THE MINI-COG: RECEIVER OPERATING CHARACTERISTICS WITH EXPERT AND NAÏVE RATERS

JAMES SCANLAN1,2* AND SOO BORSON1,2
1Alzheimer’s Disease Research Center, University of Washington, Seattle, WA, USA
2Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

ABSTRACT

Background. As elderly populations grow, dementia detection in the community is increasingly needed. Existing screens are largely unused because of time and training requirements. We developed the Mini-Cog, a brief dementia screen with high sensitivity, specificity, and acceptability. Here we describe the development of its scoring algorithm, its receiver operating characteristics (ROC), and the generalizability of its clock drawing scoring system.

Sample and methods. A total of 249 multi-lingual older adults were examined. Scores on the three-item recall task and the clock drawing task (CDT-CERAD version) were combined to create an optimal algorithm. Receiver operating characteristics for seven alternatives were compared with those of the MMSE and the CASI using expert raters. To assess the CDT scoring generalizability, 20 naïve raters, without explicit instructions or prior CDT exposure, scored 80 randomly selected clocks as ‘normal’ or ‘abnormal’ (20 from each of four CERAD categories).

Results. An algorithm maximizing sensitivity and correct diagnosis was defined. Its ROC compared favorably with those of the MMSE and CASI. CDT concordance between naïve and trained raters was >98% for normal, moderately and severely impaired clocks, but lower (60%) for mildly impaired clocks. Recalculation of the Mini-Cog’s performance, assuming that naïve raters would score all mildly impaired CDTs in the full sample as normal, retained high sensitivity (97%) and specificity (95%).

Conclusion. The Mini-Cog algorithm performs well with simple clock scoring techniques. The results suggest that the Mini-Cog may be used successfully by relatively untrained raters as a first-stage dementia screen. Further research is needed to characterize the Mini-Cog’s utility when population dementia prevalences are low. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS — dementia screening; MMSE; clock drawing; three-item recall; Cognitive Abilities Screening Instrument; naïve raters; language

INTRODUCTION

In a recent survey of 368 general practitioners, 82% believed that dementia screening of elderly persons is worthwhile, but only 24% routinely screened their own patients (Bush et al., 1997). The major barriers were lack of time (85%), fear of offending patients (58%) and inadequacies in available tests (22%). When a formal cognitive screen was used, it was usually the MMSE (Bush et al., 1997); if a short, simple screen such as clock drawing could be shown to be effective, 93% would use it.

Persuading doctors to adopt routine dementia screening thus appears to require a more efficient approach. The ideal instrument would be very brief, simple, sensitive, acceptable to older persons, and uninfluenced by low education and language barriers that weaken the utility of most available screens. The Mini-Cog is a new instrument that combines a simple memory test with clock drawing (CDT) and appears to fulfill these criteria. When used by expert raters, the Mini-Cog attained 99% sensitivity and 93% specificity in a split sample containing 50% demented persons, and required only 3 minutes to administer (Borson et al., 2000). However, it is unclear whether novice raters unfamiliar with cognitive testing will accurately...
give the Mini-Cog in ‘real world’ settings. Its structure is simple enough to be used by community volunteers and health care personnel, provided its most vulnerable element, the CDT, can be scored correctly.

Many CDT dementia screening systems have been proposed, with varying degrees of complexity (Watson et al., 1993; Shulman et al., 1986; Dastoor et al., 1991; Tuokko et al., 1992; Death et al., 1993). While all studies report the sensitivity and specificity of their proposed systems, cost–benefit, measured in ease of use, is rarely examined. If a scoring system slightly improves sensitivity or specificity but increases costs in terms of complexity, interpretation and training, then many CDT advantages are lost. As analog clock form and function are highly learned in most modern cultures, we believe that understanding how to judge CDTs as ‘normal’ and ‘abnormal’ is intuitively and widely understood. Therefore, if naïve raters gave their best estimates without an explicit scoring system, their judgments should closely resemble expert decisions. If true, this could greatly facilitate CDT use as in community-based dementia screening efforts.

