**Supplemental File A: Uncertain Interval Trichotomization Methods**

Supplement A of: “Impact of the Prevalence of cognitive impairment on the accuracy of the Montreal Cognitive Assessment” by Johannes A. Landsheer (e-mail: j.a.landsheer@uu.nl).

The most popular method in medical decision making is using dichotomization, most often by applying the maximum Youden Index so that the sum of *Se* and *Sp* is maximized. The constructors of the MOCA have used a similar dichotomization, using a balance between *Se* and *Sp* 1. In the past 2–6 as well as more recently 7–10 researchers have argued against such dichotomization and have suggested methods for defining a range of uncertain test scores that does not allow for sufficient distinction between the patients. These methods result in three ranges (trichotomization): 1. a range that indicates the presence of the targeted disease relatively the best; 2. a range of uncertain test scores that offers very limited distinction between patients; and 3. a range that indicates the absence of the disease relatively the best.

This paper proposes a practical approach: test scores can be defined as uncertain when they offer about equal likelihood to come from the two populations of patients with and without the targeted disease. The goal of the data analysis is to identify the range of test scores that has about equal probabilities to be observed for both the populations of patients with and without the targeted disease. These test results are most prone to classification errors and form a separate class in between less error-prone scores. These outer classes offer more superior positive or negative classifications. Classification errors typically concentrate around the point of intersection of the two distributions or its equivalence, the optimal threshold where the Youden index (*J* = *Se* + *Sp* -1) reaches its maximal value 11.

Without losing generality, the descriptions of the statistics below assume that lower scores indicate a positive status of the patient and higher scores a negative status. To be able to explain the applied method of threshold estimation more clearly, it is useful to explain the equivalence and difference between predictive values, standardized predictive values and post-test probabilities.

***Predictive values***. Predictive values provide the probabilities for the presence of the disease, *given* the obtained test result 12. Predictive values consequently provide information about the accuracy of the classification. Commonly, the negative predictive values (NPV) is calculated for the range of test scores used for a negative classification (test scores > dichotomous cut-point c), leading to the formulation NPV = TN / (TN + FN) and the PPV for positive classifications (the range of test scores <= c; PPV = TP / (TP + FP)), where TN concern the number of true negative, FN false negative, TP true positive and FP false positive classifications. Within the context of three-way classification, a more general definition is needed. Predictive values provide the probabilities of the patient’s negative and positive true status, given the range of test scores x. More generally, predictive values are calculated directly based on the observed frequencies in the two samples of patients with and without the targeted disease. For a range of test scores x, if f0(x) and f1(x) are the frequencies of patients without and with the targeted disease given x, the negative predictive value (NPV) can be defined as: NPV(x) = f0(x) / (f0(x) + f1(x)) and the positive predictive value (PPV) as: PPV(x) = f1(x) / (f0(x) + f1(x)). This definition also shows that NPV(x) = 1 – PPV(x) when calculated for the same range of test scores x.

These predictive values are exact for the observed patients with and without the targeted disease and are valid for the observed sample prevalence. Interpreting the predictive values of individual test scores is straightforward. For instance, when 240 true patients from a sample have score 25, and 257 patients without the targeted disease have score 25 a patient who receives MoCA test score 25, will consequently have a 240 / (240 + 257) = 0.48 chance of CI. This number is exact for the sample involved. These predictive values therefore indicate the accuracies of the classifications in the sample, given the range of applied test score(s). As such, it is an important outcome for evaluating the accuracy of classification in a sample, given the observed test score(s). For comparisons of methods, in this paper the lower limit of .8 is considered a sufficient predictive value, both for NPV and PPV.

***Standardized predictive values***. Heston’s proposal 13,14 to standardize predictive values was made in the context of a single cut-point. However, it makes sense to also use a more general definition here, and to relate standardized predictive values to the *relative* frequencies or densities of the (range of) test score(s). The densities for a range of test scores x can be defined d0(x) = f0(x) / n0 and d1(x) = f1(x) / n1, where n0 and n1 are the number of observed patients in the two samples. The standardized negative predictive value (SNPV) is defined as SNPV(x) = d0(x) / (d0(x) + d1(x)) and the standardized positive predictive value (SPPV) as SPPV(x) = d1(x) / (d0(x) + d1(x)). The two distributions are weighted equally, or in other words, the prevalence is standardized to .5. Like Se and Sp, standardized predictive values are independent of prevalence when the samples are drawn from the same two populations. The interpretation of the standardized predictive values is not as straightforward as the interpretation of the common predictive values: they provide the estimated relative probability which of the two distributions makes the observed test score most likely, the distribution of the population of patients with or the population without the disease. If, for instance, 8% of true patients have score 25, and 11% of patients without CI have score 25, a patient with test score 25 has an estimated relative probability of 8 / (8 + 11) = 0.42 to belong to the population with cognitive impairment and a probability of 0.58 to belong to the population without cognitive impairment. The estimates improve with larger samples.

