteries. Rather, it appears to be a problem inherent to binary classification that likely relates to the crudeness of using cutoff points. Such cutoff points are susceptible

Figure 1
Results of the Monte Carlo Simulations



Note. This figure visualizes the expected agreement (Cohen's κ) for a variety of cutoff points and correlations when classification is based on a single test (A) or a test battery comprised of seven tests (B; assuming intertest correlations of .4). See the online article for the color version of this figure.

to random variation; slight decreases or increases in raw scores may flip classification of individual cases. Regression to the mean effects—well known in repeated testing (e.g., Barnett et al., 2005)—may play a role as well. Regression-to-the-mean refers to the effect that individuals who initially obtain highly deviating test scores often tend to score less extreme on a repeated measurement. This might flip classification from cognitively impaired to unimpaired, thereby affecting agreement between measurements. An unequal distribution of "impaired" and "unimpaired" cases (Flight & Julious, 2015) may also contribute negatively to the observed agreement scores. Finally, differences between norm groups could potentially affect agreement between test batteries as well, although evidence from literature remains inconclusive (Collins et al., 2013; Schilder et al., 2010; Wyman-Chick et al., 2018).

Interestingly, the simulation study shows that agreement becomes worse when impairment criteria are set to be stricter (for similar observations on categorical analysis of continuous data, see MacCallum et al., 2002). While this might seem counterintuitive at first, this may be explained by the reduced sensitivity of more stringent criteria in populations in which cognitive impairment is rare: Although the absolute number of "false positives" might decrease for criteria further from the normative mean, the proportion of impaired "cases" relative to this number decreases drastically as well. Consequently, sensitivity drops, implying that chances are relatively high that classification as cognitively impaired turns out to be incorrect. From this, it may also be inferred that in populations with a higher degree of cognitive impairment (e.g., patients with Alzheimer's disease), agreement in classification when applying strict criteria would be (slightly) better.

Misclassification of cognitive impairment can have serious implications. Especially when binary classification aids decision-making, its inherent reliability problems can have serious consequences, and great caution should be taken. For example, when patients undergo neuropsychological assessment in the context of work disability (Nieuwenhuijsen et al., 2009), minor fluctuations that flip classifications and may well be the result of random error, may have severe personal impact. Furthermore, misdiagnosis can lead to patients being referred to inadequate health care or missing out on necessary treatment altogether (Gaugler et al., 2013). This may result in poorer health outcomes and higher health care costs, impacting society as a whole (Hunter et al., 2015). In addition, the social stigma that can accompany an incorrect diagnosis could cause considerable harm as well, including psychological distress for patients and their families (Herrmann et al., 2018). Finally, misclassification could impede research into treatment or prevention of diseases by including incorrectly diagnosed patients in clinical trials. These examples illustrate the importance to develop alternative approaches to classification in order to prevent misclassification.

New initiatives could perhaps apply modern statistical tools, such as dimensional models to improve the validity of classifying individuals as impaired. In contrast to crude impairment classification methods such as the ICCTF criteria, dimensional models may consider the entire cognitive profile of a patient (as assessed through a test battery) in the classification. In multivariate normative comparison (MNC), for example, this pattern of test scores is compared in its entirety to that of a normative sample (Huizenga et al., 2007). By taking into account the relations between tests, MNC can detect certain combinations of test scores that may indicate an "abnormal" cognitive profile. As the MNC is a profile analysis, only a single comparison has to be performed. This reduction in the number of comparisons also reduces the chance of false-positive or false-negative results (Huizenga et al., 2007).