Similar reasoning has been applied to the MMSE, and some studies show item subsets to perform as well as the complete scale. In one study, the four MMSE items that best discriminated normal controls from mild AD patients were two time orientation items (day, date) and two of the three conventional word recall items (‘apple’ and ‘penny’). Dementia discrimination using these items alone was comparable to that of the full MMSE (Fillenbaum et al., 1994). The MMSE three-item recall task was highly predictive of total score: 97% of subjects recalling 0/3 words had MMSE scores below the conventional cut-off point for screening (≤23), while 86% of those recalling 1–3 words scored ≥24 (Braekhus et al., 1992). In five different studies designed to distill the essential MMSE items (Magaziner et al., 1987; Roca, 1987; Klein et al., 1985; Braekhus et al., 1992; Fillenbaum et al., 1994), only three-item recall was consistently implicated in both predicting total score and distinguishing normal from mildly demented subjects. Additionally, we have previously reported (Borson et al., 2000) that three-item recall by itself was a sensitive (91%) and specific (97%) dementia discriminator. Accordingly, we combined this short learning task with the CDT to yield a dementia screen (the Mini-Cog) more sensitive than three-item recall alone and relatively uninfluenced by education or language (Borson et al., 2000). In this paper, we report the receiver operating characteristics (ROC) of different Mini-Cog scoring algorithms, and compare them with the MMSE and CASI. We also examine the accuracy of CDT scoring by untrained raters without scoring instruction, and whether Mini-Cog subject classification would be compromised by naïve rater scoring errors.

METHODS

Subjects

Full details of sample selection, subject characteristics, and diagnostic evaluations are provided elsewhere (Borson et al., 1999, 2000). Briefly, 249 community dwelling older adults (173 women, 76 men) reflecting the five major ethnic groups in the United States completed the CDT, MMSE, and CASI during initial evaluation. All subjects or their proxies gave written informed consent using a multiple-language protocol approved by the University of Washington IRB.

Subjects were initially classified as ‘probably demented’ (n=129) or ‘probably not demented’ (n=120) based on an informant’s history of cognitive decline (CERAD expanded history) and current functioning (Clinical Dementia Rating) (Hughes et al., 1982), an approach analogous to that of Jorm et al. (1991). The classification of all subjects was subsequently confirmed using formal diagnostic criteria (CERAD, DSM-IV: American Psychiatric Association, 1994; and NINCDS-ADRDA: McKhann et al., 1984), and demented subjects were further diagnosed as having probable or possible AD, another dementia, or no dementia. Subjects with uncertain/very mild cognitive impairment (CDR=0.5) were excluded. Post hoc dementia diagnoses were probable AD in 92 (71%), possible AD (mixed states, usually vascular brain disease in addition to AD) in 16 (12%), vascular dementia in 13 (10%), and other dementias in 8 (6%).

The CDT was scored using CERAD templates (Borson et al., 1999). Subjects were instructed to draw a large circle, fill in the numbers on a clock face, and set the hands at 8:20, with no time limit. Excellent inter-rater CDT reliability was obtained (intra-class correlation=0.97) by two independent raters blind to each other’s scores and to any subject information using the four-point CERAD scale (0, 1, 2, 3: 0=normal; 3=severe impairment).
While the three-item recall task can be given in both cued and uncued fashion, research suggests that uncued recall is sensitive to dementia in AD patients, and that cueing does not improve scores or dementia detection (Yuspeh et al., 1998). Accordingly, we used uncued recall in this data analysis. One point was given for each word correctly recalled.

Clock scoring by naïve raters

We randomly sampled 20 CDTs from each of the four CERAD performance levels (0=normal, 1=mild, 2=moderate, 3=severe) for which exact agreement was reached by the two expert raters. The 80 CDTs were scrambled in random sequence. Twenty volunteers with no prior cognitive testing experience were recruited from hospital clinics, social service agencies, and administrative services. Naïve raters consisted of 9 (45%) administrative/clerical workers, 7 (35%) clinical/medical employees, and 4 (20%) community case managers. These raters averaged 39±10 years of age, with 16±2 years of education. All scored each CDT as normal or abnormal without hints or templates to guide them.