Standardized predictive values are most useful to identify the range of uncertain test scores that offer a limited distinction between the populations of patients with and without the targeted disease. It should also be noted that the predictive values of two samples of patients with and without the targeted impairment (PPV and NPV) can be different from the estimates of the standardized predictive values for the two populations (SPPV and SNPV). These differences are larger when prevalence deviates more strongly from .5 (see Supplemental File B where the distributions of 30 different US ADCs are shown).

***Post-test probabilities***. Posttest probabilities 15 may seem quite different from predictive values, but they are not. The posttest probability is equal to the positive predictive value when the pretest probability is set to the sample prevalence, while the standardized positive predictive value is equal to the posttest probability when the pretest probability is set to .5. Post-test probabilities are most versatile, as they can be calculated for every possible value of prevalence.

***Uncertain test scores***. This is defined as a range of test scores with about equal densities in the two distributions of patients with and without the targeted disease. Standardized predictive values are therefore most suited to the determination of this range of uncertain test scores. How much uncertainty can be allowed is open for discussion. This paper uses an SNPV and an SPPV < .667 (odds two to one) to define test scores that are too uncertain for classification concerning the presence of CI.

***More reliable estimates and smoothing***. Even when all conditions stay the same, we cannot expect to find the same test score for a patient on second application. Due to random influences, a second test score will be slightly lower or higher. Reliable estimates of these predictive probabilities are consequently needed, and these should be corrected for this randomness to a certain degree. In test theory, this random effect is estimated with the Standard Error of Measurement (SEM), which depends directly on the reliability of the test:, where s is the standard deviation of the test scores and r the estimated reliability of the test 16,17. The true score of an individual patient lies with some probability (roughly 68%) within a range of ±1 SEM around the observed test score. This provides information about the range of test scores where the true score of the patient can be expected. The average standardized predictive values of a fixed number of consecutive test scores (in this case 5) are calculated, where each subset of test scores is modified by a forward shift, excluding the first test score and including the next test score. Such a moving average smooths the predictive values, stabilizes the estimates across different samples, and mitigates peculiarities in the sample. For the determination of thresholds, standardized predictive values are calculated for the range of ±1 SEM around each test score to obtain more reliable predictive values.

A documented function (RPV) in the language R 18 that performs all calculations is made available in the R-package UncertainInterval 19.

The first step is to define the cutoff scores for the range of uncertain test scores that disallow positive or negative classification, based on all results. For this, we need a sufficient estimate of the test reliability.

A meta-study 20 of 39 studies presented 8 studies of the repeated measurement reliability (using the intra-class correlation; ICC) of the MOCA, varying from .75 to .92. Most of these studies use small samples (<= 80), but the results are comparable for samples of normal and MCI patients. Feeney et al. 21 studied 253 older participants in a longitudinal study in Ireland and reported a reliability of .81 (ICC). Cooley et al. 22 studied healthy patients and reported low test-retest reliabilities for this group and concluded that the MOCA may be susceptible to practice effects. Another study suggests that estimates of the reliability of the MOCA may vary strongly across clinical (.75) and non-clinical populations (.5 and .63), using Cronbach’s alpha as an estimate 23. Considering the differences in sampling, the size of the samples, the different reliability indicators used and the varying results, it is difficult to draw from these results a conclusion concerning the reliability of the MoCA.

