Screening for Dementia of the Alzheimer Type in the Community: The Utility of the Clock Drawing Test

Deborah A. Cahn, David P. Salmon, Andreas U. Monsch, Nelson Butters, W. C. Wiederholt, and Jody Corey-Bloom

Departments of Neurosciences and Psychiatry, University of California, San Diego

Elizabeth Barrett-Connor

Department of Family & Preventive Medicine, University of California, San Diego

This study reports the sensitivity and specificity of the Clock Drawing Test (CDT) for detecting dementia of the Alzheimer type in a community-dwelling sample of elderly subjects. Forty-two patients with clinically diagnosed Alzheimer's disease and 237 cognitively intact subjects were administered the CDT as part of an epidemiological study of aging and dementia. Three individual measures of clock drawing performance (quantitative score, qualitative score, and combined quantitative and qualitative score) were determined for each participant. When qualitative elements such as errors and strategies were incorporated into the CDT score, the sensitivity was 84% and the specificity was 72%. The findings suggest that a CDT score which evaluates qualitative and quantitative features provides reasonably good discrimination between normal elderly individuals and DAT patients. However, the CDT appears to have limited utility as a single screening instrument in the community. Instruments such as the Dementia Rating Scale (Mattis, 1976) provide better discrimination of DAT, indicating that functions such as memory and verbal fluency need to be assessed during screening.

The Clock Drawing Test (CDT) (Goodglass & Kaplan, 1972) has been proposed as a useful neuropsychological screening instrument for dementia of the Alzheimer type (DAT).
because of its time and cost efficiency and its sensitivity to the constructional, visuoperceptual, and conceptual deficits that are commonly observed in this disorder (Shulman, Shedletsky, & Silver, 1986; Sunderland et al., 1989; Mendez, Ala, & Underwood, 1992; Watson, Arfken, & Birge, 1993). Several studies have shown that performance on the CDT is highly effective for discriminating between patients with DAT and normal elderly (NE) individuals (Wolf-Klein, Silverstone, Levy, & Brod, 1989; Tuokko, Hadjistavropoulos, Miller, & Beattie, 1992; Feedman et al., 1994). For example, Wolf-Klein et al. (1989) reported rates of sensitivity (i.e., percentage of correctly identified patients) and specificity (i.e., percentage of correctly identified normal control subjects) for the diagnosis of DAT of 86% and 92%, respectively, using only the CDT. Similarly, Tuokko et al. (1992) found that the CDT discriminated between DAT and NE subjects with sensitivity of 92% and specificity of 86%.

Although previously reported diagnostic accuracy of the CDT is quite impressive, the effectiveness of this test as a general screening instrument for dementia in the community remains questionable. All of the previous studies employed clinically-based samples of self-selected or pre-screened subjects who may not have been representative of the community-dwelling population. The use of these clinically-based samples may have led to an overestimation of the true diagnostic accuracy of the CDT when it is applied to a community-based sample in which other medical disorders and subtle cognitive impairment are prevalent. The true effectiveness of the CDT for screening for DAT in the community can only be assessed by determining its diagnostic accuracy in a representative sample of community-dwelling individuals (Beardsall & Huppert, 1991).

The efficacy of the CDT as a screening instrument for DAT in the community may be enhanced by considering the qualitative aspects of a subject’s performance, in addition to the single, quantitative rating of the drawing that is usually obtained. Qualitative analyses of the types of errors produced on other neuropsychological tests have proven to be particularly helpful in distinguishing subjects in the earliest stages of DAT from NE individuals (Bondi et al., 1994).

Furthermore, qualitative assessment of the types of errors committed on the CDT have been shown to be useful in differentiating between demented and nondemented individuals (Kozora & Cullum, 1994) and between patients with different dementing disorders (Rouleau, Salmon, Butters, Kennedy, & McGuire, 1992). Rouleau and her colleagues demonstrated that DAT patients were more likely than NE subjects and patients with Huntington’s disease (HD) to make errors that reflected a deficit in adequately representing the attributes, features, and meaning of a clock. In contrast, HD patients were more likely than DAT and NE subjects to make errors indicative of visuospatial processing deficits. The results of this study suggest that DAT patients have a propensity to make a particular and unique pattern of errors on the CDT, and that consideration of the errors produced on the task when screening for DAT in the community may augment the test’s diagnostic effectiveness, particularly when attempting to classify individuals who manifest very subtle cognitive impairment.

