Methods to improve the detection of mild cognitive impairment

William R. Shankle*†‡§, A. Kimball Romney†§¶, Junko Hara*†§, Dennis Fortier†§, Malcolm B. Dick‖, James M. Chen†§, Timothy Chan†§, and Xijiang Sun†§

Departments of *Cognitive Science and ¶Anthropology and ‖Brain Aging Research Unit, University of California, Irvine, CA 92612; and †Medical Care Corporation, Irvine, CA 92612

Contributed by A. Kimball Romney, February 12, 2005

We examined whether the performance of the National Institute of Aging's Consortium to Establish a Registry for Alzheimer's Disease's 10-word list (CWL), part of the consortium's neuropsychological battery, can be improved for detecting Alzheimer's disease and related disorders early. We focused on mild cognitive impairment (MCI) and mild dementia because these stages often go undetected, and their detection is important for treatment. Using standardized diagnostic criteria combined with history, physical examination, and cognitive, laboratory, and neuroimaging studies, we staged 471 community-dwelling subjects for dementia severity by using the Clinical Dementia Rating Scale. We then used correspondence analysis (CA) to derive a weighted score for each subject from their item responses over the three immediate- and one delayed-recall trials of the CWL. These CA-weighted scores were used with logistic regression to predict each subject's probability of impairment, and receiver operating characteristic analysis was used to measure accuracy. For MCI vs. normal, accuracy was 97% [confidence interval (C.I.) 97–98%], sensitivity was 94% (C.I. 93–95%), and specificity was 89% (C.I. 88–91%). For MCI/mild dementia vs. normal, accuracy was 98% (C.I. 98–99%), sensitivity was 96% (C.I. 95–97%), and specificity was 91% (C.I. 89–93%). MCI sensitivity was 12% higher (without lowering specificity) than that obtained with the delayed-recall total score (the standard method for CWL interpretation). Optimal positive and negative predictive values were 100% and at least 96.6%. These results show that CA-weighted scores can significantly improve early detection of Alzheimer's disease and related disorders.

Alzheimer's disease | Consortium to Establish a Registry for Alzheimer's Disease | correspondence analysis

In the United States today, ≈12% of individuals age 65 and over and ≈0.8% of persons 45–65 years old have Alzheimer's disease (AD) or a related disorder (ADRD) (1). ADRD refers to all disorders that can lead to mild cognitive impairment (MCI), which is typically followed by dementia. MCI has been defined in a variety of ways, and there is no universally accepted standard. However, all definitions share the feature of cognitive impairment (usually just one) that does not impair instrumental activities of daily living (e.g., shopping, finances, cooking, household maintenance, and finding familiar locations). Dementia is defined as the presence of two or more areas of cognitive impairment that affect instrumental activities of daily living at the very least.

The most common dementia-related disorders are AD (55–70%), cerebrovascular dementia (15–25%), Lewy body disease and Parkinson's disease (10–15%), frontal lobe dementia (5–10%), and traumatic brain injury (<5%) (2). A comprehensive multifactorial evaluation including clinical assessment, laboratory testing, and imaging is typically used to diagnose ADRD.

The earliest clinical stage of ADRD is classified as MCI. During this stage, an individual's most complex abilities may be compromised but higher-order instrumental activities of daily living such as traveling, paying bills, doing laundry, and balancing a checkbook are spared. Because there is irreversible loss of function for every month that mild to moderate AD goes untreated (3), and because cholinesterase inhibitor treatment reduces the rate of cognitive impairment in AD patients treated for 5 years by ≈50% (3, 4), it is important to detect, diagnose, and treat AD as early as possible (3, 5–9).

MCI and dementia can be measured by using a variety of standardized tools, one of which is the Clinical Dementia Rating (CDR) scale. The CDR has high interrater reliability (10, 11). The clinician using this scale interviews the patient and family, assigns a severity score to each of six CDR subcategories (memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal hygiene), and then applies standard scoring rules to obtain an overall severity score. A CDR score of 0 suggests normal aging, a score of 0.5 indicates MCI, and scores of 1, 2, and 3 indicate mild, moderate, and severe dementia, respectively. A person with mild dementia (MD) is impaired in performing instrumental activities of daily living such as traveling, shopping, paying bills, keeping house, and cooking. A person with moderate dementia is impaired in basic activities of daily living such as dressing, bathing, and toileting.

