Published in final edited form as:

Alzheimer Dis Assoc Disord. 2012 July ; 26(3): 225–231. doi:10.1097/WAD.0b013e318231cd30.

The SIST-M: Predictive validity of a brief structured Clinical Dementia Rating interview

Olivia I. Okereke, MD, SM¹, Norberto Pantoja-Galicia, PhD², Maura Copeland, MD¹, Bradley T. Hyman, MD, PhD³, Taylor Wanggaard, BA¹, Marilyn S. Albert, PhD⁴, Rebecca A. Betensky, PhD², and Deborah Blacker, MD, ScD^{1,3}

¹Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

²Department of Biostatistics, Harvard School of Public Health, Boston, MA

³Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

⁴Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD

Abstract

Background—We previously established reliability and cross-sectional validity of the SIST-M (Structured Interview and Scoring Tool–Massachusetts Alzheimer's Disease Research Center), a shortened version of an instrument shown to predict progression to Alzheimer disease (AD), even among persons with very mild cognitive impairment (vMCI).

Objective—To test *predictive* validity of the SIST-M.

Methods—Participants were 342 community-dwelling, non-demented older adults in a longitudinal study. Baseline Clinical Dementia Rating (CDR) ratings were determined by either: 1) clinician interviews or 2) a previously developed computer algorithm based on 60 questions (of a possible 131) extracted from clinician interviews. We developed age+gender+education-adjusted Cox proportional hazards models using CDR-sum-of-boxes (CDR-SB) as the predictor, where CDR-SB was determined by either clinician interview or algorithm; models were run for the full sample (n=342) and among those jointly classified as vMCI using clinician- and algorithm-based CDR ratings (n=156). We directly compared predictive accuracy using time-dependent Receiver Operating Characteristic (ROC) curves.

Results—AD hazard ratios (HRs) were similar for clinician-based and algorithm-based CDR-SB: for a 1-point increment in CDR-SB, respective HRs (95% CI)=3.1 (2.5,3.9) and 2.8 (2.2,3.5); among those with vMCI, respective HRs (95% CI) were 2.2 (1.6,3.2) and 2.1 (1.5,3.0). Similarly high predictive accuracy was achieved: the concordance probability (weighted average of the area-under-the-ROC curves) over follow-up was 0.78 vs. 0.76 using clinician-based vs. algorithm-based CDR-SB.

Corresponding author: Olivia Okereke, MD, SM, MGH Gerontology Research Unit, 149 13th St., Suite 2691, Charlestown, MA 02129, Tel: (617) 726-5571, Fax: (617) 726-5760, ookereke@partners.org.
Drs. Okereke, Pantoja-Galicia and Betensky completed all statistical analyses.

Statements of Disclosure

Dr. Pantoja-Galicia reports current employment by the U.S. Food and Drug Administration. He reports no financial disclosures or conflicts.

Dr. Copeland reports no disclosures.

Ms. Wanggaard reports no disclosures.

Dr. Betensky reports no disclosures.

Conclusion—CDR scores based on items from this shortened interview had high predictive ability for AD – comparable to that using a lengthy clinical interview.

Keywords

Alzheimer disease; mild cognitive impairment; dementia; CDR; instrument; questionnaire; validity; prediction; psychometric

Introduction

The Clinical Dementia Rating (CDR)^{1, 2} and CDR sum-of-boxes (CDR-SB, the total of CDR ratings from 6 cognitive and functional domains) are effective at distinguishing cognitive status along the spectrum encompassing normal aging, mild cognitive impairment (MCI) and mild dementia. Indeed, the CDR is a mandatory element of the National Institute on Aging-funded Alzheimer's Disease Centers/Alzheimer's Disease Research Centers (ADCs/ADRCs) Uniform Data Set (UDS)³. Formal CDR interview protocols² have been developed, as well as an expanded interview⁴ that yields high reliability and discriminative ability of the CDR within the range of even very mild cognitive change. In earlier work, we demonstrated that this expanded interview can predict clinical course, even among very mildly impaired individuals who do not meet formal MCI criteria as implemented in most clinical trials⁵. However, its average 90-minute administration time would preclude use in most larger-scale research settings, including multi-center clinical trials.

Therefore, we recently developed the SIST-M (Structured Interview and Scoring Tool-Massachusetts ADRC)⁶ – a systematic, time-efficient method (~25 minutes) for administering the CDR among very mildly impaired persons without sacrificing reliability. The SIST-M was found to have high reliability not only for the global CDR score, as has been achieved with worksheet scoring systems⁷, but also for the CDR-sum-of-boxes (CDR-SB) over a wide range of cognitive symptoms. Key strengths of the SIST-M include: 1) facilitation of reliable quantification of cognitive change, even among people with very mild levels of impairment – which may be of particular value in future trials of preventive interventions targeted at persons with very mild cognitive changes; 2) structuring of the interview and detailed symptom probes of the SIST-M-IR aid in obtaining reports on subtle cognitive changes; 3) the independent-report format of the SIST-M-IR. We previously found that agreement between the SIST-M and expanded interview was good to superior for global CDR and CDR-SB. Our current objective was to evaluate whether the predictive validity of the SIST-M for subsequent development of AD is comparable to that of the expanded clinician interview, particularly among participants in the spectrum of mild impairments—a critical group for early intervention studies. To achieve this, we utilized conventional survival analysis, as well as newer methods of time-dependent Receiver Operating Curve (ROC) analysis⁸.