Data analyses

A series of Mini-Cog scoring algorithms was tested using various combinations of three-item recall and the CDT, to construct an ROC curve that was superimposed upon ROC curves generated for the MMSE and CASI using all possible cut-off points. In the formulation of Mini-Cog algorithms, preference was given to robust and simple combinations. The best available three-item/CDT combination was assessed against the conventional MMSE and CASI conventional cut-off points (MMSE=23/24 of 30 possible points (Folstein et al., 1975); CASI=80/81 of 100 possible points (Teng et al., 1994). Judgements of CDTs by naïve raters were compared with those of expert raters. Because naïve raters were simply asked to judge the CDTs as ‘normal’ or ‘abnormal’ and experts used the entire CERAD scale (0–3) we defined agreement as ‘normal’=CERAD 0 and ‘abnormal’=CERAD 1–3.

RESULTS

Algorithm construction

Cross-tabulations of three-item recall scores and CERAD CDTs were examined as the basis for constructing a classification algorithm when both tests were combined into a single instrument (Table 1). Various decision algorithms were constructed (see Table 2). Algorithm numbers correspond to those represented on the ROC curve as single points (Fig. 1) for comparison with the MMSE and CASI (represented as continuous curves). Results show that at least one recall/CDT algorithm had superior sensitivity relative to the CASI and MMSE. The conventional cut-off points for the

Table 1. Correct classification of subjects by three-item recall and CDT scores (frequency and % shown)

<table>
<thead>
<tr>
<th>CERAD CDT score</th>
<th>Number of words correctly recalled</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>ND (freq.)</td>
</tr>
<tr>
<td></td>
<td>DEM (freq.)</td>
</tr>
<tr>
<td>1</td>
<td>ND (freq.)</td>
</tr>
<tr>
<td></td>
<td>DEM (freq.)</td>
</tr>
<tr>
<td>2</td>
<td>ND (freq.)</td>
</tr>
<tr>
<td></td>
<td>DEM (freq.)</td>
</tr>
<tr>
<td>3</td>
<td>ND (freq.)</td>
</tr>
<tr>
<td></td>
<td>DEM (freq.)</td>
</tr>
</tbody>
</table>

CDT: 0=normal, 1=abnormal; ND=probably not demented (psychiatric judgement); D=probably demented (psychiatric judgement). Bold numbers indicate clearly demented (by Mini-Cog algorithm 7). Italic numbers indicate demented by Mini-Cog (would be classified as non-demented by naïve CDT raters). The remainder are clearly non-demented (by Mini-Cog).
MMSE and CASI corresponded relatively well with their points of optimal function (the best balance of sensitivity and specificity) in this sample.

When the performance of algorithm 7 (Table 3, henceforth referred to as the Mini-Cog) was compared with the other dementia screens (three-item alone, CDT alone, CASI, and MMSE), it performed best in sensitivity (99%) and diagnostic value (96%) and had acceptable specificity (93%), though it was less efficient in classifying non-demented subjects than the CASI (96%). The MMSE was the least sensitive (91%) and specific (92%) (Borson et al., 2000). When separate logistic regression analyses predicting correct diagnosis and controlling for the effects of education and non-native language (English/non-English) were performed the results were striking. The Mini-Cog was most predictive ($\chi^2 = 273$), followed by the CASI ($\chi^2 = 233$) and the MMSE ($\chi^2 = 201$; see Borson et al., 2000, for detailed description).

**DISCUSSION**

We have previously reported that the combination of the three-item recall task with the CDT results in superior dementia prediction relative to either of its parts and to the CASI and MMSE in a very heterogeneous community sample (Borson et al., 1999). Inspection of the ROC curves shown here further explicates those findings. First, the conventional cut-off points for the MMSE (23/24) and CASI (80/81) worked as expected in heterogeneous samples of older adults, producing the best sensitivity/specificity combination attainable with these instruments in this sample. Examining our algorithms (Table 2 and Fig. 1) we see that numbers 1–3 tend to maximize sensitivity, while 4–6 maximize specificity and 7 (the Mini-Cog) is a weaker for category 1 (60%) (Table 3). The 40% of naïve ratings discordant for mildly impaired CDTs, as judged by research raters, were uniformly rated as normal. Overall concordance between naïve and expert raters was 89%.