Although current methods of detecting moderate dementia in community-based clinical practices are reasonably accurate, they do not sensitively detect MCI and often do not detect MD. This insensitivity is because a person with MCI or very MD experiences subtle memory problems greater than normally expected with aging but may not show other symptoms of dementia such as impaired judgment or reasoning. In fact, >67% of individuals are moderately demented at the time of first diagnosis (12, 13). The difficulty in detecting MCI and, in many cases, MD is largely because of the insensitivity of the most commonly used screening test in clinical practice, the MiniMental Status Examination (MMSE). The MMSE is a brief test of several cognitive abilities with a maximum score of 30 points. One of the larger studies designed to differentiate individuals with MCI from those with normal aging showed that the MMSE detected only 30% of 244 subjects classified as MCI according to a CDR score of 0.5 (14, 15). More sensitive screening tests that can be practically applied in community health-care settings are therefore needed.

The National Institute of Aging, founded in 1986, has brought together 24 major medical research centers in the Consortium to Establish a Registry for AD (CERAD). The consortium has developed an extensive battery for evaluating and diagnosing persons with the MCI and dementia stages of ADRD. The

Freely available online through the PNAS open access option.

Abbreviations: AD, Alzheimer's disease; ADRD, AD and related disorders; CA, correspondence analysis; CDR, Clinical Dementia Rating; CERAD, Consortium to Establish a Registry for Alzheimer's Disease; C.I., confidence interval; CWL, CERAD 10-word list; MCI, mild cognitive impairment; MD, mild dementia; MMSE, Mini-Mental Status Exam; ROC, receiver operating characteristic.

‡To whom correspondence should be addressed. E-mail: rshankle@mccare.com.

§W.R.S., A.K.R., J.H., D.F., J.M.C., T.C., and X.S. have a financial interest in Medical Care Corporation.

SOCIAL SCIENCES

Table 1. Sample characteristics for normal aging, MCI, and MCI/MD by source

Diagnosis	CDR	*n*	Samples		Sex, %		Age, yr		Education, yr	
			University	Community	Male	Female	Mean	SD	Mean	SD
Normal aging	0	119	29	90	33	68	58.6	14.8	15.4	4.0
MCI	0.5	95	50	45	44	56	71.6	9.5	14.8	3.6
MD	1	257	119	138	40	60	76.4	8.6	12.6	5.0
Normal aging vs. MCI	0–0.5	214	79	135	39	61	64.9	14.1	15.1	3.8
Normal aging vs. MCI/MD	0–1	471	198	273	40	60	73.5	11.8	13.1	5.0

CERAD battery includes demographic data on subject and informant, clinical history and examinations, extensive neuropsychological exams, laboratory and imaging studies, and neuropathological studies. One of the subtests of this battery, the CERAD 10-word list (CWL), has been shown to be one of the more sensitive tests for detecting MCI (16). The CWL consists of three immediate-recall trials of a 10-word list, followed by an interference task lasting several minutes, and then a delayed-recall trial with or without a delayed-cued-recall trial. The CWL is usually scored by recording the number of words recalled in each of the four trials. A single cutoff score for the delayed-recall trial, with or without adjustment for demographic variables, is typically used to determine whether cognitive impairment exists. This approach, however, may ignore other important information contained in the CWL. For example, the measurement of attention, working memory, learning, retention, and serial position effects may be important in identifying MCI. The research summarized herein tests the hypothesis that additional information contained in the CWL can more accurately distinguish MCI from normal aging.

Methods

Our study used a case-control design to determine the value of using the full performance profile of the CWL for discriminating normal aging and MCI. We also measured classification accuracy for either MCI or MD because these stages usually are not detected in community settings. We used correspondence analysis (CA), a technique that creates weighted scores from the individual performance profiles. CA produces an optimally weighted combination of values, somewhat similar to principal components analysis, which is appropriate for dichotomous data, and maximizes the correlation between dependent (classification) and independent (predictor) variables (17). These CA-weighted scores were then used in a logistic regression to predict each subject's probability of cognitive impairment, which was then used with their true classification to construct nonparametric receiver operating characteristic (ROC) curves. Each ROC curve was used to measure overall classification accuracy and select a cutoff point corresponding to an optimal sensitivity and specificity on the curve (18). These results were validated with split-sample and randomization methods. Finally, we compared these results with those of other studies reported in the scientific literature for MCI vs. normal aging.