In view of the lack of knowledge concerning the efficacy of the CDT as a screening instrument for DAT in the community, and the possibility that consideration of qualitative aspects of performance may enhance diagnostic accuracy, the present study had two goals. The first was to empirically derive for three measures from the CDT (i.e., a quantitative score, a qualitative score, and a combined quantitative-qualitative score) the effective cut-off scores for differentiating between DAT patients and normal elderly individuals in a community-based population sample.

Because all previous studies of the efficacy of the CDT for identifying patients with DAT employed clinically-based, self-selected, or pre-screened subjects, the present study was the
first to examine the true diagnostic accuracy of this test in the general population. This information is critical if the CDT is to be used as a screening instrument for DAT in general clinical practice or in large scale epidemiological studies of dementia.

The second goal of the present study was to examine the CDT performance of subjects who were identified by a diagnosing neurologist as being “at risk” for DAT because of subtle cognitive impairment (detected on neuropsychological evaluation or mental status testing) in the absence of clear functional decline. Although it is usually not possible to diagnose such subjects on the basis of psychometric performance alone (Storandt & Hill, 1989), the qualitative aspects of their performance on the CDT may provide important information about the underlying nature of their mild cognitive deficits. An error profile similar to that of patients with DAT may be indicative of the early stages of a dementing disorder, whereas a profile similar to that of normal elderly may indicate a more benign age-associated cognitive decline.

**METHOD**

**Study Population**

The Rancho Bernardo Study population has been described in detail elsewhere (Criqui, Barrett-Connor, & Austin, 1978; Wiederholt et al., 1993). Briefly, between 1972 and 1974, 5052 adults (representing 82% of the population of all adult residents aged 30 to 79) of Rancho Bernardo, CA, were surveyed for heart disease risk factors. The entire cohort has been followed with regard to health status to the present. Beginning in 1988, the 1800 locally-residing surviving members of the cohort, who were aged 65 and older, were invited to a clinic visit, and of these, 1349 completed a screening evaluation for cognitive functioning. One hundred ninety-nine subjects who were identified by the screening evaluation as having possible cognitive impairment according to operationally defined criteria (Wiederholt et al., 1993) attended a subsequent clinic visit to determine the actual clinical diagnoses. To control for the possibility of false negative errors, a sample (n = 203) of NE subjects was randomly selected for the comprehensive evaluation from those subjects who had no cognitive impairment according to the screening evaluation, with the same age and gender distribution as the sample who had possible cognitive impairment according to the screening evaluation. Thus, a total of 402 individuals attended a clinic visit where they were given the comprehensive evaluation, which included neurological examination, neuropsychological assessment, standard medical history and examination, and, in some cases, CT scans of the brain.

**Determination of Neurological Diagnosis**

The diagnosis of DAT was made according to the NINCDS–ADRDA criteria (McKhann et al., 1984). The examining neurologist made a diagnosis based on his/her examination, medical history, functional status, and the results of laboratory tests and CT scans. A second neurologist independently reviewed all data from the clinical examination and the laboratory results. Both examiners were blind to the specific scores of the neuropsychological tests but were informed of a neuropsychologist’s opinion as to whether the subject had two or more areas of cognitive deficit. Any discrepancy in the diagnosis between the examining neurologist and the reviewing neurologist was discussed in a monthly meeting attended by all participating investigators, and a consensus diagnosis was determined. The diagnosis of “at risk” (AR) for DAT was made when the subject demonstrated impairment in two or more areas of cognition on either the neurological or neuropsychological examination or both but had no obvious impairment in activities of daily living and had not demonstrated
functional decline by clinical assessment or by the score on the Pfeffer Outpatient Disability Scale (Pfeffer, Kurosaki, & Harrah, 1981).

Fifty-two individuals met NINCDS-ADRDA criteria for possible or probable Alzheimer’s Disease. Of the individuals diagnosed with DAT, 48 had been identified as possibly cognitively impaired by the screening evaluation and 4 had been identified as not cognitively impaired. Two-hundred-thirty-nine individuals were diagnosed as being cognitively-intact. Of these NE individuals, 168 had been identified as not cognitively impaired by the screening evaluation, and 71 had been identified as possibly cognitively impaired. Eighty-two individuals were identified by the comprehensive examination as being AR for Alzheimer’s disease.