Within scientific research, another alternative could be to abandon the use of arbitrary predefined cutoff points and apply subgrouping analyses instead (e.g., Agelink van Rentergem et al., 2021). These methods determine whether subgroups can be distinguished from multivariate data—for instance, one cognitively impaired and one cognitively intact subgroup—and which criterion best separates them. An example of subgrouping analysis is latent profile analysis (LPA), which aims to identify subgroups in which individuals share a comparable "profile" (Barvas et al., 2021; Spurk et al., 2020), or in our case, a certain set of neuropsychological test scores. LPA assumes that sets of test scores originate from one or more multivariate distributions and determines the optimal number of groups that can be discriminated by estimating the optimal number of these distributions. Hence, instead of using a predefined criterion to separate groups, the criterion that best separates them is determined in a data-driven and more meaningful manner. Possibly, this could provide clinicians with evidence-based criteria for individual classification.

Finally, another option could be to replace dichotomization with a thrichotomization method (Landsheer, 2018), for example, by defining an interval of uncertainty: A range of inconclusive test scores that hardly provide any information about the absence or presence of cognitive impairment and will often result in incorrect classification. This ensures more careful handling of these inconclusive results and, by avoiding misclassification of the individuals that fall within this range, improves the rate of correctly classifying the remaining individuals (e.g., related to cognitive impairment detection, see Landsheer, 2020). Further research is required to examine which alternative method yields the most reliable and valid results.

Often, binary classification remains unavoidable. Given that there is no optimal solution to this problem yet, great caution should be taken when applying impairment cutoffs in research and clinical practice. When conducting research, it is important to adhere to standardized methods (which include the use of standardized tests and cutoff points) in order to limit other potential sources of variation and ensure optimal comparison between studies. In clinical practice, diagnosis of cognitive impairment must never be based solely on the extent to which test scores deviate. Neuropsychological test scores should be considered as components of a bigger picture and evaluated with the clinical judgment of an experienced neuropsychologist in context of other variables, including self-reported questionnaires or anamnestic examination. 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 or other domains within the field of neuropsychology.

References

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., Huizenga, H. M., & the ANDI Consortium. (2020). The factor structure of cognitive functioning in cognitively healthy participants: A meta-analysis and meta-analysis of individual participant data.

Neuropsychology Review, *30*(1), 51–96. https://doi.org/10.1007/s11065-019-09423-6

Agelink van Rentergem, J. A., Deserno, M. K., & Geurts, H. M. (2021). Validation strategies for subtypes in psychiatry: A systematic review of research on autism spectrum disorder. *Clinical Psychology Review*, *87*, Article 102033. https://doi.org/10.1016/j.cpr.2021.102033

Ahles, T. A., & Root, J. C. (2018). Cognitive effects of cancer and cancer treatments. *Annual Review of Clinical Psychology*, *14*(1), 425–451. https://doi.org/10.1146/annurev-clinpsy-050817-084903

Alpherts, W., & Aldenkamp, A. P. (1995). *FePsy: The iron psyche, manual. Heemstede: Instituut voor epilepsiebestrijding*.

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman & Hall.

Artemiadis, A., Bakirtzis, C., Chatzittofis, A., Christodoulides, C., Nikolaou, G., Boziki, M. K., & Grigoriadis, N. (2021). Brief international cognitive assessment for multiple sclerosis (BICAMS) cut-off scores for detecting cognitive impairment in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, *49*, Article 102751. https://doi.org/10.1016/j.msard.2021.102751

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Barvas, E., Mattavelli, G., Zappini, F., Giardina, F., Ottaviani, D., & Papagno, C. (2021). Cognitive phenotypes in Parkinson's disease: A latent profile analysis. *Neuropsychology*, *35*(4), 451–459. https://doi.org/10.1037/neu0000737

Bernstein, L. J., McCreath, G. A., Komeylian, Z., & Rich, J. B. (2017). Cognitive impairment in breast cancer survivors treated with chemotherapy depends on control group type and cognitive domains assessed: A multilevel meta-analysis. *Neuroscience and Biobehavioral Reviews*, *83*, 417–428. https://doi.org/10.1016/j.neubiorev.2017.10.028

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*(1), 31–46. https://doi.org/10.1093/arclin/acn001

