Cognitive Test Performance in Detecting, Staging, and Tracking Alzheimer's Disease

Joseph J. Locascio, PhD; John H. Growdon, MD; Suzanne Corkin, PhD

Objectives: To identify the specific cognitive deficits that characterize Alzheimer's disease (AD) and determine which cognitive tests, or combination of tests, are best for detecting AD (ie, distinguishing patients with AD from normal control subjects), staging AD (ie, distinguishing different severities of dementia), and tracking disease progression.

Subjects: Patients with AD (n=123) and normal control subjects (n=60) of comparable age, education, and gender distribution.

Setting: Outpatient care.

Measures: Ten cognitive tests of memory, language, visuospatial abilities, and reasoning; the Information, Memory and Concentration subtests of the Blessed Dementia Scale, and the total score on an activities of daily living questionnaire.

Design: Patients with AD were tested every 6 to 24 months over a span of up to 3.5 years.

Results: Patients with AD were significantly inferior to normal control subjects on all cognitive tests. The scores of patients with AD worsened over time. Delayed recall of stories and figures showed sharp deterioration to an early floor, consistent with the finding that these tests discriminated patients with mild AD from normal control subjects well but were poor for staging. Confrontation naming, semantic fluency, and immediate recognition of geometric figures showed steady linear decline across time for patients with AD, consistent with these tests being found best for staging dementia severity.

Conclusions: We postulate that the pathologic bases of impairment in delayed recall are atrophy of cholinergic ventral forebrain neurons and partial deafferentation of the hippocampus, both of which occur early in the course of AD. Worsening language and visuospatial abilities likely reflect progressive loss of neocortical neurons and their connections.

(Arch Neurol. 1995;52:1087-1099)

A primary goal of behavioral research in Alzheimer's disease (AD), and our objective here, is to identify tests that (1) clearly discriminate patients with AD from age- and education-matched normal control subjects (NCS), (2) are sensitive to differences in dementia severity, and (3) track the progression of the disease. Global rating scales (“mental status examinations”) have been designed to detect the presence and severity of dementia, and numerous clinical studies have sought to describe the specific nature of cognitive deficits associated with AD. Of these reports, one consistent finding has been that beginning early in the disease, patients with AD are impaired relative to NCS on tests of delayed recall of verbal material. Less information has been published on the nature and progression of language and visuospatial impairments in AD, although there is the suggestion that they show a more gradual steady decline throughout the course of the disease than memory abilities, which tend to deteriorate rapidly to an early floor.

With respect to assessing change over a span of time, most prior studies on longitudinal change in cognitive test performance in patients with AD examined a small number of subjects, followed subjects for a relatively short period of time, or studied only a few cognitive functions. We hoped to overcome limitations of these studies by (1) employing a large sample of patients with AD and capturing the full range of cognitive change in AD across a broad spectrum of demen-
SUBJECTS, METHODS, AND MATERIALS

METHODS

The data for this study were extracted from the registry that contained information on all patients with AD and NCS examined in the Memory Disorders Unit at Massachusetts General Hospital, Boston, from 1985 through 1993. The criteria for selection were: (1) all currently available data from NCS whose first language was English and who had no history of dementia or other neurologic or psychiatric disease; and (2) all currently available data from patients diagnosed as having probable AD, and who, on at least two occasions, took the set of cognitive tests. The NCS were generally spouses or caregivers of patients, and some were volunteers from the Geriatric Education Center of the Harvard Medical School (Boston, Mass.) Division on Aging. The NCS signed consent forms. The extracted data set contained a total of 323 records (test sessions) from 123 patients with AD (data at first testing: mean age, 70.8 ± 8.3 years; education, 13.3 ± 3.5 years; and gender, 76 female and 47 male subjects), and 90 records from 60 NCS (data at first testing: age, 69.8 ± 11.2 years; education, 13.3 ± 3.1 years; and gender, 35 female and 25 male subjects). Of the patients with AD, 13 had died by the time of this analysis; an autopsy was performed on eight of these cases; and all eight had a neuropathologic diagnosis of AD. For two other patients in the registry with a clinical diagnosis of probable AD who met the requirements of the study, the diagnosis was not confirmed at autopsy; they were not included in the census of 123 patients with AD. It is reasonable to assume that other patients with probable AD will not be diagnosed as having AD at autopsy. Of the 123 patients with AD, 103 were native English speakers, and four spoke English as a second language from childhood (language information was not obtained for 16 patients). Each record in the extracted data set corresponded to a different cognitive testing session during which all or most of the cognitive tests were administered. Of the 60 NCS, 11 had two or more testing sessions (records) in this data set, with an average of 2.5 years between the first two testing sessions. All 123 patients with AD had at least two testing sessions (with a mean of 1 year between the first two sessions); 50 patients had three or more testing sessions, 19 patients had four or more testing sessions, five patients had five or more testing sessions, and three patients had six testing sessions. Patients with AD underwent follow-up for a mean of 1.8 ± 1.1 years; the range of follow-up extended from 0.5 to 6.3 years. Duration of AD (data at first testing, mean, 3.3 ± 2.0 years; range, 0.2 to 11 years) was based on the recollections of caregivers and was not considered a reliable variable for analysis.