Sample Classification and Characteristics. The sample was drawn from two subject pools: a university dementia research clinic and a community dementia clinic. The university sample consisted of subjects who had been either referred to the University of California, Irvine, Alzheimer's Disease Research Center or recruited for normal-aging studies between 1988 and 1997. The community sample consisted of subjects who were self-referred, referred by their physicians, or participating in a normal-aging study. University sample subjects were thoroughly evaluated with complete medical history, patient and caregiver interviews, general physical and neurological exams, and 2 h of cognitive testing using the CERAD neuropsychological test battery plus the Wechsler tests for immediate and delayed recall of visual and verbal information and the Blessed Information, Memory, and Concentration Test. Also measured were abilities to perform instrumental and basic activities of daily living and presence or absence of major depression by standardized diagnostic criteria. Of the subjects with MCI in the university sample, 92% had a cause associated with progressive decline (AD, cerebrovascular disease, Lewy body disease, or frontal lobe dementia).

Because of practical constraints, not all subjects from the community sample received similarly extensive testing. However, all community sample subjects were staged with the CDR and were given a portion of the CERAD battery (Boston naming, F-A-S letter fluency, animal category fluency, figural fluency, constructional praxis, trails A and B, symbol digit modalities test, mental calculations, and CWL). Subjects identified with cognitive or functional impairment received the same diagnostic evaluation as those from the university sample except that when the etiologic diagnosis was not certain, a fluorodeoxyglucose positron-emission tomography brain scan was obtained. The same neurologist performed CDR staging and diagnosed impairment for both samples. Of the subjects with MCI in the community sample, 98% had a cause associated with progressive decline.

Subjects from either the community or the university sample who showed any evidence of cognitive and/or functional impairment were further evaluated with routine dementia laboratory testing and magnetic resonance brain imaging. When diagnosis was unclear, functional brain imaging with either hexamethyl-propyleneamine oxime plus xenon single-photon emission with computed tomography or positron-emission tomography scans were performed. These evaluations were used in conjunction with standardized criteria to identify the underlying causes of the cognitive impairment or dementia (19–21).

To stage dementia severity with the CDR, a neurologist and neuropsychologist independently assigned scores to each subject by using the interview-based approach previously described. Any differences in their CDR score assignments for a given subject were resolved through discussion and, if necessary, case review. CDR staging did not use the results of performance on the CWL. The combination of comprehensive tests used in the evaluation of the two samples was designed to ensure accurate classification of individuals by severity level. Table 1 summarizes sample characteristics.

Analysis. The primary outcome (dependent) variable used in this study was severity stage (normal, MCI, MD, or moderate dementia) based on the CDR. Primary predictor (independent) variables were the subjects' CA-weighted scores derived from the 40 CWL items, 10 words for each of the three immediate-recall trials and one delayed-recall trial. For the MCI vs. normal-aging group, we also compared the total CWL recall scores with those obtained from the CA-weighted scores. This latter comparison provided a measure of the increase in test performance that can be obtained by CA compared with unadjusted total scores. Demographic measures of age, gender, and education were

included after the primary analysis to determine their potential contribution to classification accuracy.

The three immediate-recall trials of the CWL measure attention and working memory (dorsolateral prefrontal cortex) plus learning and serial position effects. The delayed-recall trial measures retention and delayed-retrieval effects to assess the functional capacity of the entorhinal cortex and hippocampus, which are the first cortical structures to be affected by AD.

To code each subject's CWL data for use with CA, each word (item response) that a subject recalled in each trial was assigned a value of 1; words not recalled in a trial were assigned a value of 0. This coding created a row of 40 0s and 1s for each subject's CWL data. However, to create weighted scores for recalling and not recalling a given word, CA requires two binary response variables (recalled = yes/no, not recalled = yes/no) to be created for each word in each trial. The sum of these two binary variables for each word is always 1, and each subject's row total is always 40. The input data indicator matrix used with CA therefore consisted of 471 subject rows by 80 binary response columns.

CA was then applied to this indicator matrix to create a set of orthogonal, weighted scores for each subject. The CA-weighted scores maximize the correlation between the CDR score (rows) and the CWL item responses (columns). Kendall and Stuart (17) mathematically demonstrated that these CA-weighted scores give the best linear solution to explaining the total variance of the data, always performing as well as or better than the raw scores from which they were derived.

The CA-weighted scores were then used as input to develop and test the classification model for each severity-group comparison against the normal-aging group. Because the dependent variable is binary (impaired vs. normal), it is reasonable to model it with a binomial distribution in which all subjects in a group have a uniform *a priori* probability of being assigned to that group. An appropriate classification model is therefore a log-linear (logistic) odds-ratio regression with some number of independent predictor variables. Age, gender, and years of education were tested in the model with and without the CA-weighted subject scores to determine the impact of demographic variables on classification accuracy.