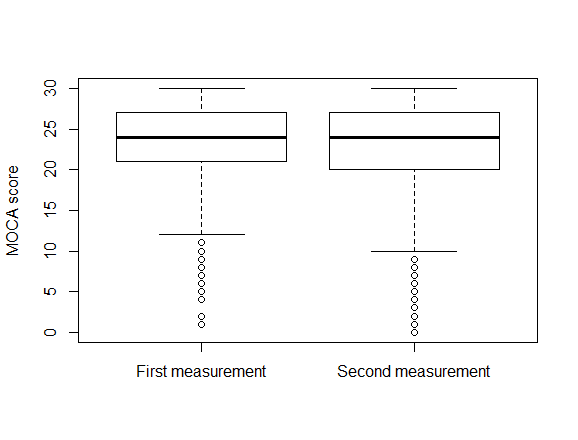
The test-retest reliability estimate used in this study is based on 2038 patients for whom both the first and second MoCA measurements are available. The time lapse between these two measurements is highly variable, with a minimum of 5 days and a maximum of 1064 days. The mode and the median of the time lapse are close to a year (median 387 days, mode 364 days). There is hardly any systematic difference between the distributions of the two measurements (Figure A1), although the second measurement is more variable, especially for those patients with lower scores. The ICC based on these two measurements is .87. 

Figure A1: Boxplot to compare the first two measurements (n = 2038).

The test-retest reliability for individuals who were re-evaluated between 335 and 395 days after initial testing offers an ICC of .88 (n = 882). For this study, the lower estimate (.87) is used as an estimate of reliability.

To reduce random effects, the standardized predictive values are calculated for a range of scores around the obtained score. The size of this range is set to (approximate) the score ±1 \* SEM. This somewhat smooths the estimates and stabilizes them. Table A1 shows the smoothed standardized predictive values for these more reliable score ranges. The estimate of SEM is 2.27 and the ranges are score ± 2. This covers a confidence level of 62.1% for the reliable test score. Reliable standardized predictive values cannot be calculated for the most extreme values (test scores 0, 1, 29 and 30) and are consequently extended from the nearest calculable value. These test scores represent the highest and lowest standardized predictive values. In practice, this should therefore rarely pose a problem for the determination of the most uncertain test scores, as classification errors are typically found around the Youden threshold (in this case 23) and not in the tails of the distributions. The ranges of test scores are indicated with subscripts in Table A1. Applying the odds of a correct classification as 2 against 1 means the minimum level of SNPV or SPPV is .667.

Table A1: Reliable Standardized Negative and Positive Predictive Values

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Score | 0 | 1 | 024 | 135 | 246 | 357 | 468 | 579 | 6810 | 7911 | 81012 |
| RSNPV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| RSPPV | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Score |  | 91113 | 101214 | 111315 | 121416 | 131517 | 141618 | 151719 | 161820 | 171921 | 182022 |
| RSNPV |  | 0 | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.07 | 0.11 | 0.16 | 0.20 |
| RSPPV |  | 1 | 1 | 0.99 | 0.98 | 0.97 | 0.95 | 0.93 | 0.89 | 0.84 | 0.80 |
| Score |  | 192123 | 202224 | 212325 | 222426 | 232527 | 242628 | 252729 | 262830 | 29 | 30 |
| RSNPV |  | 0.27 | 0.37 | 0.46 | 0.55 | 0.66 | 0.75 | 0.81 | 0.86 | 0.86 | 0.86 |
| RSPPV |  | 0.73 | 0.63 | 0.54 | 0.45 | 0.34 | 0.25 | 0.19 | 0.14 | 0.14 | 0.14 |

RSNPV: Reliable estimate of the Standardized Negative Predictive Value

RSPPV: Reliable estimate of the Standardized Positive Predictive Value

The obtained results of this method are shown in Table A2. The row “Correct classifications” represents the common NPV and PPV for the negative and positive classifications (NPV = .85; PPV = .922).

Table A2: Obtained results, using reliable predictive values and trichotomization with classification odds of 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Negative Classifications | Uncertain | Positive Classifications |
| Scores | 26-30 | 22-25 | 0-21 |
| n | 1877 | 1379 | 1763 |
| Total sample | 37.4% | 27.5% | 35.1% |
| Correct classifications | 85.0% | - | 92.2% |
| True negative status | 67.1% | 27.1% | 5.8% |
| True positive status | 10.6% | 27.8% | 61.6% |
| Realized odds | 5.7 | 1.14 | 11.8 |

The scores 22 to 25 lead to odds (d1(x) / d0(x)) of 1.14 , indicating the difficulty of basing classifications on that range of scores. Although 27.5% of all test scores are found in that range, no less than 55.3% of all errors are in that range when applying the optimal cutoff score of 23. When using another, non-optimal dichotomous cutoff score, this percentage of errors will be even larger. For instance, when using cutoff score 25 (score 25 and less indicate CI), 60.6% of all errors are found in the range 22 to 25.