The BDS is a brief mental status examination that measures the presence and severity of dementia. The IMC section of the BDS tests a subject's memory and knowledge of personal information and public events. A score of 0 to 2 is considered normal, for a mean age of 65 years, and an increasing number of errors indicates increasing impairment up to a maximum score of 37. The ADL questionnaire includes items concerning the patient's ability to function in areas of self-care, household activities, and employment. It is completed by the caregiver, usually the spouse. Impairment is expressed on a continuum of 0 (no impairment) to 100 (maximal dependency). These data were selected such that the BDS and ADL assessments were the closest within 90 days of the cognitive test session; data from a given assessment were associated with only one test record. The BDS, ADL, and cognitive test were completed on the same day for about 50% of the patient records. The BDS and ADL scores were unavailable for 17 of 323 patient records. Mean IMC and ADL scores at the date of initial testing for the 123 patients with AD were 11.6 ± 5.9 and 28.7 ± 16.0, respectively. The IMC data were also obtained from 59 NCS (mean score, 0.85 ± 1.27).

To analyze test performance across stages of dementia severity, we divided all records for those with AD into three groups of approximately equal size based on IMC score. One group (containing 99 records) had IMC scores of 10 or less and were considered to represent mild AD; another group with scores of 11 to 16 (99 records) were considered to represent moderate AD, and the third group, with scores of 17 or greater (92 records), were considered to represent severe AD. Age levels were comparable across these subgroups and the NCS (Table 1), but the education level of the patients with AD was lower than that of the NCS (P < .001). Therefore, we adjusted for education in the analyses described below.

The series of cognitive tests assessed explicit memory, language, visuospatial abilities, and abstract reasoning.

centration (IMC) section of the Blessed Dementia Scale (BDS), and the total score on an activities of daily living (ADL) questionnaire.

RESULTS

RELIABILITY AND INTERCORRELATIONS

To gain estimates of test reliability, we correlated test scores across sessions for patients with AD and NCS separately. Correlations across temporally adjacent sessions were often on the order of .3 to .6, and were generally statistically significant for patients with AD for the first few time periods, which is where sample sizes were largest. Measures with evidence of good reliability were the
(Table 2). Detailed discussions of the nature and administration of these cognitive tests, as well as the BDS and ADL, are available elsewhere. All cognitive tests were scored on a numeric scale, with high scores indicating proficiency and low scores indicating impairment.

DATA ANALYSIS

Detecting and Staging AD

Most of these analyses followed two general approaches. In the first approach, analyses were based on all available data even though such data contained more than one record for most subjects. This type of analysis made full use of the richness of the data (maximal information); however, P values from significance tests were not strictly accurate because of the nonindependence of some of the observations. In the second approach, analyses were restricted to data chosen so that only one record (usually the earliest or most complete) per subject was analyzed. This type of analysis provided the valid P values necessary to confirm the statistical significance of the analyses of the complete data.

Within each of these two general approaches, conventional linear discriminant analysis, regression analyses, and graphic techniques identified cognitive tests that best discriminated patients with AD from NCS, and differentiated groups of patients with AD that varied in dementia severity. In addition, using correlation, regression, and graphic methods, cognitive test scores were related to continuous measures of AD severity (IMC and ADL) to obtain a finer delineation of the relationship between cognitive impairments and severity of dementia than could be ascertained from analyses of broad categories of severity.

Assessment of Change

We measured change in test performance by comparing each subject’s score for each task in his or her first session to scores for that task in each subsequent session. Specifically, for each record (test session), we computed change from the initial level of a measure for the given subject to the level at the time of testing for that record and divided this change by the amount of time in years elapsing between the initial test and the given test session.

Regression analyses were also employed in which the score for a given cognitive test (or IMC or ADL) was the dependent (predicted) variable regressed on length of time in the study (ie, from initial testing) for the subject as of the date that score was obtained, using all available records as individual observations. Trends across time were adjusted for the subject’s initial level of the measure in question, i.e., multiple regression analysis was employed where the initial score was a simultaneous predictor along with time in the study, so that the coefficient for time indicated change for subjects with the same or similar initial scores. Curvilinear effects (quadratic and log) were also tested as were interactions of time and initial score so that nonparallel change dependent on initial score could be examined. Further, it was of interest to see whether and how patterns of change differed as a function of the total length of time a subject was in the study. Thus, in the case of patients with AD, analyses were conducted separately for groups of individuals with a different total span of time in the study. The time spans that were separately analyzed were as follows: up to 2 years, up to 3 years, and up to 4 years. Only a relatively few observations corresponded to spans beyond 4 years, making estimates unreliable for those periods. For analyses that documented change across longer spans of time in the study, ie, up to 3 and 4 years, subjects with a relatively short total span of time in the study were excluded (<1 and <1.5 years, respectively) to minimize attitudinal bias. Because of the smaller sample size and fewer retestings for NCS, only one regression analysis was conducted for each cognitive test for those in this group, using all available data pooled across all years.

