Adaptive Cognitive Testing in Cerebrovascular Disease and Vascular Dementia

Hans Wouters, Inge de Koning, Aeilko H. Zwinderman, Willem A. van Gool, Ben Schmand, Maarten Buiter, Robert Lindeboom

Departments of a Clinical Epidemiology, Biostatistics and Bioinformatics, and b Neurology, Academic Medical Center, and c Department of Psychology, University of Amsterdam, Amsterdam, and d Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands

Introduction

Computerized adaptive testing (CAT) [1–3] tailors a standard cognitive test to each individual patient’s ability by selecting only the items with appropriate difficulty. For the screening and measurement of decline in global cognitive ability associated with dementia, CAT has the potential to combine brevity with precision.

Currently, patients’ levels of global cognitive ability are generally estimated with the total score of a standardized cognitive test that screens a variety of cognitive domains including memory, language and executive functioning. Among the best known of such tests are the Mini Mental State Examination (MMSE) [4], and the more extensive Cambridge Cognitive Examination (CAMCOG) [5]. For screening purposes, cutoff points with optimal sensitivity and specificity are used to detect dementia [5–7]. To monitor cognitive decline, reported periodic rates of decline can be used as benchmarks [8, 9].

However, a limitation of the total score of such tests is that in order to compare patients quantitatively, every item of the test has to be administered. This makes the testing with time consuming instruments such as the CAMCOG burdensome for patients and inefficient for clinicians. Therefore, clinicians prefer the much shorter MMSE instead of the CAMCOG. However, previous re-
search has shown that the MMSE has poor accuracy in patients with mild degrees of global cognitive impairment [10] or patients with predominantly subcortical disturbances, while the CAMCOG was shown to be more precise [7, 11]. Thus, longer instruments with desired precision such as the CAMCOG lack the efficiency of brief but less precise cognitive tests such as the MMSE.

CAT provides a potential solution to this problem. It combines brevity with precision in assessing global cognitive ability. CAT is enabled by the Rasch probabilistic theory of measurement. In the basic Rasch model, the probability of responding correctly is regressed onto an estimated dimension of global cognitive ability on which both patients on the basis of their ability level and items on the basis of the level of ability that they measure (briefly their ‘difficulty’ level) are hierarchically arranged. Individuals with higher ability will have a higher chance of correctly responding to any item, whereas an easy item will probably be answered correctly by anyone, regardless of the severity of the cognitive deterioration. Using this assumption and contrary to the total score approach, CAT does not need to administer every item of a long instrument to generate a precise estimate of a patient’s ability. Rather, it estimates a patient’s ability during testing. It selects a more difficult item if the patient has responded correctly to the previous item, and it selects an easier item if the patient has responded incorrectly. In doing so, CAT is likely to reduce the testing time substantially. By zooming in on a patient’s ability, a reduction of the testing burden and annoyance in patients is also likely to be achieved as CAT avoids the administration of too easy and too difficult items. At the same time CAT maintains the precision of a longer more precise instrument such as the CAMCOG.

Our previous analysis of CAMCOG data supported this potential of CAT. Even though there was a reduction of about 60% in number of items needed, the abilities estimated by CAT were in excellent agreement with the patient abilities estimated with the whole CAMCOG and had similar diagnostic accuracy [12]. However, these findings were obtained in patients with primarily Alzheimer’s disease (AD). As vascular dementia (VaD) is the second cause of dementia in the industrialized world and cerebrovascular disease (CVD) is associated with cognitive impairment [13], it is also important to know whether the previous results can be generalized to patients with CVD and VaD.

In this study, we examine whether the CAT CAMCOG is also valid in VaD and CVD patients. Compared to AD patients, these patients have a distinct profile of cognitive impairment. We therefore compared the precision and validity of the CAT developed in AD patients with that of a CAT we developed in the present study based on data of VaD and CVD patients. Both CATs consisted of items from the CAMCOG. Specifically, we hypothesized that when applied to CVD patients, the performance of the CAT developed for AD patients would be as good as the CAT developed in VaD and CVD patients.

Subjects and Methods

Participants
This was a simulated CAT algorithm using existing data from patients that were recruited from the Rotterdam Stroke Data bank, a registry of patients with a transient ischemic attack (TIA), ischemic stroke or intracerebral hemorrhage who were admitted to the Department of Neurology of the University Hospital of Rotterdam, The Netherlands [7]. Inclusion and exclusion criteria and the research procedure are described elsewhere [7, 14, 15]. In brief, patients were enrolled if they were 55 years of age or older, had no conditions compromising routine testing (e.g. aphasia, neglect) and if they consented to participate. All patients were examined with the CAMCOG. The diagnosis of VaD was based on the NINDS-AIREN diagnostic criteria [16] that were applied without using the CAMCOG results.