The model resulting from logistic regression was then used to obtain the predicted probability of being impaired (STATA 8.0, predict), which when combined with the subject's true classification, was used as input to the ROC analysis to compute overall classification accuracy and its exact, asymptotic, binomial 95% confidence interval (C.I.). The advantage of the ROC method is that it generates an entire curve of sensitivity-specificity values that can be used to select the best combination for a given purpose. The area under the ROC curve is a measure of the overall accuracy of any given classification method. A method with 100% sensitivity and 100% specificity has an area under the ROC curve of 1.00. Therefore, the closer this area is to 1.00, the more accurately the given method classifies the groups of interest.

To test the robustness of the results for both MCI and MCI–MD vs. normal aging, we used three different validation methods. First, we trained the model on the university sample and tested it on the community sample. Second, we trained the model on the community sample and tested it on the university sample. Third, we randomly assigned 67% of the full sample of impaired (MCI or MCI–MD) and normal-aging subjects for training and used the remaining 33% for testing. We randomly assigned the full sample 25 times in this way to derive 95% C.I. for sensitivity, specificity, and overall classification accuracy.

For comparative purposes, the randomization validation method was also used to measure classification results based on the total recall scores summed across all four trials of the CWL as well as on the total scores of the delayed-recall trial. The most common method of interpreting the CWL results is simply to use the total delayed-recall score.

Finally, to examine the quality of the fit of the classification model to the data, we looked for covariate patterns in which subjects were not consistently classified to either the normal-aging or the MCI group. Each subject can be described by a pattern of covariance between predicted classification and predictor variables. If some of the subjects with a given covariate pattern do not share the predicted classification, then this pattern will weaken the fit of the classification model. To identify such covariate patterns, we plotted the Hosmer and Lemeshow χ^2 statistic (22) against the probability of being classified as MCI for each subject. This statistic measures the decrease in the Pearson χ^2 goodness-of-fit statistic that would be caused by deleting all subjects having a given covariate pattern. A model with a very good fit shows two curves, one containing all normal subjects and one containing all MCI subjects, with no data points (covariate patterns) lying off of these curves. Data points lying off of these curves indicate covariate patterns belonging to subjects who are not consistently classified as either all normal or all MCI. When such anomalous data points appear, they indicate that the classification model cannot resolve all covariate patterns and could be improved.

To assess how well the MCI vs. normal-aging classification model performed relative to other screening tests, we compared the results with those reported in the scientific literature (see Table 4). To do this, we ran a PubMed search for MCI or cognitive impairment no dementia to identify and review all articles published on or before November 30, 2004, which studied normal aging vs. MCI and published sensitivity, specificity, and sample size (14, 23–29). Studies were excluded if they did not report these values for a normal-aging vs. MCI sample or included subjects with MD or moderate dementia because they would spuriously raise sensitivity.

Results

With logistic regression, only the first two CA-weighted scores were significant predictors of a subject's CDR score. They accounted for 16.5% of the 18% of the variance explained by age and education (gender had no effect) with the remaining 1.5% explained by age alone. However, age did not improve accuracy above that obtained with just the CA-weighted scores. Therefore, we only used the first two CA-weighted scores of each subject to compute their predicted probability of being impaired (STATA 8.0, predict p), which was then used with their true classification to perform the ROC analysis.

Table 2 displays the results of the ROC analysis using the first two split-sample validation methods described (university and community validation samples). We determined sensitivity and specificity by selecting the cutoff point along each ROC curve that maximized sensitivity and kept accuracy maximal or near maximal to avoid excessively reducing specificity. The sensitivity for MCI was 95%, specificity for normal aging was 88%, and overall accuracy was 97%.

Table 3 displays the results of the validation method with 25 randomized samples. The 95% C.I. for these results is somewhat narrower than those for the community and university validation samples displayed in Table 2. Because the randomization validation method gives a more formal, precise estimate of the 95% C.I., its values will be used from here on. Sensitivity for MCI was 94%, specificity for normal aging was 89%, and overall accuracy of the ROC curve was 97%. Sensitivity for detecting MCI–MD was 96%, specificity for detecting normal aging was 91%, and overall accuracy of the ROC curve was 98%. Fig. 1 shows the overall accuracy of the ROC curve for distinguishing MCI from normal aging with the randomization validation method.