In these analyses, the data for each subject came from multiple test sessions. This lack of independence, especially autocorrelated error, across observations violated assumptions of conventional significance tests, so that P values could only be considered approximate. However, estimates of changes in scores over time should not be biased because of these problems. In place of significance tests, indexes of “goodness of fit” provided information on how well estimated change models were congruent with actual data. These indexes indicated the percentage of the variance in the dependent variable (test scores) that was accounted for by the predictors (time in study, initial score) in a given change model. Higher variance accounted for by time in the study indicated that actual scores varied closely with time (similarly for other predictors). Variance accounted for by time in the study and initial test score combined, as well as the unique contribution of time in the study, were computed.

IMC (.73 to .84), ADL (.70 to .85), the Stroop Color Naming Test (.68 for patients with AD and .72 for NCS), the Mental Rotation test (.71 for AD), Boston Naming Test (.84 to .88 for patients with AD and .77 for NCS), Raven’s Colored Progressive Matrices (.68 to .86 for patients with AD), and the Money Road Map test (.66 to .84 for patients with AD). Another way to establish reliability was to determine the correlation between different tests that are so similar that they could be considered alternate forms of a single test. For example, using all available records, for story recall tests, Immediate Stories Verbatim and Gist (Table 2) scores correlated about .8, and Delayed Stories Verbatim and Gist scores correlated slightly higher. For the Easy and Hard Recall subtests of the Geometric Figure test, the correlation was about .4; for the Easy and Hard Figure Copying subtests, the correlation was .76 for patients with AD. Analyses using only one record per subject confirmed the statistical significance of these correlations.

Other correlational analyses showed that for patients with AD, age, education, and duration of illness did not correlate with IMC or ADL scores; duration of illness also did not correlate with cognitive test scores. However, there were significant negative correlations between age and many of the cognitive tests for the NCS and, to a lesser degree, for those with AD. For both groups, education was positively and significantly correlated with some cognitive test scores. These latter findings made it important to adjust for age and education in the analyses of cognitive test scores.
Table 1. Clinical Characteristics of Patients With AD and NCS at the Time of Initial Testing*

<table>
<thead>
<tr>
<th>Group</th>
<th>Male-to-Female Ratio</th>
<th>Mean Age ±SD, y</th>
<th>Mean Education ±SD, y</th>
<th>Mean ADL±SD</th>
<th>Mean IMC:±SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCS (n=60)</td>
<td>25:35</td>
<td>68.8±11.2</td>
<td>15.1±3.1</td>
<td>11±4.5</td>
<td>0.65±1.27</td>
</tr>
<tr>
<td>Patients with AD (n=123)</td>
<td>47:76</td>
<td>70.7±8.5</td>
<td>13.1±3.5</td>
<td>28±13.6</td>
<td>11±4.5</td>
</tr>
<tr>
<td>Mild(1)</td>
<td>29:27</td>
<td>70.7±8.6</td>
<td>14.0±3.4</td>
<td>22±12.5</td>
<td>7±1.2</td>
</tr>
<tr>
<td>Moderate(2)</td>
<td>9:30</td>
<td>70.4±8.9</td>
<td>13.3±3.1</td>
<td>29±13.8</td>
<td>13±1.6</td>
</tr>
<tr>
<td>Severe(3)</td>
<td>8:12</td>
<td>72.3±9.6</td>
<td>11.7±4.1</td>
<td>56±15.9</td>
<td>21±5.2</td>
</tr>
</tbody>
</table>

*AD indicates Alzheimer's disease; NCS, normal control subjects; ADL, an activities of daily living scale; IMC, Information, Memory, and Concentration subscale of the Blessed Dementia Scale; and N/A, not applicable.

| Subgroups formed on basis of IMC (see "Subjects, Methods, and Materials" section). Subgroup sample sizes do not sum to 123 because of missing IMCs at initial testing for 10 patients. |

For patients with AD, the IMC and ADL scores correlated negatively (and significantly in the earliest record per subject analyses) with virtually all cognitive test. The IMC and ADL intercorrelated positively among patients with AD for all records (r=.63; with age and education partialled, r=.67) and in analysis of the earliest record per patient (r=.57, P<.0001; with age and education partialled, r=.60, P<.0001).

DETECTING AND STAGING AD

The performance of patients with AD as a group was significantly inferior to that of the NCS on all cognitive tests (P<.003, except for P=.063 for easy figure copying, based on earliest record per subject; P values adjusted for multiple tests using a bootstrap method(2)). Analyses of variance indicated a main effect of gender for the Money Road Map test only (P=.0001; females lower), and no interactions between gender and group (AD, NCS).

Based on all records, Delayed Story Recall, i.e., New York University Stories Test-Delayed Recall variation (Table 2), was the single best test that distinguished patients with AD from NCS; 69% (mild AD) to 85% (severe AD) of its variance was accounted for by subject group (Table 3). For two tests used in combination, Delayed Recall of a Story and Delayed Recall of a Geometric Figure were the best pair at distinguishing patients with AD from NCS, with a 96% to 98% sensitivity and a 91% to 99% specificity (depending on dementia severity; Figure 1). Using more than two tests in combination generally did not increase the discrimimability of groups to any appreciable degree.