Informed consent was obtained from all patients or from close relatives in case of impaired judgment in the patient. The local medical ethics committee approved the study.

Instruments
The CAMCOG [5] has 67 items including those of the MMSE (maximum total score 107) to assess cognitive ability in the domains of orientation, memory, language, attention/executive functioning, praxis, calculation, abstract thinking and perception. Studies have shown the CAMCOG to be more effective than the MMSE in detecting mild dementia [11] and dementia in patients with a TIA or stroke [7, 14].

Computerized Adaptive Testing and Rasch Analysis
The CAT algorithm was developed to tailor the standard test to each individual patient by selecting for each individual patient only those items with difficulties appropriate for that patient’s global cognitive ability. This algorithm was based on a set of CAMCOG items (the C-47) of which the difficulties were previously estimated with data from the AMSTEL study, a population based study with primarily AD patients [17] using the one-parameter logistic model (OPLM), a Rasch type of model.

In brief, the OPLM regresses the probability to respond correctly onto a single dimension. The unit of measurement of this dimension is the log-transformed odds of responding correctly (log odds). It is a dimension of global cognitive ability as all CAMCOG items are arranged on this dimension. The OPLM arranges both the CAMCOG items and individuals to whom the CAMCOG was administered on this dimension. It estimates the level of ability as measured by the item (the item difficulty) and the ability level as possessed by the individual. The probability of responding correctly is a function of the item difficulty relative to
the individual’s ability level. A highly able individual will respond correctly to any item, even the most difficult ones, whereas an easy item will be correctly responded to by even less able individuals. In addition, the OPLM weights every item for its capability to discriminate between patients with different levels of global cognitive ability. Validity analyses to test the unidimensionality assumption implemented in the OPLM software showed that the difficulties of 47 CAMCOG items, briefly denoted as the C-47, could be validly estimated [17]. Using these items, CAT was subsequently used to estimate the CVD and VaD patient C-47 total scores. The CAT program presented six initial items to provisionally estimate a patient’s ability along with a standard error. They were selected in the following order:
- item 138 ‘naming/encoding pictures’,
- item 139 ‘animal fluency’,
- item 146 ‘delayed recall of encoded pictures of item 138’,
- item 171 writing an address,
- item 147 ‘recognition of objects item 138’,
- item 178 ‘recall written address of item 171’.

This was done because a standard CAT does not select recall items (items 146, 147 and 178) after the items containing the previously encoded stimuli (items 138 and 171). To better resemble actual testing, the above sequence was therefore used. Subsequently, the CAT updated the ability estimate and the standard error after each response made by the patient by selecting an item with a difficulty that was nearest to the patient’s provisionally estimated ability. Thus, a correct response to an item yielded the presentation of a more difficult item, while an incorrect response yielded the presentation of an easier item, and so on. With the administration of each subsequent item, the estimate of the patient’s global cognitive ability level became more precise (a smaller standard error) thereby increasing the adequacy of the item selection. The estimation process of a patient’s ability terminated (a) when the standard error reached a value of 0.15 log odds, which reflects a high level of precision [18] or (b) when the number of items reached a predetermined item maximum. Four item maxima (all including the six initial items) of 40, 35, 30 or 25 items were examined, which reflect an item reduction of at least 33, 42, 50 and 58% compared to the original CAMCOG.

Statistical Analyses
The following analyses were performed to examine whether the adaptive version of the C-47 can be used instead of the C-47 or the original CAMCOG. To facilitate the interpretation of the results, the OPLM ability estimates expressed in log odds were rescaled on a more familiar metric ranging from 0 ‘low ability’ to 100 ‘high ability’. Intraclass correlations were calculated to examine the agreement of the CAT estimated scores with the actual total C-47 scores. Pearson’s correlations were calculated to examine the association between the CAT estimated scores and the total scores from the original CAMCOG.

The bias, i.e. overestimation and underestimation of the total score by CAT, was examined by plotting the difference between the CAT estimated C-47 total scores and the actual C-47 total scores against the actual C-47 total scores. Pearson’s correlations were calculated to examine the magnitude of the bias. In the absence of bias the differences between the CAT estimated scores and C-47 total scores are not systematically higher or lower than the C-47 total scores and a significant association is absent.