In terms of positive predictive value (PPV, the probability that a person with a positive test result has MCI) and negative

Table 2. Full and split sample results based on the CWL CA-weighted scores

Comparison	*n*	ROC overall accuracy	95% C.I.	Sensitivity	Specificity
MCI vs. normal	214	97	(94, 99)	95	88
Community validation	135	97	(93, 99)	91	89
University validation	79	98	(93, 100)	98	90
MCI/MD vs. normal	471	98	(97, 99)	97	88
Community validation	273	97	(95, 99)	93	90
University validation	198	99	(97, 100)	98	90
MD vs. normal	376	99	(98, 100)	96	99
Community validation	228	95	(91, 97)	94	92
University validation	148	100	(98, 100)	99	97

predictive value (NPV, the probability that a person with a negative test result is normal), we derived their optimal results by combining the full range of sensitivity/specificity data points on the ROC curve with prevalence estimates for normal aging and MCI in persons <65 and ≥65 years old. We then selected the sensitivity/specificity data point that gave the best combination of PPV and NPV. The prevalence values used for normal and MCI in the <65 group were 0.975 and 0.025, respectively, and the corresponding values for those in the ≥65 group were 0.89 and 0.11. For the <65 group, the optimal PPV and NPV were 100% and 99.3%, respectively, and for the ≥65 group, they were 100% and 96.6%.

The total recall score summed over all four CWL trials gave a mean sensitivity of 85% for MCI (95% C.I. 83–87%) and a mean specificity of 91% for normal aging (95% C.I. 90–92%). These results were 9% less sensitive than those obtained with the CA-weighted scores with minimal change in specificity. The delayed-recall total score gave a mean sensitivity of 82% for MCI (95% C.I. 79–85%), and mean specificity of 91% for normal aging (95% C.I. 89–92%). These results were 12% less sensitive than those obtained with CA-weighted scores with minimal change in specificity.

Examination of the quality of the fit of the classification model to the data using the Hosmer and Lemeshow χ^2 statistic showed that all data points lay either on the curve containing the normal-aging covariate patterns or on the curve containing the MCI covariate patterns. This finding means that the subjects belonging to each covariate pattern are all classified the same, whether MCI or normal-aging. The model therefore gives a consistent classification of each covariate pattern in the data set and is consistent with the high sensitivity and specificity found.

As shown in Table 4, the use of CA-weighted scores for the CWL item response data produced at least a 9% higher sensitivity (94%) than any other published screening study of MCI vs. normal aging. The 89% specificity for normal aging was exceeded only by tests with sensitivity of 85% or lower. The combination of very high sensitivity and no loss of specificity is reflected in the very large area under the ROC curve (97%), which indicates very accurate classification of MCI and normal aging. Additionally, CA-weighted scores compared with total delayed-recall scores (the most commonly used method for interpreting the CWL) were 12% more sensitive in detecting MCI, while having about the same specificity. CA-weighted scores significantly improve normal vs. MCI classification with the CWL.

Following the search methods described, we found no published studies of MCI vs. normal aging for the following commonly used screening tests: ADAS-Cog, Buschke Selective Reminding Test, California Verbal Learning Test, CANS-MCI, Cognitive Abilities Screening Instrument (CASI), IQCODE, Memory Impairment Screen, Minnesota Cognitive Acuity Screen (MCAS), MiniCog, New York University delayed paragraph recall, Rey Auditory Verbal Learning Test, Short Test of Mental Status, SISCO, Telephone Instrument for Cognitive Screening (TICS), and Wechsler Logical Memory Scale.

To look for differences in the difficulty of recalling the 10 words of the CWL, we plotted the CA-weighted column scores for each trial (Fig. 2). This plot provided a way to examine to what degree primacy and recency effects (words at the beginning and end of the list, respectively) on word recall difficulty were represented in the CA-weighted column scores. If there were no effects of word position in the list (serial position effect), then all 10 words in a given trial would have the same recall difficulty. If there were no effects of the number of trials to learn a given word, then all trials would have the same recall difficulty for that word.

Fig. 2 shows that word recall difficulty is heterogeneous both across trials and across word positions. For the immediate-recall (learning) trials, trial 1 is most difficult for all words except words 9 and 10 (the recency effect). The delayed-recall trial is the most difficult of all, presumably because of the absence of further encoding of the 10-word list for several minutes before performing retrieval. Words in the middle of a list are more difficult to recall than words from the beginning (primacy effect) or the end (recency effect). A simple summative score of the number of words recalled would lose this information, thus reducing classification accuracy.