Brooks, B. L., & Iverson, G. L. (2010). Comparing actual to estimated base rates of "abnormal" scores on neuropsychological test batteries: Implications for interpretation. *Archives of Clinical Neuropsychology*, *25*(1), 14–21. https://doi.org/10.1093/arclin/acp100

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Clapp, J. D., Luta, G., Small, B. J., Ahles, T. A., Root, J. C., Graham, D., Hurria, A., Jacobsen, P. B., Jim, H., McDonald, B. C., Stern, R. A., Saykin, A. J., Mandelblatt, J. S., & the Thinking and Living with Cancer (TLC) Study. (2018). The impact of using different reference populations on measurement of breast cancer-related cognitive impairment rates. *Archives of Clinical Neuropsychology*, *33*(8), 956–963. https://doi.org/10.1093/arclin/acx142

Collins, B., Mackenzie, J., & Kyeremanteng, C. (2013). Study of the cognitive effects of chemotherapy: Considerations in selection of a control group. *Journal of Clinical and Experimental Neuropsychology*, *35*(4), 435–444. https://doi.org/10.1080/13803395.2013.781995

Culbertson, W. C., & Zillmer, E. A. (2001). *Tower of London-Drexel (TOL-DX) technical manual*. Multi-Health Systems.

Feenstra, H. E., Vermeulen, I. E., Murre, J. M., & Schagen, S. B. (2018). Online self-administered cognitive testing using the Amsterdam cognition scan: Establishing psychometric properties and normative data. *Journal of Medical Internet Research*, *20*(5), Article e192. https://doi.org/10.2196/jmir.9298

Feenstra, H. E. M., Murre, J. M. J., Vermeulen, I. E., Kieffer, J. M., & Schagen, S. B. (2018). Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan. *Journal of Clinical and Experimental Neuropsychology*, *40*(3), 253–273. https://doi.org/10.1080/13803395.2017.1339017

Flight, L., & Julious, S. A. (2015). The disagreeable behaviour of the kappa statistic. *Pharmaceutical Statistics*, *14*(1), 74–78. https://doi.org/10.1002/pst.1659

Gaugler, J. E., Ascher-Svanum, H., Roth, D. L., Fafowora, T., Siderowf, A., & Beach, T. G. (2013). Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: An analysis of the NACC-UDS database. *BMC Geriatrics*, *13*(1), Article 137. https://doi.org/10.1186/1471-2318-13-137

Goldman, J. G., Holden, S. K., Litvan, I., McKeith, I., Stebbins, G. T., & Taylor, J. P. (2018). Evolution of diagnostic criteria and assessments for Parkinson's disease mild cognitive impairment. *Movement Disorders*, *33*(4), 503–510. https://doi.org/10.1002/mds.27323

Herrmann, L. K., Welter, E., Leverenz, J., Lerner, A. J., Udelson, N., Kanetsky, C., & Sajatovic, M. (2018). A systematic review of dementia-related stigma research: Can we move the stigma dial? *The American Journal of Geriatric Psychiatry*, *26*(3), 316–331. https://doi.org/10.1016/j.jagp.2017.09.006

Huizenga, H. M., Smeding, H., Grasman, R. P., & Schmand, B. (2007). Multivariate normative comparisons. *Neuropsychologia*, *45*(11), 2534–2542. https://doi.org/10.1016/j.neuropsychologia.2007.03.011

Hunter, C. A., Kirson, N. Y., Desai, U., Cummings, A. K. G., Faries, D. E., & Birnbaum, H. G. (2015). Medical costs of Alzheimer's disease misdiagnosis among US medicare beneficiaries. *Alzheimer's & Dementia*, *11*(8), 887–895. https://doi.org/10.1016/j.jalz.2015.06.1889

Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, *10*(1), 120–124. https://doi.org/10.1037/0894-4105.10.1.120

Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., & de Haan, E. H. (2000). The corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, *7*(4), 252–258. https://doi.org/10.1207/S15324826AN0704_8

Kløve, H. (1963). *Grooved pegboard test user instructions*. Lafayette Instruments.