Of these subjects, 21 had been identified by the screening evaluation as not cognitively impaired, and 61 had been identified as possibly cognitively impaired. Twenty-nine patients met criteria for dementia due to other neurological conditions. The current study focuses on the clock drawing performance of the NE, DAT, and AR individuals.

Ten DAT, 8 AR, and 2 NE subjects were excluded from the current analyses because of incomplete data. Table 1 shows the demographic characteristics and Mattis Dementia Rating Scale (DRS) (Mattis, 1976) scores for the NE, DAT, and AR groups. One-way analysis of variance (ANOVA) revealed that the three groups differed significantly in age, $F(2, 350) = 24.18, p < 0.0001$, and level of dementia as determined by the DRS, $F(2, 350) = 190.11, p < 0.0001$, but not in years of education, $F < 1$. Post-hoc Tukey tests ($\alpha = 0.05$) revealed that the DAT patients and AR subjects did not differ significantly in age, and that both of these groups were significantly older than the NE group ($p < 0.05$). The mean DRS score of the NE group was significantly higher than that of the AR group ($p < 0.05$) and that of the DAT group ($p < 0.05$). The mean DRS score of the AR group was significantly higher than that of the DAT group ($p < 0.05$).

Procedure

As part of the extensive neuropsychological test battery, the CDT was administered to each subject with the instructions to “Draw a clock, put in all of the numbers, and set the hands for 10 after 11.” Two trained independent raters scored the drawings according to the criteria developed by Rouleau et al. (1992). Briefly, a quantitative score is derived by adding the scores of three separate features: the clockface (0–2 points), the placement of the hands (0–4 points), and the placement of the numbers (0–4 points). Thus, a “perfect” clock would receive a quantitative score of 10. A qualitative analysis of error types is also performed in which the examiner notes the presence of conceptual, perseverative, stimulus bound, and spatial arrangement errors. A description of these different error types is presented next.

### Table 1

<table>
<thead>
<tr>
<th>Gender (F/M)</th>
<th>NE (n = 237)</th>
<th>At Risk (n = 74)</th>
<th>DAT (n = 42)</th>
</tr>
</thead>
<tbody>
<tr>
<td>140/97</td>
<td>45/29</td>
<td>25/17</td>
<td></td>
</tr>
<tr>
<td>Age in years</td>
<td>78.4 (6.8)</td>
<td>83.3 (5.5)</td>
<td>83.5 (4.4)</td>
</tr>
<tr>
<td>Years of education</td>
<td>13.9 (2.7)</td>
<td>13.9 (2.7)</td>
<td>13.6 (3.0)</td>
</tr>
<tr>
<td>DRS (max = 144)</td>
<td>136.9 (5.3)</td>
<td>127.9 (6.6)</td>
<td>115.8 (12.5)</td>
</tr>
</tbody>
</table>

Note. Standard deviations are presented in parentheses.
1. **Stimulus bound response.** The tendency of the drawing to be dominated or guided by a single stimulus. There may be three types of stimulus bound errors:
   (a) The hands may be set for 10 to 11 instead of 10 after 11.
   (b) The time is written beside the 11 or between 10 and 11 on the clock.
   (c) The hands are absent or are pointed toward “10” and/or “11”. This type of error is also considered to be a conceptual error.

2. **Conceptual deficit.** This error type reflects a loss or deficit in accessing knowledge of the attributes, features, and meaning of a clock. Included in this category are misrepresentation the clock itself (a clockface without numbers or inappropriate use of numbers) and misrepresentation of the time on the clock (the hands are either absent or inadequately represented or the time is written on the clock).

3. **Perseveration.** The continuation or the recurrence of activity without an appropriate stimulus. In clock drawing, this error is seen in the presence of more than two hands and abnormal prolongation of numbers (writing beyond “12”).

4. **Neglect of left hemispace.** All attributes of the clock are written on the right half of the clock face. Possible neglect of right hemispace was also evaluated, but this type of error was never observed.

5. **Planning deficit.** This error type is represented by gaps before the 12, 3, 6, or 9, depending on the strategy used in drawing.