Discussion

As mentioned, the delayed-recall score of the CWL is reported to be somewhat sensitive for detecting the earliest stages of ADRD and has been used by the National Institute of Aging CERAD centers for >20 years (16, 30). The sensitivity of delayed recall is high because it measures entorhinal and hippocampal cortical function, where the earliest neuropathological changes in AD occur (31). In detecting subtle entorhinal or hippocampal dysfunction, measuring encoding may be more

Table 3. Results of the randomization validation method based on CA-weighted scores of the CWL

Comparison	*n*	ROC overall accuracy	95% C.I.	Sensitivity	95% C.I.	Specificity	95% C.I.
Normal vs. MCI	214	97	(97, 98)	94	(93, 95)	89	(88, 91)
Normal vs. MCI/MD	471	98	(98, 99)	96	(95, 97)	91	(89, 93)

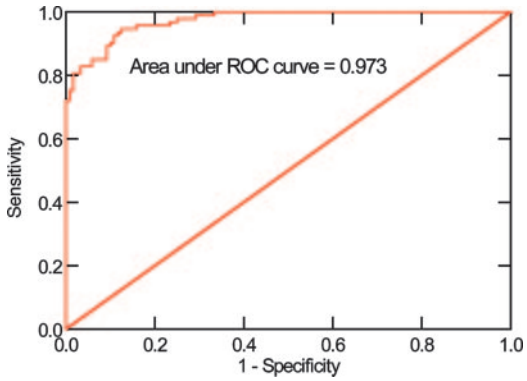

Fig. 1. Overall accuracy of classifying MCI vs. normal. The nonparametric ROC curve for MCI (sensitivity) vs. the false-alarm rate for normal aging (1 − specificity) was generated by applying CA and logistic regression to the item response indicator matrix of the CWL data by using the methods described in the text. A perfect fit has an area of 1.0. The fit of this ROC curve is 0.973 (95% C.I. 0.97, 0.98), giving an overall accuracy of 97.3%.

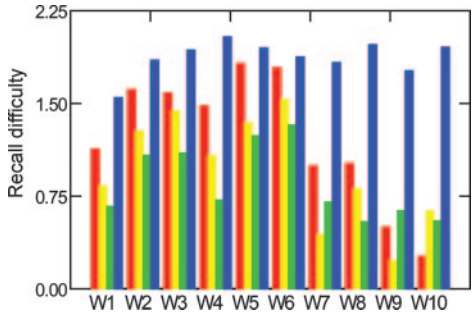

Fig. 2. Effect of word position and trial number in recalling words from the CWL. The difficulty of recalling each word in each trial is shown as a bar graph according to the serial position in the word list (*x* axis) and trial number coded by color (red, yellow, green, and blue). Shown are the effects of serial position (primacy, recency) and repeated exposure across trials. There is marked heterogeneity of recall difficulty as a function of both serial position and repeated exposure (trial number).

important than retrieval because analysis of the "people and doors" test showed no difference in classification accuracy between delayed recall (which requires that a word be previously encoded to be retrieved) and delayed-recognition (which eliminates retrieval and simply measures whether the word was encoded) (32). Disorders such as cerebrovascular disease, depression, and Lewy body disease, in which delayed recall is impaired but delayed recognition is intact, indicate a dysfunction of retrieval that is presumably caused by disrupted connections to the entorhinal-hippocampal circuit without damage to the circuit itself.

Encoding occurs during the immediate-recall trials of the CWL, and its persistence is measured with delayed-recall and delayed-recognition tasks. Our results suggest that the immediate-recall trials have encoding and/or retrieval information that enhances the power of delayed-recall measures to detect MCI. The relatively higher sensitivity of the present study's results compared with other studies of normal vs. MCI is likely caused by the use of all of the encoding and retrieval measures in the CWL data, including weighting of each word's relative importance by its position in the list and by the trial in which it is recalled. The CA-weighted column scores of the CWL data measure the difficulty of encoding and retrieval for each word in each trial. A simple summation or cutoff score of the number of words recalled across the four trials would not account for such weightings of encoding and retrieval difficulty.

When we separated MCI subjects into AD and non-AD diagnoses, sensitivity was higher for the AD group, suggesting that increasingly precise measures of encoding can improve detection of early entorhinal-hippocampal dysfunction in AD. A larger sample would enable more complete analysis of MCI AD and MCI non-AD.

Efficacy of CA. With ≈95% of the MCI subjects having a diagnosis that would produce progressive decline, the high sensitivity in the study means that many non-AD diagnoses also show early changes in encoding and/or retrieval that differ from normal aging. The implication of an abnormal screening result based on the randomization validation method is that it is correct in ≈94% of MCI cases, most of which are progressive, and incorrect (a false positive result) in ≈11% of normal-aging subjects. The implication of a normal screening result is that it is correct in ≈89% of all normal-aging subjects and incorrect (a false negative result) in ≈6% of MCI cases. If one uses the optimal positive and negative predictive value results derived from the ROC curve and population prevalence estimates of normal and MCI, the findings are even more striking. In this case, the probability