In contrast to distinctions between NCS and those with AD, most of the cognitive tests were poor at predicting stage-of-dementia severity (Table 3). No test or pair of tests was particularly good at discriminating mild from moderate AD; Delayed Story Recall was the best single predictor, but only 10% of its variance was accounted for by these two groups. The Boston Naming and Verbal Fluency tests together were best for discriminating mild from severe AD, but still only 43% of the variance of their linear combination was accounted for by the groups; sensitivity and specificity were not high, at 77% and 84%, respectively. The power to distinguish moderate from severe dementia was even less; Verbal Fluency was the single best predictor, but with just 22% of variance accounted for. The Boston Naming and Benton Visual Retention tests were the best pair of predictors for this comparison, but they increased the variance accounted for to only 31%.

With statistical adjustments for education and age simultaneously, the results for detecting and staging AD were essentially the same as those found without performing the statistical adjustment.

Figure 2 illustrates the relatively good discrimination between NCS and patients with mild (and more severe) AD found with Delayed Story Recall (Verbatim and Gist) as well as with Geometric Figure (Easy) Recall. Floor effects were also evident for these tests. The relatively good staging of AD provided by the Boston Naming, Benton Visual Retention, and, to some extent, the Verbal Fluency tests is reflected in the fairly steady linear decline across increasing severity of AD for those scores.

The tests found most useful for discriminating patients with AD from NCS, the Delayed Story Recall and Geometric Figure Recall tests, correlated moderately among patients with AD (for all records, r=.36 to .4), as did the tests best for staging, the Boston Naming, Verbal Fluency, and Benton Visual Retention tests (r=.3 to .5), whereas these two sets of measures did not cross-correlate well (r=.11 to .26).

To confirm the statistical significance of the above results, we ran stepwise discriminant analyses using one record per subject (the earliest having the least missing values). All groups formed for these analyses had similar age and education distributions. The sample size for the normal group was always 29 subjects; for all analyses, the sample size for those with mild AD ranged from 25 to 34 subjects; for those with moderate AD, 23 to 27 subjects; and those with severe AD, 10 subjects. The results of these analyses confirmed the highly significant (P<.002) power of Delayed Story Recall and Geometric Figure Recall in discriminating those with AD from NCS. The analyses also attested to the relatively poor ability of the cognitive tests to separate stages of AD but showed that the Boston Naming and Verbal Fluency tests significantly (P<.05) discriminated patients with mild AD from patients with severe AD.

RELATIONSHIP OF COGNITIVE TEST SCORES TO IMC SCORES

These analyses examined the relationship between cognitive test performance and AD severity where the latter
was assessed by a continuous variable, the IMC section of the BDS. Some of the analyses attempted to estimate effects across the full range of normalcy and severity of AD by including those NCS for whom IMC data were obtained.

For analyses employing all available records for the patients with AD and NCS, bivariate regressions and scatterplots with each cognitive test as the dependent (ordinate) variable and the IMC as the predictor (abscissa) showed Delayed Story Recall (Verbatim and Gist) and Geometric Figure Recall with substantial curvature and a bottoming-out effect at moderate AD severity levels, and relatively high percentage variance due to curvilinear effects. The Boston Naming, Verbal Fluency, and Benton Visual Retention tests had primarily straight-line relationships across all levels of AD severity with no floor- ing out (relatively small percentage variance owing to curvilinear effects) (Figure 3). Complementary analyses performed with only one record (the earliest and the latest, separately) per subject verified the statistical significance of these results.

**BOSTON NAMING TEST: EFFECTS OF WORD FREQUENCY, PICTURE FAMILIARITY, AND VERBAL FLUENCY**

Given the prominence of the staging results for the Boston Naming Test, we speculated that these staging effects may be partly a reflection of the fact that the test was constructed to require word responses of widely varying frequency of occurrence in the English language and that it employed pictures of varying familiarity to the subjects. We hypothesized that low-frequency words and unfamiliar pictures might discriminate NCS from subjects with mild AD, whereas higher-frequency words and familiar pictures might serve better in staging (M. Caterina Silveri, MD, oral communication, 1994). We reanalyzed the data from the Boston Naming Test, separating items into those with “high” vs “low” word frequency as well as “familiar” vs “unfamiliar” pictures. We found no differential discrimination vs staging capabilities for items on the Boston Naming Test corresponding to high-frequency vs low-frequency words nor for familiar vs unfamiliar pictures.