6. **Nonspecific spatial error.** A deficit in the spatial layout of numbers, without any specific pattern in spatial disorganization.

7. **Numbers written on the outside of the clock.** Numbers written either around the perimeter of the circle or on the circle itself.

8. **Numbers written counterclockwise.** Arrangement of the numbers with “12” at the top of the clockface and then continuing around in a counterclockwise fashion.

In the present study, the following three measures were derived:

1. **Quantitative CDT score.** Assesses the presence and correctness of the features of the clock (Maximum = 10 points).

2. **Qualitative CDT score.** The number of error types, as described by Rouleau et al. (1992) that were committed (Maximum = 8 error types). Because the error types were scored only for their presence or absence, the total number of error types committed provided a summary score of the qualitative errors.

3. **Global CDT score.** The CDT quantitative score minus the qualitative error score. This score takes into account not only presence and correctness of the features of the clock, but also the number of error types made and strategies used in the construction of the clock.

The two raters demonstrated high interrater reliability ($r = 0.95$, $p < 0.0001$) for the Quantitative CDT score. This level of interrater reliability is consistent with the results of previous studies (Nussbaum, Fields, & Starratt, 1992). A kappa statistic was calculated for each of the eight error types by comparing the observed agreement with the agreement expected by chance (50%) alone.

Of the 8 kappa values, 7 were in the range indicating excellent agreement beyond chance, and 1 (nonspecific spatial arrangement) was in the range of fair to good agreement beyond chance (Landis & Koch, 1977).

**Statistical Methodology**

A Receiver Operating Characteristic curve (ROC) analysis was performed to quantify test accuracy and to characterize the distribution of test scores in the study samples. For
the present analyses, each CDT score was treated as a possible cut-off score and the corresponding sensitivity (i.e., percentage of DAT patients scoring below this score) and specificity (i.e., percentage of NE subjects scoring above this score) for the diagnosis of DAT were calculated. An ROC curve was then generated for each of the three CDT measures by plotting sensitivity against specificity for every possible cut-off score. Because it was decided a priori that sensitivity and specificity were to be of equal importance, the optimal cut-off score was the point at which the sum of both measures reached a maximum value. In the present study, the criterion standard for correct classification as a DAT patient or NE subject was the clinical diagnosis made by the two independent senior staff neurologists.

RESULTS

Because the DAT, NE, and AR subject groups differed in age, the possible effect of age on CDT performance was of some concern. Therefore, simple regression analyses were performed to examine the relationship between age and CDT performance in cognitively intact NE subjects.

Because these analyses revealed no significant relationship between age and quantitative CDT score ($r = -0.08, p < 0.05$), qualitative error score ($r = 0.07, p < 0.05$) or global CDT score ($r = -0.08, p < 0.05$), all subsequent analyses were performed using the raw scores.

The means and SDs of the three CDT scores achieved by each group are presented in Table 2. Separate ANOVA comparing the DAT, NE, and AR groups on each measure revealed significant main effects of Group for the Quantitative CDT score, $F(2, 350) = 59.2, p < 0.0001$; the Qualitative CDT score, $F(2, 350) = 47.6, p < 0.0001$; and the Global CDT score, $F(2, 350) = 64.6, p < 0.0001$. Post-hoc Tukey tests ($\alpha = 0.05$) indicated that DAT patients scored significantly worse than NE and AR subjects, and AR subjects scored significantly worse than NE subjects on all three CDT measures (all $p$'s $< 0.05$).

The results of the ROC analyses indicated that the optimal cut-off score for the Quantitative CDT measure ($\leq 7$) produced 88% sensitivity and 63% specificity for the diagnosis of DAT. The optimal cut-off score for the Qualitative measure ($\geq 1$ error) was similarly effective, producing 81% sensitivity and 68% specificity. The Global CDT score, which takes into account both quantitative and qualitative aspects of performance, was the most effective of the three CDT measures for the diagnosis of DAT with an optimal cut-off score of $\leq 6$ producing 83% sensitivity and 72% specificity. Thus, the combination of qualitative and quantitative measures on the CDT resulted in a statistically significant increase in specificity relative to a simple quantitative rating from 63% to 72% ($\chi^2 = 4.26; p < 0.04$), with no appreciable decline in sensitivity (88–83%).