Table 4. Studies of sensitivity and specificity of tests for MCI vs. normal

MCI vs. normal-aging test	*n*, MCI	*n*, normal	*n*, total	Sensitivity	Specificity
CWL with CA-weighted scoring*	95	119	214	94	89
CWL total delayed-recall score*	95	119	214	82	91
Object delayed recall with proactive interference	53	53	106	85	89
Modified MMSE†	24	52	76	83	90
Cognitive Capacity Screening Exam (CCSE)	47‡	267	314	74	85
	84	267	351	88	84
MMSE + CCSE	47‡	267	314	83	80
DemTect	97	97	194	80	92
Computerized Dementia Screen (Korean)	41	103	144	76	94
Clock drawing test	48	41	89	75	76
MMSE	24	52	76	71	85
	47	267	314	61	80
7-min test	25	35	60	28	?
Functional activities questionnaire	244	198	442	20	99

*Data from the present study using the randomization validation method (see *Methods*).
†The MMSE was modified by including multiple delayed recall trials.
‡This MCI sample was restricted to those with a National Institute of Neurological Disorders and Stroke/Alzheimer's Disease and Related Disorders Association diagnosis of possible AD.

that a person with a positive screening test result has MCI is 100% and the probability that a person with a negative screening test result is normal is at least 96.6%. These rather straightforward interpretations can provide guidance for busy clinicians as to whether to proceed with diagnosis and treatment.

Potential Sources of Error. Although this is a case-control study, the sample of MCI subjects reflects the same proportions of disorders commonly seen in dementia. The percentage of MCI cases with progressive decline was comparable to that reported for dementing disorders (33). The patterns of missing data in the two samples we studied can introduce error if those patterns are not missing at random (34). A third source of potential error is common to almost all medical research and is caused by the difficulty of randomly sampling a population. One source of error that was reduced by the inclusion of a community sample was the selection bias that occurs in academic, tertiary referral centers. A source of error that arises from the use of community and academic samples is caused by their differences in how extensively subjects are evaluated. Community samples may misclassify more subjects because they are less extensively studied, which may explain the slightly lower sensitivities and specificities of the community validation sample. Another potential source of error in both samples was caused by the use of fewer diagnostic tests (the more expensive ones) in the normal-aging subjects. Such an error source is difficult to completely eliminate because of associated costs and would increase the false negative rate and lower sensitivity. However, the sensitivity and specificity of our results is comparable to those obtained from positron-emission tomography studies, suggesting that including these more expensive tests would neither significantly increase false negative rates nor lower sensitivity.

Conclusions

CA can substantially improve the CWL, which has been internationally validated and is sensitive for detecting early stages of ADRD. Compared with the usual method of scoring and interpreting the CWL, the CA-weighted scores derived from the item responses increased sensitivity in detecting MCI by 12% while preserving high specificity. Following the literature search method for identifying normal vs. MCI studies, we found our results to be the highest reported. Because most other screening tests rely primarily on total scores that are not adjusted to maximize their explanation of the variance, they could be improved by incorporating the methods presented here.

We gratefully acknowledge useful criticism and feedback from Dr. Robert Katzman (University of California at San Diego) and Dr. Susan C. Weller (University of Texas Medical Branch, Galveston). We thank the Medical Care Corporation, which specializes in the early detection, diagnosis, and treatment of ADRD, for making its assessment tool available for the research done in this study. This research was supported in part by the Medical Care Corporation, an Alzheimer's Association grant (to W.R.S.), a center grant from the National Institute of Aging, and National Science Foundation Grant SBR-9631213 (to A.K.R. and William H. Batchelder).