Another explanation for the linear decline of performance on the Boston Naming Test with increasing severity of AD concerns its relationship to verbal fluency. Because a component of performance on confrontational naming is related to verbal fluency abilities, it may be that performance on the Boston Naming Test shows a linear decline only because of a linear decline in fluency capabilities, the latter being demonstrated with our measure of the same. In fact, scores on the Boston Naming Test correlated significantly with Verbal Fluency in our study (for all records, r=.5 for patients with AD and r=.32 for the NCS; in one record per subject analyses, P<.001 for both correlations). We performed an analysis of covariance where the Boston Naming Test was the dependent variable, the groups were NCS and subjects with mild, moderate, or severe AD, and our measure of verbal fluency was a covariate. The linear component of change for the Boston Naming Test across groups of increasing severity remained stronger than curvilinear components, whether fluency was
Table 3. Cognitive Tests That Discriminate Patients With AD, NCS, and Stages of AD Severity Using All Available Records

<table>
<thead>
<tr>
<th>Group Comparison</th>
<th>Best Single Predictor (% Variance)</th>
<th>Best Pair of Predictors (% Variance; Sensitivity; Specificity)</th>
<th>Other Good Predictors$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal control vs mild AD</td>
<td>Story Recall (Delay Verbatim) (68.6)</td>
<td>Story Recall (Delay Verbatim), Geometric Figure (Easy) Recall (72.5; 86; 91)</td>
<td>Verbal Fluency, Story Recall (Immediate Verbatim and Delay Gist)</td>
</tr>
<tr>
<td>Normal control vs moderate AD</td>
<td>Story Recall (Delay Gist) (83.3)</td>
<td>Story Recall (Delay Gist), Geometric Figure (Easy) Recall (83; 98; 95)</td>
<td>Verbal Fluency, Stroop Color Naming, Benton Visual Retention, Story Recall (Delay Verbatim)</td>
</tr>
<tr>
<td>Normal control vs severe AD</td>
<td>Story Recall (Delay Gist) (85)</td>
<td>Story Recall (Delay Gist), Geometric Figure (Easy) Recall (88; 98; 98)</td>
<td>Benton Visual Retention, Verbal Fluency, Boston Naming, Picture Arrangement (WAIS-R); Story Recall (Immediate Gist)</td>
</tr>
</tbody>
</table>

Discrimination

Staging

<table>
<thead>
<tr>
<th>Mild AD vs moderate AD</th>
<th>Story Recall (Delay Gist) (10.2)</th>
<th>Story Recall (Delay Gist), Picture Arrangement (WAIS-R); (16.6; 61; 85)</th>
<th>Geometric Figure (Hard) Recall, Story Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild AD vs severe AD</td>
<td>Boston Naming (33)</td>
<td>Boston Naming; Verbal Fluency (43; 77; 84)</td>
<td>Story Recall (Immediate Gist and Recognition), Benton Visual Retention, Geometric Figure (Easy) Recall</td>
</tr>
<tr>
<td>Moderate AD vs severe AD</td>
<td>Verbal Fluency (21.7)</td>
<td>Benton Visual Retention; Boston Naming (31; 58; 79)</td>
<td>Geometric Figure (Hard) Copying; Story Recall (Immediate Verbatim and Gist)</td>
</tr>
</tbody>
</table>

$AD$ indicates Alzheimer’s disease; NCS, normal control subjects; and WAIS-R, Wechsler Adult Intelligence Scale–Revised.

$\dagger$ The best single predictor and percentage variance of variable accounted for by group membership.

$\ddagger$ Best pair of predictors used together and percentage variance of their optimal linear combination accounted for by group membership. Also, the sensitivity and specificity based on the best pair of predictors (percent).

$\S$ Other cognitive tests with relatively good prediction power, singly or in combination.

covaried or not ($P<.0001$ for linearity in both cases), and covariate adjusted means continued to show a linear trend.

**CHANGE OVER TIME**

Test-retest scores were available for 11 NCS. Their cognitive test scores did not change, aside from occasional small increases for relatively low initial scorers and decreases from ceilings, i.e., regression to the mean.

For patients with AD, scores on all cognitive tests were significantly ($P<.06$) lower at baseline than were baseline scores for NCS and worsened further over time (Table 4). For all cognitive tests, as well as the IMC and ADL, $P$ values were below .05 for time as a predictor (initial score held constant) and usually were at the .001 or .0001 level. Change over time in the cognitive test scores for the patients with AD generally fit one of the three patterns discussed below (Figure 4).

Change Pattern A: Curved-Floor Effect

Scores for cognitive tests following this pattern declined in a decelerating curve to a floor effect, with early decline being steeper for subjects with higher initial scores. (A small portion of the latter effect may be the result of regression to the mean.) This pattern characterized story recall (especially delayed) and geometric figure recall data.

For Delayed Recall of Stories (Verbatim and Gist), rate of change was almost universally negative for the patients with AD. The rate of decline per year for those with AD tended to be a half to a third of their initial test score. A clear floor effect (Figure 4) was apparent in scatter-plots, and, for a given initial test score, rates of decline per year were less across longer spans of time. The linear effect of time and its interaction with initial level scores accounted for 23% to 30% (depending on time span examined) more variance in the cognitive test scores across subjects and time, beyond that due to differences in subjects' initial test scores. The total percentage variance accounted for by time and initial score combined was 57% to 73% (the lower percentages occurred in assessing longer time spans). Scores for Immediate Story Recall (Verbatim) for patients with AD showed change patterns similar to those for Delayed Story Recall, but the rate of change and percentage variance figures were lower (Table 4). Patterns of change for patients with AD on Geometric Figure (Easy and Hard) Recall were similar to those for Story Recall. The rate of decline per year for patients with AD was about a quarter to a third of their initial level. The effect of time and its interaction with initial score accounted for 8% to 18% of the variance in scores (across subjects and testing) beyond that accounted for by the initial score itself, bringing the total variance accounted for to 59% to 66% (Table 4).