Areas under the curve (AUCs) were calculated to compare the diagnostic accuracy of the CAT estimated scores and C-47 total scores as well as the total score of the original CAMCOG using the clinical diagnosis of VaD as the reference standard. Subtypes of the clinical diagnosis of dementia were AD with CVD, possible VaD and probable VaD. The sensitivity and specificity of the C-47 and CAT were compared with those of the original CAMCOG. Optimal sensitivity and specificity values were those values where the sum of sensitivity and specificity estimates was maximal.

Each of the analyses was repeated for the four predetermined maxima of items to be used by the CAT. Subsequently, we compared the CAT developed for AD with an alternative CAT developed in this sample of VaD and CVD patients with CAMCOG item difficulties based on the responses of this sample (the ‘alternative CAT’). For that goal, the goodness of fit of the CAMCOG items to the OPLM model was examined. Opposite to the usual direction to find significant p values (e.g. <0.05), the null hypothesis in the OPLM analysis is defined as the absence of misfit and good fitting items have p values >0.05. For the original C-47 based CAT to be useful as a general screener, it should have at least the same precision as the newly developed alternative CAT. We therefore replicated each of the above analyses for the alternative CAT and compared the results with those of the original CAT administered C-47.

Results
Demographic and Clinical Characteristics
Of the 284 patients, 55 (19%) were diagnosed with dementia, 6 (2%) participants had mild cognitive disturbances that did not meet the criteria for dementia (cognitive impairment no dementia, CIND). Of the demented patients 35 (64%) had probable VaD, 5 (9%) had possible VaD, and 15 (27.3%) had AD with concomitant CVD. The percentage of women was higher in the group of demented patients than in the group of non-demented patients (55 vs. 37%). On average, the demented patients were less educated (mean years of education: 7.5, SD 2.9) than the non-demented group of patients (mean years of education: 9.0, SD 3.0) and older (mean age: 73.6, SD 7.24) than the non-demented (68.6 years, SD 8.0) (table 1).

Test Reduction
With CAT there was substantial test reduction in all of the four conditions of predetermined maxima of items needed with CAT, as often sufficient reliability in the estimation of ability was reached before the predetermined maximum of items. Compared to the C-47 and the original CAMCOG (between parentheses), the mean percentage reduction in number of items was 36.4% (50.2%) for 40 items, 41.2% (53.9%) for 35 items, 46.4% (58.0%) reduction for 30 items and 52.1% (62.5%) for 25 items.
Correlations and Agreement

Even though there was a substantial test reduction, the CAT estimated C-47 total scores were in very high agreement with the C-47 total scores, irrespective of the four predetermined maxima of items as indicated by intraclass correlation: 0.98 (40 items), 0.98 (35 items), 0.98 (30 items), 0.97 (25 items). The CAT estimated C-47 total scores also correlated substantially with the total scores of the original CAMCOG being 0.94 for the four maxima of items.

Figure 1 displays the relationship of the CAT estimated total scores and the original CAMCOG scores. The dotted line represents the total score of the original whole CAMCOG as predicted by the CAT tailored item selection ($R^2 = 0.88$).

The scatter plots of the differences between the CAT estimated scores and C-47 scores (y) and C-47 scores (x) did not show bias of the CAT estimated total scores in the four predetermined maxima of items as indicated by small and insignificant correlations (r): 40 items ($r = -0.06$, $p = 0.30$), 35 items ($r = -0.06$, $p = 0.31$), 30 items ($r = -0.05$, $p = 0.38$) and 25 items ($r = -0.03$, $p = 0.60$). Furthermore, the deviations of the CAT estimated total C-47 scores from those of the C-47 were only minor with 95% of the deviations being at most 10% of the C-47 range which is 87 points.

Diagnostic Accuracy

ROC curve analysis (fig. 2) showed that the diagnostic accuracy as indicated by the AUC was slightly lower for the C-47 (0.94) and CAT (all four maxima being 0.94) than for the original CAMCOG (0.96). The sensitivity of the C-47 and CAT were slightly lower than for the original CAMCOG, while the specificity was in the same range. Optimal sensitivity versus specificity was 0.89 versus 0.92 for the original CAMCOG and 0.84 versus 0.94 for the C-47. Across the four predetermined maxima of items, the optimal sensitivity versus optimal specificity was constantly 0.82 versus 0.92 for a cutoff point of 76/77 points which translates to a predicted original CAMCOG score of 78/79 (fig. 1). Repeating the analysis without the patients who had AD with CVD did not change the results importantly. The AUCs ranged from 0.92 to 0.93 for the four maxima of the CAT, the AUC for the C-47 was 0.93 and 0.95 for the original CAMCOG.