| TABLE 2 | Mean Score of the Normal Elderly (NE), At Risk, and Dementia of the Alzheimer Type (DAT) Subjects on the Three Clock Drawing Test Measures |
|---------|-----------------|-----------------|-----------------|
|         | NE               | At Risk          | DAT             |
| Quantitative CDT Score (max = 10) | 8.7 (1.2) | 7.4 (1.7) | 6.4 (2.1) |
| Qualitative CDT Score (max = 8) | 0.43 (0.72) | 1.1 (1.1) | 1.8 (1.5) |
| Global CDT Score (max = 10) | 8.3 (1.8) | 6.3 (2.5) | 4.5 (3.2) |

Note. Standard deviations are presented in parenthesis.
The clock drawing performance of the AR subjects was examined to determine their diagnostic classification based upon the three CDT cut-off scores derived for the DAT and NE subjects. The Quantitative CDT cut-off score of ≤7 classified 50% of the AR subjects as normal and 50% as DAT. The Global CDT cut-off score of ≤6 classified 51% of AR subjects as normal and 49% as DAT. When qualitative features alone were evaluated, a cut-off score of ≥1 error classified 35% of AR subjects as normal and 65% as DAT. Thus, AR subjects were as likely to be classified as DAT as they were normal on the basis of their Quantitative and Global CDT scores. However, the Qualitative CDT score classified more AR subjects as DAT than as normal. In addition, the pattern of error types produced by the AR subjects was very similar to that of the DAT patients.

The percentage of NE, DAT, and AR subjects that made each type of error on the CDT are presented in Figure 1. Chi-square analyses comparing the three groups revealed that there was a significant effect of diagnosis (p ≤ 0.001) for the following error types: Conceptual, Perseverative, Stimulus-bound, Planning, and Nonspecific Spatial. The three groups did not differ significantly in the number of errors reflecting neglect of left hemispace nor in writing the numbers outside of the clockface. Too few subjects made counter clockwise errors to perform an analysis. Post-hoc Chi-square analyses comparing the DAT and NE subjects demonstrated that more DAT patients than NE subjects committed Planning (p < 0.01), Nonspecific spatial (p < 0.0001), Stimulus bound (p < 0.0001), Conceptual (p < 0.0001), and Perseverative (p < 0.0001) errors. The AR and DAT groups differed only on nonspecific spatial (p < 0.05) errors, with a greater percentage of DAT than AR subjects making this type of error. Comparison of the AR and NE subjects showed that more AR subjects than NE subjects committed Conceptual (p < 0.01), Stimulus bound (p < 0.0001), Perseverative (p < 0.01), and Planning (p < 0.0001) errors. Examples of the types of errors made on the CDT by four different AR subjects are presented in Figure 2. An example of a Conceptual deficit, in which the subject was unable to convey the time properly on the clock is shown in the upper left quadrant. A Stimulus bound error, in which the subject drew the hands pointing at “10” and “11” is shown in the upper right quadrant. Perseverative behavior is shown in the lower left quadrant; the lower right quadrant illustrates planning difficulty.

DISCUSSION

The results of the present study indicate that a clock drawing score that takes into account both qualitative and quantitative features can provide reasonably good sensitivity and specificity for discriminating DAT from normal elderly individuals in a community-based population. The Global CDT score, based on the scoring procedure described by Rouleau et al. (1992), provided sensitivity of 83% and specificity of 72% for differentiating DAT patients from normal elderly subjects. These findings suggest that when used as a single screening instrument, the CDT will be only marginally effective in detecting DAT in the community.