We evaluated the projected Mini-Cog performance as if it had been administered by naïve raters to our entire subject sample. We made the following assumptions: that the naïve raters properly administered the three-item recall task, that subjects received appropriate CDT instructions, that all mildly impaired CDTs were scored normal (rather than the 40% found here), and all other CDTs were scored identically to research ratings (rather than the 98–99% found here). Under these circumstances, Mini-Cog sensitivity would decrease from 99 to 97%, and specificity would increase from 93 to 95%, both well within effective screening ranges. The percentage of subjects with mildly impaired CDTs in this sample was small (10%), and 67% of these recalled 0/3 words.

**CDT classification by naïve subjects**

CDT judgments by naïve raters (normal or abnormal) were compared with expert judgments using CERAD categories (0, 1, 2, 3). Concordance between naïve and expert raters was extremely high for CERAD categories 0, 2 and 3 (>98%) and weaker for category 1 (60%) (Table 3). The 40% of naïve ratings discordant for mildly impaired CDTs, as judged by research raters, were uniformly rated as normal. Overall concordance between naïve and expert raters was 89%.

We evaluated the projected Mini-Cog performance as if it had been administered by naïve raters to our entire subject sample. We made the following assumptions: that the naïve raters properly administered the three-item recall task, that subjects received appropriate CDT instructions, that all mildly impaired CDTs were scored normal (rather than the 40% found here), and all other CDTs were scored identically to research ratings (rather than the 98–99% found here). Under these circumstances, Mini-Cog sensitivity would decrease from 99 to 97%, and specificity would increase from 93 to 95%, both well within effective screening ranges. The percentage of subjects with mildly impaired CDTs in this sample was small (10%), and 67% of these recalled 0/3 words.

**DISCUSSION**

We have previously reported that the combination of the three-item recall task with the CDT results in superior dementia prediction relative to either of its parts and to the CASI and MMSE in a very heterogeneous community sample (Borson et al., 1999). Inspection of the ROC curves shown here further explicates those findings. First, the conventional cut-off points for the MMSE (23/24) and CASI (80/81) worked as expected in heterogeneous samples of older adults, producing the best sensitivity/specificity combination attainable with these instruments in this sample. Examining our algorithms (Table 2 and Fig. 1) we see that numbers 1–3 tend to maximize sensitivity, while 4–6 maximize specificity and 7 (the Mini-Cog) is a weaker for category 1 (60%) (Table 3). The 40% of naïve ratings discordant for mildly impaired CDTs, as judged by research raters, were uniformly rated as normal. Overall concordance between naïve and expert raters was 89%.

We evaluated the projected Mini-Cog performance as if it had been administered by naïve raters to our entire subject sample. We made the following assumptions: that the naïve raters properly administered the three-item recall task, that subjects received appropriate CDT instructions, that all mildly impaired CDTs were scored normal (rather than the 40% found here), and all other CDTs were scored identically to research ratings (rather than the 98–99% found here). Under these circumstances, Mini-Cog sensitivity would decrease from 99 to 97%, and specificity would increase from 93 to 95%, both well within effective screening ranges. The percentage of subjects with mildly impaired CDTs in this sample was small (10%), and 67% of these recalled 0/3 words.
The four-category CERAD system we used, while screening cut-offs were set at 3 of 4 possible points. The essence of dementia for screening purposes. Combining these two elements seems to capture the presentation, global attention, and executive functions. Including long-term memory, visuo-spatial representation, global attention, and executive functions. Combining these two elements seems to capture the essence of dementia for screening purposes.