Alternative CAMCOG Item Set

Sixty-five CAMCOG items were available for analysis. We excluded dichotomous items that were only (in)correctly responded to (Q134, Q140, Q176) or items with 97% of the responses in one of the categories (Q125, Q126, Q130, Q137, Q141, Q150, Q162, Q172). Fifty-four items were therefore considered for analysis. Polytomous items were dichotomized. Missing values were treated as incorrect responses, as suggested by Fillenbaum et al. [19]. The amount of missing values per item was 6.5% at most.

Seven items (of 54) had misfit to the OPLM ($p < 0.05$) (Q122, Q135–136, Q154, Q158, Q163, Q183) and were excluded. The remaining 47 items were selected to form the alternative CAT. Eleven items were present in the alternative CAT that were absent in the original C-47 (Q142, Q143, Q152, Q161, Q164, Q168, Q169, Q179, Q180, Q181, Q185) and 11 items were absent in the alternative CAT that were present in the C-47 (Q122, Q126, Q130, Q135, Q136, Q141, Q150, Q154, Q162, Q163, Q183).

The figure of test reduction by the CAT based on the alternative CAT was the same as it ranged from 35.0 to 49.2% compared to the C-47 and from 49.1 to 60.2% compared to the original CAMCOG. The intraclass correlations (all 0.98) were in the same range while the correlations were somewhat higher (all 0.96). The scatter plots did not show bias for the four item maxima as indicated by small and insignificant correlations for the three highest item maxima: 40 items ($r = 0.10, p = 0.09$), 35 items ($r = 0.11, p = 0.07$), 30 items ($r = 0.11, p = 0.07$) and a significant but small correlation for the maximum of 25 items ($r = 0.12, p = 0.05$). Again the deviations were minor as 95% of the deviations were at most 9% of the total range of the scale and this figure was consistent for the four predetermined maxima of items. AUCs were also similar: 0.94 for the alternative CAMCOG item set and for the alternative CAT being 0.94 across the predetermined maxima of items.

Table 1. Clinical and demographic characteristics

<table>
<thead>
<tr>
<th></th>
<th>Study participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>demented (n = 55)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>30 (55)</td>
</tr>
<tr>
<td>CIND, n (%)</td>
<td>15 (27.3)</td>
</tr>
<tr>
<td>Dementia subtype, n (%)</td>
<td>5 (9.1)</td>
</tr>
<tr>
<td>AD + VaD</td>
<td>35 (63.6)</td>
</tr>
<tr>
<td>Possible VaD</td>
<td></td>
</tr>
<tr>
<td>Probable VaD</td>
<td></td>
</tr>
<tr>
<td>CAMCOG score M (SD; range)</td>
<td>63.2 (14.1; 65)</td>
</tr>
<tr>
<td>MMSE score M (SD; range)</td>
<td>17.6 (4.1; 20)</td>
</tr>
<tr>
<td>Education, years M (SD; range)</td>
<td>7.6 (2.9; 15)</td>
</tr>
<tr>
<td>Age, years M (SD; range)</td>
<td>73.6 (7.2; 32)</td>
</tr>
</tbody>
</table>

Table 1. Clinical and demographic characteristics

<table>
<thead>
<tr>
<th></th>
<th>Study participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>demented (n = 55)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>30 (55)</td>
</tr>
<tr>
<td>CIND, n (%)</td>
<td>15 (27.3)</td>
</tr>
<tr>
<td>Dementia subtype, n (%)</td>
<td>5 (9.1)</td>
</tr>
<tr>
<td>AD + VaD</td>
<td>35 (63.6)</td>
</tr>
<tr>
<td>Possible VaD</td>
<td></td>
</tr>
<tr>
<td>Probable VaD</td>
<td></td>
</tr>
<tr>
<td>CAMCOG score M (SD; range)</td>
<td>63.2 (14.1; 65)</td>
</tr>
<tr>
<td>MMSE score M (SD; range)</td>
<td>17.6 (4.1; 20)</td>
</tr>
<tr>
<td>Education, years M (SD; range)</td>
<td>7.6 (2.9; 15)</td>
</tr>
<tr>
<td>Age, years M (SD; range)</td>
<td>73.6 (7.2; 32)</td>
</tr>
</tbody>
</table>
Discussion