Change Pattern B: Straight Line, Varying Slope

Scores for cognitive tests following this pattern declined in a straight line but with steeper slopes for subjects with higher initial scores. This pattern was characteristic of data for the Verbal Fluency, Benton Visual Retention, Stroop Color and Word, and Wechsler Adult Intelligence Scale–Revised Picture Arrangement tests. There was no floor effect for these tests, except for some tendency in the case of the Picture Arrangement test (Table 4 and Figure 4).
Change Pattern C: Straight Line

Scores for cognitive tests following this pattern declined in a straight-line manner independent of initial score. This pattern was exhibited by the Boston Naming Test, which showed a trend of decreasing about three points per year across all time spans analyzed, and the Mental Rotation Test. There was little or no evidence of a curvilinear relationship with time for either of these measures (Table 4 and Figure 4).

Atypical Change Patterns

Raven’s Coloured Progressive Matrices, as well as the Money Road Map test, did not show strong or simple change over time. For Raven’s Coloured Progressive Matrices, change over time for patients with AD was not substantial compared with random error fluctuation. The change that did occur did not appear to be heavily dependent on initial score (Table 4). On the Money Road Map test, patients with AD scored significantly lower at initial testing than did the NCS, even though 10% of the patients with AD and most of the NCS had the maximum score. Trends across time for patients with AD were weak and primarily showed regression to the mean for those who scored high initially (Table 4).

Measures of AD Severity

The IMC and ADL showed linear increases essentially independent of initial level. Rate of change for the IMC for patients with AD across time in the study was about a three-point increase per year for the first 2 years, with the rate tapering somewhat to an average of two to three points per year across the third and fourth years of the study, apparently showing some ceiling effect (Table 4 and Figure 4). The ADL score increased about 10% per year with a slight but insignificant tendency to show a diminished rate of change as the study progressed (Table 4).

COMMENT

COGNITION

The results of our study present a comprehensive and cohesive picture of cognition in AD. Patients with AD performed significantly worse than did the NCS on all cognitive tests employed, but the tests differed in their detection, staging, and tracking abilities. Patients with AD at any level of dementia severity were discriminated best from NCS by tests of explicit memory, ie, those for delayed story recall and delayed geometric figure recall. Early in the course of the disease, the performance of patients with AD on these tests deteriorated sharply and reached a floor. Thus, these tests were good for diagnosing dementia but were poor for staging severity of dementia. Staging was best achieved, though not as well as discrimination of NCS from subjects with AD, with the Boston Naming, Verbal Fluency, and Benton Visual Retention tests. In the analyses of change over a span of time, those tests that were best for staging dementia severity were the same tests that had a relatively linear decline during progression of illness. The linear relationship of the Boston Naming, Verbal Fluency, and Benton Visual Retention tests to IMC scores among patients with AD further adds to a unified picture of the staging and tracking capabilities of these tests.

With regard to the number of tests providing optimal discrimination, the NCS could be discriminated from those with moderate or severe AD with a single cognitive test (Delayed Recall of Stories [Gist]). All other pairwise discriminations (those with mild AD vs NCS, and among AD severity groups) were improved somewhat by using a combination of two or more tests (Table 3 and Figure 1).

In summary, the results of our study indicate that tests of recall of recent information, whether verbal or pictorial, are best for detecting mild or more severe AD, but are poor for staging the disease because of early floor effects. Tests of confrontation naming, word fluency, and immediate pictorial recognition are best for staging AD.

Our observations confirm and extend results of other
Figure 2. Box-whisker plots displaying distributions of scores for cognitive tests for normal control subjects (NCS) and patients with Alzheimer's disease (AD), with mild, moderate, and severe dementia (using all available records). In a box-whisker plot, the bottom and top edges of the box are at the 25th and 75th percentiles, respectively, with whiskers extending as far as the data extend, up to 1.5 interquartile ranges from the box. Values more extreme up to three interquartile ranges are indicated with a zero, and scores even more extreme are indicated with a dot. The central horizontal line represents the sample median. For a given cognitive test, boxes and whiskers that show relatively little overlap indicate strong differences between groups. NYU indicates New York University.

studies on cognition in AD.13,18-40 In a series of studies employing the data set from the Consortium to Establish a Registry for Alzheimer's Disease, Welsh et al14,15 found that a test of delayed recall of verbal material discriminated patients with mild AD from age-, education-, and gender-matched NCS but was poor for staging, whereas the converse was true for a measure of fluency (for animals), similar to the findings of the present study. In a 2-year longitudinal study,16 scores on the Boston Naming Test and a test of visual recognition (of household objects) declined significantly in patients with AD, whether they were initially rated as having early or advanced AD, whereas measures of verbal recall showed deterioration only in those patients originally judged as having early AD. These results are consistent with those of the present study, showing a linear decline throughout levels of AD severity for the Boston Naming and Benton Visual Retention tests, and a bottoming out for measures of verbal recall. Other studies have also found performance on confrontation naming tasks to decline steadily with the progression of AD.23,34 Hill et al34 reported that recognition but not recall memory was the best test to distinguish mild AD from severe AD. They also found a large decrease in confrontation naming and visuospatial abilities in moderate dementia. Not unlike our results, Munsch et al34,35 found that tests of verbal fluency discriminated patients with AD from age- and education-matched NCS; however, the power to discriminate was greater in the case of fluency for categories (semantic fluency) than for words beginning with specified letters (symbolic fluency). Those authors interpreted this difference as likely due to category fluency’s being relatively more dependent on intactness of semantic knowledge, which deteriorates severely in AD.35 Because the present study assessed only fluency for animals, we cannot generalize the staging effects we found to symbolic fluency. In addition, studies by Chan et al44,45 investigated differences in the semantic network of patients with AD and NCS based on their fluency for animals and their ratings of animal similarities and found that the structure of semantic memory was abnormal in AD.