Landsheer, J. A. (2018). The clinical relevance of methods for handling inconclusive medical test results: Quantification of uncertainty in medical decision-making and screening. *Diagnostics*, *8*(2), Article 32. https://doi.org/10.3390/diagnostics8020032

Landsheer, J. A. (2020). Impact of the prevalence of cognitive impairment on the accuracy of the montreal cognitive assessment: The advantage of using two MoCA thresholds to identify error-prone test scores. *Alzheimer Disease and Associated Disorders*, *34*(3), 248–253. https://doi.org/10.1097/WAD.0000000000000365

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Mayo, S. J., Lustberg, M., Dhillon, H. M., Nakamura, Z. M., Allen, D. H., Von Ah, D., Janelsins, M., Chan, A., Olson, K., Tan, C. J., Toh, Y. L., Oh, J., Grech, L., Cheung, Y. T., Subbiah, I. M., Petranovic, D., D'Olimpio, J., Gobbo, M., Koeppen, S., … Peters, K. B. (2020). Cancer-related cognitive impairment in patients with non-central nervous system malignancies: An overview for oncology providers from the MASCC neurological complications study group. *Supportive Care in Cancer*, *29*, 2821–2840. https://doi.org/10.1007/s00520-020-05860-9

Nieuwenhuijsen, K., de Boer, A., Spelten, E., Sprangers, M. A., & Verbeek, J. H. (2009). The role of neuropsychological functioning in cancer survivors' return to work one year after diagnosis. *Social and Behavioral Dimensions of Cancer*, *18*(6), 589–597. https://doi.org/10.1002/pon.1439

Nightingale, S., Dreyer, A. J., Saylor, D., Gisslén, M., Winston, A., & Joska, J. A. (2021). Moving on from HAND: Why we need new criteria for cognitive impairment in people with HIV and a proposed way forward. *Clinical Infectious Diseases*. Advance online publication. https://doi.org/10.1093/cid/ciab366

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276. https://doi.org/10.2466/pms.1958.8.3.271

Schilder, C. M., Seynaeve, C., Linn, S. C., Boogerd, W., Gundy, C. M., Beex, L. V., van Dam, F. S., & Schagen, S. B. (2010). The impact of different definitions and reference groups on the prevalence of cognitive impairment: A study in postmenopausal breast cancer patients before the start of adjuvant systemic therapy. *Psycho-Oncology*, *19*(4), 415–422. https://doi.org/10.1002/pon.1595

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, Article 103445. https://doi.org/10.1016/j.jvb.2020.103445

Strassnig, M., Bowie, C., Pinkham, A. E., Penn, D., Twamley, E. W., Patterson, T. L., & Harvey, P. D. (2018). Which levels of cognitive impairments and negative symptoms are related to functional deficits in schizophrenia? *Journal of Psychiatric Research*, *104*, 124–129. https://doi.org/10.1016/j.jpsychires.2018.06.018

Tulsky, D., Zhu, J., & Ledbetter, M. F. (1997). *WAIS-III/WMS-III technical manual*. Psychological Corporation.

van den Burg, W., Saan, R. J., & Deelman, B. G. (1985). *15-woordentest. provisional manual*. University Hospital, Department of Neuropsychology.

Wefel, J. S., Vardy, J., Ahles, T., & Schagen, S. B. (2011). International cognition and cancer task force recommendations to harmonise studies of cognitive function in patients with cancer. *Lancet Oncology*, *12*(7), 703–708. https://doi.org/10.1016/S1470-2045(10)70294-1

Wyman-Chick, K. A., Martin, P. K., Weintraub, D., Sperling, S. A., Erickson, L. O., Manning, C. A., & Barrett, M. J. (2018). Selection of normative group affects rates of mild cognitive impairment in Parkinson's disease. *Movement Disorders*, *33*(5), 839–843. https://doi.org/10.1002/mds.27335

Received October 14, 2021
Revision received April 21, 2022
Accepted April 21, 2022 ■