Positive classifications based on scores 0 to 21 are far less error-prone, with realized odds of 11.8 (d1(x) / d0(x)). Making correct negative classifications based on the MoCA scores 26 to 30 is slightly more difficult (odds = d0(x) / d1(x) = 5.7) with an 85% rate of correct classifications. Since the same range of scores is used for negative classifications, as proposed by Nasreddine et al. 1 (test result > 25 indicate CI), the accuracy results are equal: Sp (.671) and NPV (.85) are equal. In comparison to the dichotomous cutoff score of 25, the rate of positive classifications when the true status is positive (Se = .616) is lower. The reason for this is that for the calculation uncertain scores are considered as unambiguous errors. However, considering uncertain test scores as errors is hardly realistic.

**References**

1. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 2005;53(4):695–699.

2. Coste J, Jourdain P, Pouchot J. A gray zone assigned to inconclusive results of quantitative diagnostic tests: application to the use of brain natriuretic peptide for diagnosis of heart failure in acute dyspneic patients. *Clin Chem*. 2006;52(12):2229–2235.

3. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. *Int J Epidemiol*. 2003;32(2):304–313.

4. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. *J Clin Epidemiol*. 1990;43(1):109.

5. Greiner M, Sohr D, Göbel P. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. *J Immunol Methods*. 1995;185(1):123–132.

6. Simel DL, Feussner JR, Delong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making*. 1987;7(2):107–114.

7. Landsheer JA. Interval of Uncertainty: An Alternative Approach for the Determination of Decision Thresholds, with an Illustrative Application for the Prediction of Prostate Cancer. *PloS One*. 2016;11(11):e0166007.

8. Landsheer JA. The Clinical Relevance of Methods for Handling Inconclusive Medical Test Results: Quantification of Uncertainty in Medical Decision-Making and Screening. *Diagnostics*. 2018;8(2):32. doi:10.3390/diagnostics8020032

9. Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. *Bmj*. 2012;345(e6717):1:10.

10. Shinkins B, Perera R. Diagnostic uncertainty: dichotomies are not the answer. *Br J Gen Pr*. 2013;63(608):122–123.

11. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*. 2005:73–81.

12. Gallagher EJ. The problem with sensitivity and specificity…. *Ann Emerg Med*. 2003;42(2):298–303.

13. Heston TF. Standardizing predictive values in diagnostic imaging research. *J Magn Reson Imaging*. 2011;33(2):505-505. doi:10.1002/jmri.22466

14. Heston TF. Standardized predictive values. *J Magn Reson Imaging*. 2014;39(5):1338.

15. Sonis J. How to use and interpret interval likelihood ratios. *Fam Med*. 1999;31:432–437.

16. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory.* Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887; 1986.

17. Harvill LM. Standard error of measurement. *Educ Meas Issues Pract*. 1991;10(2):33–41.

18. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2014. http://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf. Accessed May 8, 2014.

19. Landsheer JA. *UncertainInterval: Uncertain Area Methods for Cut-Point Determination in Tests*.; 2017. https://cran.r-project.org/web/packages/UncertainInterval/index.html. Accessed February 7, 2017.

20. Ozer S, Young J, Champ C, Burke M. A systematic review of the diagnostic test accuracy of brief cognitive tests to detect amnestic mild cognitive impairment. *Int J Geriatr Psychiatry*. 2016;31(11):1139–1150.

21. Feeney J, Savva GM, O’Regan C, King-Kallimanis B, Cronin H, Kenny RA. Measurement error, reliability, and minimum detectable change in the Mini-Mental State Examination, Montreal Cognitive Assessment, and color trails test among community living middle-aged and older adults. *J Alzheimers Dis*. 2016;53(3):1107–1114.

22. Cooley SA, Heaps JM, Bolzenius JD, et al. Longitudinal change in performance on the Montreal Cognitive Assessment in older adults. *Clin Neuropsychol*. 2015;29(6):824–835.

23. Bernstein IH, Lacritz L, Barlow CE, Weiner MF, DeFina LF. Psychometric evaluation of the Montreal Cognitive Assessment (MoCA) in three diverse samples. *Clin Neuropsychol*. 2011;25(1):119–126.