Our report is one of the few longitudinal studies to examine change over a period of time in a comprehensive set of cognitive tests administered to patients with AD, and the only one to examine closely different patterns of change in test performance over a period of time. Much of the prior information about cognition in longitudinal studies centered on brief tests of mental status. There is broad consensus46-48 that mental status scores worsen in a steady linear fashion with time, regardless of the specific test. The current study confirmed our previous finding of a linear increase in dementia severity over time in the IMC portion of the BDS and extended the findings by documenting a similar linear increase in functional dependence, as measured by the ADL scale. The IMC and ADL scores were highly correlated \( r = 0.33, P < 0.0001 \) and worsened steadily at predictable rates, regardless of the level of impairment noted at initial measurement. Brief mental status and ADL scores are useful as global measures of deterioration but do not provide details about specific cognitive functions nor offer insights into suspected neuropathologic areas. Our analyses in this report go beyond
these summary conclusions and document a distinctive pattern of change for each cognitive test.

Scores in the NCS group were stable over the period studied. Although we tested only a small number of NCS, our observation is consistent with other reports of stability in the test scores of subjects without dementia. It was no surprise to find that scores for patients with AD were significantly poorer on all tests than were those for NCS and that the score for all tests for patients with AD deteriorated. Our study indicates, however, that test scores do not decline in parallel, and that cognitive tests differ in detecting the progress of dementia. By uncovering different patterns of decline on cognitive tests, our study specified which tests are most sensitive to change over a period of time.

MEASUREMENT PROPERTIES

It is important to assess whether the different patterns of change in cognitive test scores are due in whole or in part to dissimilarities in the tests’ measurement properties rather than differential change in the respective cognitive functions they purport to assess. One issue is whether the effective differences between the tests are simply in their level of difficulty. It may be that tests that show floor effects are tests that are too difficult, whereas tests that show steady linear decline over time have a more appropriate range of difficulty for assessment of AD progression over the entire course of the disease. It seems unlikely, however, that differences in the behavior of the tests are solely because of different test difficulty levels. Tests that differ from other tests only in being more difficult would be expected to show floor effects like those found here, but this explanation does not readily account for why tests like Delayed Recall of Stories (Verbatim and Gist) and Geometric Figure Recall are best at differentiating NCS from patients with AD; they should also be difficult for NCS. Further, the only two types of tests studied here that are intended specifically to measure memory for recently acquired information both show the same effects, in spite of the fact that one measures verbal, and the other nonverbal, recall. In addition, for the Geometric Figure Recall Tests, an easy and a hard figure were employed, and although the easy figure appeared to show slightly better discrimination capabilities, the behavior of the two tests was similar (same floor effects, curvilinear trends, discrimination, and moderate correlations), suggesting that the complexity of the figure to be recalled had little bearing on the test’s discrimination or staging capabilities, or lack of same. The New York University Stories Delayed Recognition Test (Table 2) could also be viewed as an easier form of the New York University Delayed Recall test, and this test showed a similar floor effect once the effects of guessing were taken into account.

Another possible methodologic artifact concerns the Boston Naming Test’s relative usefulness for staging and its linear tracking of progression of illness. These characteristics may be partly a reflection of the fact that it was constructed to require word responses of widely varying frequency of usage in the English language or the fact that its pictures vary in how familiar they are to subjects. Items with low-frequency word responses and/or unfamiliar pictures may be good for detecting AD, whereas those with high-frequency words and familiar pictures may be better for staging. Reanalyses of our data showed no support for these hypotheses.

In considering measurement issues, reliability is an obvious consideration. Although differences in reliabil-
Progression of Illness

The pattern of change in test scores over time fits well with the distribution and temporal appearance of lesions in the brains of those with AD. The neurons in the mediotemporal lobe are damaged early in the course of illness.\textsuperscript{5,6,9,10} Damage is especially severe in layers 2 and 4 of the entorhinal cortex and in the CA 1/subicular zone of the hippocampal formation.\textsuperscript{8,9} Layer 2 of the entorhinal cortex provides a major source of cortical afferents to the hippocampus via the perforant pathway.\textsuperscript{11} Neurons in layer 4 of the entorhinal cortex receive input from the CA 1/subicular zone and thereby direct hippocampal output to cortical targets.\textsuperscript{7} Destruction of the entorhinal relay neurons, which partially disconnects the hippocampus from the neocortex, is a postulated cause of amnesia in AD.\textsuperscript{8} The steep slope of declining scores for tests of delayed recall of recently presented verbal and visual information is consistent with damage to the structures in the mediotemporal lobe early in the course of AD.