The sensitivity and specificity for the diagnosis of DAT obtained with the optimal cut-off scores in the present community-based sample are lower than previously reported with clinically-based or volunteer samples (Wolf-Klein et al., 1989; Tuokko et al., 1992; Freedman et al., 1994). As with other neuropsychological tests (Cahn et al., 1995), the lower discrimination between DAT and NE observed here may have occurred because NE subjects in the community demonstrate greater variability in CDT performance than healthy, self-selected volunteer subjects. It is likely that this increased variability in CDT performance is due to the presence of NE individuals in the community who have medical disorders that may affect test performance or who have subtle cognitive decline that falls between dementia and normal cognitive functioning.
The CDT measure that provided the best discrimination between DAT and NE individuals incorporated both a quantitative rating and an analysis of the qualitative features of the CDT. This finding suggests that evaluation of the types of errors produced on the CDT may aid in the early detection of dementia. Such an error analysis has been shown previously to be extremely useful for differentiating DAT from normal aging and from other dementing disorders (Fuld, 1983; Jacobs, Salmon, Tröster, & Butters, 1990; Rouleau et al., 1992; Freedman et al., 1994). The predominant types of errors committed by the DAT patients in the present study were Conceptual, Stimulus bound, and nonspecific spatial arrangement of numbers. Rouleau and colleagues (1992) have suggested that Conceptual errors, in which the subject demonstrates a misrepresentation of the clock or difficulty in setting the hands may occur because of loss of semantic knowledge concerning the meaning and attributes of the clock (Butters, Granholm, Salmon, Grant, & Wolf, 1987; Rouleau et al., 1992; Chan et
Screening for Dementia With the Clock Drawing Test

FIGURE 2. Upper left quadrant: Conceptual deficit in the clock drawing of an at risk (AR) subject, Upper right quadrant: Stimulus-bound error in the clock drawing of an AR subject, Lower left quadrant: Perseverative behavior in the clock drawing of an AR subject, Lower right quadrant: Planning difficulty in the clock drawing of an AR subject Screening for Dementia.

al., 1993). Stimulus bound errors in which the patient is pulled to the perceptual features of the stimulus (i.e., setting the hands at 10 and 11), may represent a tendency to process information on a more perceptual than semantic level (Kaplan, 1988). This type of error has been attributed to frontal system dysfunction which disrupts abstract thinking.

In the current study, the optimal cut-off score of the quantitative CDT measure classified 50% of AR subjects as DAT and 50% as normal. Furthermore, when quantitative ratings and qualitative features were considered together, the classification of AR subjects remained relatively unchanged. The inconsistency with which the CDT quantitative rating classified AR subjects is similar to the observation of Storandt and Hill (1989) that the performances of questionably demented subjects overlaps with those of normal control and mildly-demented subjects on objectively, or quantitatively, scored neuropsychological tests.

When the types of errors produced by the AR subjects on the CDT were used for classification, the majority of these subjects were classified as DAT. Close examination of the error types revealed that the AR subjects produced a pattern of errors similar to that of the DAT patients. In particular, the DAT and AR subjects were equally prone to make Planning, Stimulus-bound, Conceptual, and Perseverative errors. Furthermore, both the DAT and AR groups committed more of these types of errors than the NE subjects. It is possible that changes in such subtle qualitative features of clock drawing may be an early indication of a
progressive dementing disorder. It should be emphasized, however, that only longitudinal follow-up of the AR subjects will determine whether these error types are predictive of future cognitive deterioration.

Although the CDT by itself was effective in differentiating between DAT and NE individuals in the present study, more accurate classification has been reported with other single or composite neuropsychological measures. For example, measures of memory and, in particular, rapid forgetting, may be more appropriate screening instruments as memory impairment is the most prominent and early manifestation of the disorder (Eslinger, Damasio, Benton, & Allen, 1985; Moss & Albert, 1988; Welsh, Butters, Hughes, Mohs, & Heyman, 1991; Welsh, Butters, Hughes, Mohs, & Heyman, 1992). Also, the Mattis Dementia Rating Scale (Mattis, 1976) has been recently demonstrated to have excellent classification accuracy, particularly because it assesses memory and language functions (Monsch et al., 1995).

Finally, it is important to note the limitations to the generalizability of the current findings. The Rancho Bernardo community is a predominantly white, highly-educated, middle to upper-middle class population. The cut-off scores and rates of sensitivity and specificity that were derived with the three CDT measures might not be applicable to less well-educated or non-White communities. While some investigators suggested that performance on clock drawing tasks is independent of socioeconomic factors (Shulman, Shedletsky, & Silver, 1986; Watson, Arfken, & Birge, 1993) recent evidence suggests that educational background may affect clock drawing ability (Ainslie & Murden, 1993; Cahn, Wiederholt, Salmon, & Butters, 1994). Additional studies should be conducted to evaluate the diagnostic accuracy of the CDT in community-dwelling populations with different socio-demographic characteristics.

REFERENCES


Screening for Dementia With the Clock Drawing Test