Virtually all CDT studies show that greater impairment in clock drawing is associated with more severe dementia and lower MMSE scores (Death et al., 1993; Brodaty and Moore, 1997; Juby, 1999). However devised, the cut-offs for these systems are usually set so that a small number of errors (10–25%) is consistent with non-demented status. For example, in the 10-point Shulman scale (Shulman et al., 1986), a score of 8 was consistent with clinically non-demented status, while a score of ≤7 indicated possible dementia. With an alternative four-point system tested by Lam in a Chinese sample (Lam et al., 1998), dementia screening cut-offs were set at 3 of 4 possible points. The four-category CERAD system we used, while simpler than some alternative systems, uses a comparable cut-off point (0=normal, 1–3=demented). The CDT judgments by naïve raters were strikingly similar to those of expert raters for CERAD categories 0 (normal), 2 (moderate) and 3 (severe). The lower sensitivity of naïve than expert raters to mild CDT impairment (CERAD category 1) appeared unlikely to compromise Mini-Cog performance, even if failure to detect mild impairment was assumed to be complete. Using the full Mini-Cog algorithm, most would be classified as demented, even if all their CDTs were scored as ‘normal’, because the majority of subjects with mild CDT impairment (67%) scored poorly on three-item recall (0 of 3 words recalled) and would be classified as demented on that basis alone. This is consistent with our general observation that most of the work of dementia discrimination is being performed by the three-item recall task, with the CDT acting to supplement and refine these judgements in less obvious cases. This is also consistent with the idea that three-item memory task is tapping into the first and most powerful factor in MMSE factor analyses. Additionally, we note that while the interpretation errors of mild CDT problems by naïve raters would decrease Mini-Cog sensitivity, it has the simultaneous effect of increasing specificity, resulting in no net change in overall correct detection. In an unselected population sample, we would predict that CDT-naïve raters using the Mini-Cog would have slightly lower sensitivity, slightly higher specificity, and very similar overall correct detection relative to more expert raters. In short, while naïve raters were less likely to identify mildly impaired CDTs as ‘abnormal’, we think that this would have minimal effects on overall Mini-Cog accuracy.

These results suggest that the Mini-Cog’s ‘minimalist’ approach could be successfully used in settings where time is short, trained personnel are lacking, and/or language barriers exist. More lengthy and complex methods may be counter-productive, if physicians and other staff are unlikely to use them. The length and complexity of alternative instruments tends to defeat the goal of cost-effective dementia detection (Borson et al., 2000), while the Mini-Cog speed and simplicity may promote it.

As is true for all first-stage screening tests, individuals who screen positive on the Mini-Cog cannot be presumed demented without additional evidence. Screening unselected elderly populations will result in increases in false-positive tests (lower specificity and diagnostic value) relative to the Mini-Cog developmental sample, which was enriched for probable cases. If we were to extrapolate the effects of using naïve raters in a population with substantially fewer impaired subjects, more false
negative and fewer false positives would be detected. This would have the advantage of decreasing the number of unnecessary diagnostic work-ups of non-demented older adults, but the disadvantage of reducing the detection of dementia in its earliest stages. However, we are encouraged by preliminary work comparing the Mini-Cog with the MMSE in a population-based sample in Pennsylvania (the MoVIES sample: Ganguli et al., 1993), which suggests that the Mini-Cog remains at least as sensitive and accurate as the MMSE, even when the CDT scoring system differs somewhat from that used here.

Full diagnostic dementia assessment of false-positive subjects is costly in time, money, and emotional distress. Previous studies have found that combining cognitive screening with informant history improves dementia detection over either method alone (Gurland et al., 1995). Our approach, though different in detail, was similar in concept. For screening of elderly persons living alone without accessible informants, the Mini-Cog alone may be acceptable. However, a two-step screening approach, first using the Mini-Cog and then a brief informant interview, such as the short IQ-CODE (Jorm, 1994), warrants further population-based testing. If this method retains high sensitivity and minimizes false-positive screens, it may be the most cost-effective strategy for improving dementia detection and accurately selecting those in need of a complete diagnostic assessment.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the consultative contributions of E. Teng and the invaluable assistance of S. Neher, J. Cashman, M. Brush, M. McLaughlin, Sta Maria, J. Roques, E. Gil, D. Beekly, R. Barnhart, G. van Belle, and the staff of the Chinese Information and Service Center and Sea Mar Community Care Center, Seattle, WA, the Korean Women’s Association, Tacoma, WA, and the Washington State Aging and Adult Services Administration.

REFERENCES