We speculate that increasing neuronal damage in neocortical regions accounts, at least in part, for the progression in dementia severity. Among neocortical regions, high-order cortices are more severely affected in AD than are modality-specific cortices, which are more affected than primary motor or sensory areas.\textsuperscript{10,11} The abnormalities that correlate best with the signs and symptoms of AD are neurofibrillary tangles (NFTs),\textsuperscript{2,3} decrease in neuronal number,\textsuperscript{2,3} and loss of cortical synaptic connections.\textsuperscript{3,4} In a cross-sectional neuropathologic study, Arriagada et al\textsuperscript{5} found that the number of NFTs in temporal, parietal, and frontal cortices increased significantly with the duration of dementia. Further, there was a significant positive correlation between NFTs scores and the total number of cortical NFTs. Terry et al\textsuperscript{10} reported a 40% decrease of large pyramidal neurons in the frontal and temporal cortices in AD. Hyman et al\textsuperscript{7} found a similar reduction in neurons in the superior temporal sulcus that closely paralleled the duration and severity of dementia in patients with AD. Synaptic loss has been postulated as the physical basis of cognitive impairment in AD, and Terry et al\textsuperscript{10} found a highly significant inverse correlation between the number of synapses in the midfrontal cortex and the IMC score. These data support the view that as the number of cortical neurons and synapses decreases, and NFTs increase, cognitive impairments other than deficits in explicit memory should emerge. In this sense, therefore, patterns of change over time in cognitive tests reflect the underlying spread of neuronal dysfunction that causes progressive dementia.

Staging

The results of the present study suggest that the tests that best characterize stages of dementia severity in AD are those that measure aspects of cognitive function other than explicit memory. From an anatomic standpoint, these functions depend on brain regions other than the struc-
in AD. Our data fit this prediction perfectly: there were progressive impairments in language and visuospatial abilities that we hypothesize result from progressive neuronal dysfunction and death in temporal, parietal, and frontal brain regions.

CLINICAL IMPLICATIONS

To the extent that there is good correspondence between specific cognitive functions and regional brain networks, impaired performance on specific cognitive tests in AD provides a window into the location and extent of the neuropathologic lesions characteristic of AD. Further, the pattern and rate of change in test scores may provide an indirect measure of the distribution of lesions and pace with which their number and extent increases. We found that measures of explicit memory, specifically recalling a brief story and drawing a geometric figure after a 10-minute delay, were the most sensitive tests to distinguish patients with AD from NCS. This finding underscores the central importance of memory loss in the clinical diagnosis of AD. We conclude that tests of explicit memory, including delayed recall, should be the mainstay of the diagnostic examination of individuals suspected of having AD. Such tests could also be used when conducting clinical studies of therapeutic interventions whose goal is to reverse dementia and restore normal cognition. Such tests of explicit memory would not be useful, however, in longitudinal studies of cognition in AD because of floor effects early in the course of dementia.

Cognitive tests were not as efficient in denoting different stages of AD severity as they were in distinguishing patients with AD from NCS. This finding implies that it is difficult to stratify subjects with AD according to dementia severity on the basis of specific cognitive tests. Our results indirectly support the practice of staging dementia severity by global measures, such as the BBS, Mini-Mental State, or Clinical Dementia Rating tests. Despite relatively poor precision in staging, the best aids are the Boston Naming and Verbal Fluency tests that, alone or in combination, were best at distinguishing mild from severe AD as well as moderate from severe AD.

Tests that show linear decline over a period of time are best for charting the course of disease. In addition to the IMC of the BBS and the ADL, the Boston Naming, Verbal Fluency, and Benton Visual Recognition tests all meet this criterion. Changes in these test scores provide accurate information on whether there is progressive dementia and on the speed of deterioration. For detecting change over a period of time, such tests as these that have a linear decline are superior to tests of explicit memory, which have early floor effects. For this reason, such tests should be included as outcome measures in any therapeutic trial that claims to alter or slow the course of cognitive deterioration.

Accepted for publication June 3, 1995.

This research was funded by grants 3-R37-AG06605 and AG05134 from the National Institutes of Health, Bethesda, Md. Dr. Corkin’s work is supported by grants AG06605 and AGNS08117 from the National Institutes of Health.

We thank Jennifer Cohen, Gena Desclous, Elizabeth Hoffman, Marnie Leclerc, Ari Marcus, Beth Souza, Massachusetts General Hospital, and Avital Rodal, Massachusetts Institute of Technology (MIT), Cambridge, for data collection and management; Mark Mapstone, MA, MIT, for assistance in the preparation of tables and figures; Emily Rossie, Jeremiah Jamison, and Gayle Schneider, MIT, for compiling the reference list; and Nancy Hiney, MA, MIT, for help with proofreading.

Reprint requests to Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, E10-003A, Cambridge, MA 02139 (Dr. Locascio).

REFERENCES