We found that with CAT, it may be possible to administer the CAMCOG much more efficiently, i.e. precise and less time consuming. A reduction in number of items is possible up to 50% compared to the C-47, i.e. 47 CAMCOG items selected with Rasch analysis [17] and up to 60% compared to the original CAMCOG, without losing much reliability and diagnostic accuracy. The CAT estimated scores were in high agreement with the C-47. Furthermore, the CAT estimated scores also correlated substantially with the original CAMCOG. Using the CAT
estimates, the original CAMCOG total scores could be predicted well. Also, the cutoff point with optimal sensitivity and specificity that we found is consistent with the optimal cutoff point as previously found [15] and differs only in one point from that reported originally [5]. Furthermore, the overall accuracy of the CAT was slightly lower than the accuracy of the CAMCOG [15] and the accuracy of the memory section of the CAMCOG in the detection of AD [6].

These findings are consistent with our previous findings concerning the CAT administration of the C-47 in a sample of AD patients [12], and are important as they pertain to patients with a different etiology, i.e. vascular cognitive impairment and VaD. As this etiology is presumed to be the second most important cause of dementia and given that dementia at older ages is often characterized by a mixture of AD and CVD, our results may be clinically relevant. In CVD patients, the CAT developed on AD patients performed in the same range of precision and accuracy as the alternative vascular cognitive impairment specific CAT developed in the current sample. This makes the CAT version of the C-47 promising as a general screen of cognitive impairment due to AD and VaD. More research is needed to replicate the findings in samples with other forms of dementia.

One could argue that both studies of the utility of CAT are limited by the fact that they were conducted only on previously collected data, thus they do not really represent a test of the CAMCOG administered with CAT. However, retrospective validation is needed to provide a sound rationale for using CAT prospectively in new patients series in clinical research. The findings of this study fulfill a sensible prerequisite for prospective studies, which may subsequently document sufficient evidence for introducing the CAT in clinical practice.

Another apparent limitation of the findings in our studies is that they were obtained with the CAMCOG, which is a measure of global cognitive impairment. The use of global measures in patients with CVD seems to contradict the focal nature of the disease at least in stroke patients. Indeed, the use of the total scores of any test version, whether adaptive or not, presupposes a single dimension of global cognitive ability or impairment. However, the fact that every test version had satisfactory accuracy negates such doubts. Also, patients with severe focal deficits, like severe aphasia or neglect, were excluded. Therefore, our sample may not be representative for all stroke patients. Furthermore, support for the presupposition of global cognitive impairment also comes from findings showing that in addition to the total CAMCOG score, the domain specific scores of memory and language are substantially correlated with global brain damage [20].

One might criticize the small number of patients with mild degrees of cognitive impairment if the objective of the CAT is to screen for milder cases. However, 20 of the demented patients and 50 of the nondemented participants were mildly impaired defined as a CAMCOG total score ranging from average to 1 SD below average.

A difficulty of using CAT in cognitive testing are the delayed recall items of previously encoded stimuli usually naming items. In our previous validation [12], we let the CAT select five initial items randomly, one from each quintile of the distribution of item difficulties after which the CAT estimated a provisional ability level. However, this procedure may result in the selection of a ‘recall item’ where the patient is expected to have reproduced a list of stimuli presented earlier, whereas this did not happen.

We therefore decided to have the delayed recall items be administered as initial items following the naming items. As distractor items, we used an item to assess executive functioning (Q160), and another one to assess praxis (Q171). Together the initial items measured at least language, praxis, memory and executive functioning to have the most important cognitive domains assessed.

Finally, it should be stressed that the use of CAT pertains to cognitive screening and monitoring rather than to in-depth neuropsychological testing for diagnostic purposes. The results of our study show that adaptive cognitive testing can be used reliably in the screening for global cognitive impairment in patients with CVD, but the use of a screening instrument should never be a replacement of a formal neuropsychological examination.

In conclusion, our results support the potential of CAT to combine efficiency with precision in the screening and measurement of decline in global cognitive ability associated with CVD and dementia.

Acknowledgements

We would like to thank the Dutch Council for Educational Research and Prof. Verhelst for making the OPLM software available to us. Funded with a grant from the Internationale Stichting Alzheimer Onderzoek (ISAO).
References