1. Kawas, C. & Katzman, R. (1999) in *Alzheimer's Disease*, eds. Terry, R. D., Katzman, R., Bick, K. L. & Sisson, R. A. (Lippincott, Philadelphia), pp. 95–117.
2. Clarfield, A. M. (2003) *Arch. Intern. Med.* **163,** 2219–2229.
3. Doody, R. S., Geldmacher, D. S., Gordon, B., Perdomo, C. A. & Pratt, R. D. (2001) *Arch. Neurol.* **58,** 427–433.
4. Farlow, M. R. (2002) *Int. J. Clin. Pract.* Suppl., **127,** 37–44.
5. Desai, A. & Grossberg, G. (2001) *Exp. Opin. Pharmacother.* **2,** 653–666.
6. Farlow, M., Potkin, S., Koumaras, B., Veach, J. & Mirski, D. (2003) *Arch. Neurol.* **60,** 843–848.
7. Farlow, M., Anand, R., Messina, J., Jr., Hartman, R. & Veach, J. (2000) *Eur. Neurol.* **44,** 236–241.
8. Cummings, J. L. (2003) *Am. J. Geriatr. Psychiatry* **11,** 131–145.
9. Geldmacher, D. S., Provenzano, G., McRae, T., Mastey, V. & Ieni, J. R. (2003) *J. Am. Geriatr. Soc.* **51,** 937–944.
10. McCulla, M. M., Coats, M., Van Fleet, N., Duchek, J., Grant, E. & Morris, J. C. (1989) *Arch. Neurol.* **46,** 1210–1211.
11. Marin, D. B., Flynn, S., Mare, M., Lantz, M., Hsu, M. A., Laurans, M., Paredes, M., Shreve, T., Zaklad, G. R. & Mohs, R. C. (2001) *Int. J. Geriatr. Psychiatry* **16,** 745–750.
12. Callahan, C. M., Hendrie, H. C. & Tierney, W. M. (1995) *Ann. Intern. Med.* **122,** 422–429.
13. Gifford, D. R. & Cummings, J. L. (1999) *Neurology* **52,** 224–227.
14. Shankle, W. R., Datta, P. & Pazzani, M. (1996) *Alzheimer's Res.* **2,** 95–99.
15. Shankle, W. R., Mani, S., Pazzani, M. J. & Smyth, P. (1997) in *Intelligent Data Analysis in Medicine and Pharmacology*, eds. Keravnou, E., Lavrac, N. & Zupan, B. (Kluwer, Boston), pp. 73–86.
16. Larumbe, R. (1997) *Rev. Med. Univ. Navarra* **41,** 6–11.
17. Kendall, M. G. & Stuart, A. (1961) *The Advanced Theory of Statistics* (Griffin, London).
18. Green, D. M. & Swets, J. A. (1966) *Signal Detection Theory and Psychophysics* (Wiley, New York).
19. Kosunen, O., Soininen, H., Paljarvi, L., Heinonen, O., Talasniemi, S. & Riekkinen, P. J., Sr. (1996) *Acta Neuropathol.* **91**, 185–193.
20. Chui, H. C., Victoroff, J. I., Margolin, D., Jagust, W., Shankle, R. & Katzman, R. (1992) *Neurology* **42,** 473–480.
21. American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders* (Am. Psychiatric Assoc., Washington, DC), 4th Ed, Text Revision.
22. Hosmer, D. W. & Lemeshow, S. (1980) *Commun. Stat.* **A9**, 1043–1069.
23. Loewenstein, D. A., Acevedo, A., Luis, C., Crum, T., Barker, W. W. & Duara, R. (2004) *J. Int. Neuropsychol. Soc.* **10,** 91–100.
24. Loewenstein, D. A., Barker, W. W., Harwood, D. G., Luis, C., Acevedo, A., Rodriguez, I. & Duara, R. (2000) *Int. J. Geriatr. Psychiatry* **15,** 434–440.
25. Xu, G., Meyer, J. S., Thornby, J., Chowdhury, M. & Quach, M. (2002) *Int. J. Geriatr. Psychiatry* **17,** 1027–1033.
26. Kalbe, E., Kessler, J., Calabrese, P., Smith, R., Passmore, A. P., Brand, M. & Bullock, R. (2004) *Int. J. Geriatr. Psychiatry* **19,** 136–143.
27. Cho, B., Yang, J., Kim, S., Yang, D. W., Park, M. & Chey, J. (2002) *J. Neurol. Sci.* **203–204**, 109–114.
28. Yamamoto, S., Mogi, N., Umegaki, H., Suzuki, Y., Ando, F., Shimokata, H. & Iguchi, A. (2004) *Dement. Geriatr. Cogn. Disord.* **18,** 172–179.
29. Kalman, J., Boda, K., Bende, Z. & Janka, Z. (2003) *Orv. Hetil.* **144,** 1929–1938.
30. Batchelder, W., Chosak, J. & Shankle, W. R. (1997) *J. Gerontol. B Psychol. Sci. Social Sci.* **52,** 206–215.
31. Squires, L. (2000) in *The Oxford Handbook of Memory*, eds. Craik, F. & Tulving, E. (Oxford Univ. Press, Oxford), pp. 490–491.
32. Greene, J. D., Baddeley, A. D. & Hodges, J. R. (1996) *Neuropsychologia* **34,** 537–551.
33. Clarfield, A. M. (1988) *Ann. Intern. Med.* **109,** 476–486.
34. Little, R. J. A. & Rubin, D. B. (1987) *Statistical Analysis with Missing Data* (Wiley, New York).