Methods

Participants

Participants were drawn from a longitudinal study examining preclinical predictors of AD that selected non-demented subjects demonstrating a range of cognitive and functional impairment⁹. Older men and women (n=379) were recruited from the community in three successive cohorts through print advertisements indicating that a research study was seeking individuals both with and without memory difficulty (i.e., rather than from a clinical or other medical referral source): Cohort 1 (N=165) from 1992-93, Cohort 2 (N=119) from 1997-98, and Cohort 3 (N=95) from 2002-06. Volunteer respondents then underwent a multistage screening procedure. To be included participants had to: be 65 years or older (with the

exception of 7 individuals aged 57-64 years); be non-demented; have a Clinical Dementia Rating (CDR) rating of normal (CDR=0) or mildly impaired (CDR=0.5)²; have a knowledgeable informant to provide collateral reports regarding cognitive symptoms; be free of significant underlying medical, neurological, or psychiatric illness (based on standard laboratory tests and a clinical evaluation). All participants and their informants provided informed consent at the time of enrollment in accordance with the guidelines of the Human Research Committee and Institutional Review Board of the Massachusetts General Hospital, Boston, MA.

Overview of methods for participant assessment

At baseline, the study procedures included a medical evaluation (i.e., medical history and physical examination, electrocardiogram, and standard blood tests, including extended chemistry, liver function tests, complete blood count, thyroid stimulating hormone, vitamin B12 and folate), the CDR interview, comprehensive neuropsychological testing, brain magnetic resonance imaging (MRI) and single photon emission computed tomography (SPECT) scans, and blood sample collection for genetic analysis. Thereafter, participants were followed annually with physical and neurologic exams, CDR interviews, and brief neuropsychological testing. Comprehensive neuropsychological testing, MRI, and sometimes SPECT were repeated when participants crossed certain clinical thresholds, including the development of dementia.

Clinical interviews

An expanded, semi-structured interview was administered at baseline, and annually thereafter, to assess the degree of clinical impairment and to generate an overall CDR rating and CDR-SB⁴. Each interview was administered by a skilled clinician (e.g., psychiatrist, neurologist, psychologist, or physician's assistant) and took approximately 90 minutes to complete. The mean inter-rater reliability of the overall CDR ratings was high ($R=0.99$, $p<0.0001$), as was the inter-rater reliability ($R=0.90$) of the 6 CDR subcategories (Memory, Orientation, Judgment and Problem-solving, Community Affairs, Home and Hobbies, Personal Care)⁴.

Application of the SIST-M algorithm

We recently developed the SIST-M⁶ (available at <http://madrc.mgh.harvard.edu/structured-interview-scoring-tool-massachusetts-adrc-sist-m>), which features 60 items derived from the larger set of 131 items in the expanded interview⁴; each SIST-M item can be graded using the same ordinal categories that denote no to mild symptoms on the CDR (e.g., 0, 0.5, 1). A computer algorithm was created in a development cohort ($n=147$ of 165 participants in Cohort 1; 18 participants had data missing on the expanded [131-item] questionnaire) to facilitate validity testing. This algorithm is a complex, hierarchical design that uses a combination of the grade of each item (i.e., 0, 0.5, 1), the frequency with which different grades of items were observed within a CDR domain (or its key sub-domains), and the relative clinical importance, or “weight,” of select items. Thus, the computer algorithm simulates, in effect, what would have happened if the clinician interviews had been conducted using only the 60 items of the SIST-M. Additional details of the development of this algorithm are provided elsewhere⁶. High concordance of algorithm-based CDR-SB with original raters' CDR-SB was confirmed in a replication cohort ($n=200$ of 214 participants in Cohorts 2 and 3; 14 participants had missing data on the expanded questionnaire): ICC (95% confidence interval[CI])=0.89 (0.86,0.91)⁶.

Neuropsychological testing

At baseline, a 22-test neuropsychological battery was administered to all participants, as described previously⁹. Component tests included several measures of episodic memory, such as the California Verbal Learning Test (CVLT) Total Learning and Delayed Retention scores¹⁰ and the Free and Cued Selective Reminding Test¹¹, as well as tests of general cognitive function (Mini-Mental State Exam[MMSE]¹²), executive function (Trail Making Test¹³), working memory and phonemic fluency. The neuropsychological battery was administered in a separate session from the clinical evaluation described above, and was scored in a blinded fashion with respect to CDR ratings or diagnostic information.

Using baseline data from the whole sample, we performed a linear regression for each neuropsychological test score using age, gender, and educational attainment as predictors, and saved the residuals from this regression. Among participants who were cognitively normal at baseline (i.e., global CDR=0.0), we computed the mean and standard deviation of these residuals. Standardized scores – or z-scores – were then calculated for all participants, using the mean and standard deviation of the score distribution of normal participants. Thus, a z-score of -1.0 indicates that the participant's performance was one standard deviation below the expected mean for a cognitively intact person of the same age, gender, and education level.

Sample for analysis

Of the 379 longitudinal study participants, we had available data to generate algorithm-based CDR ratings for 347 persons. In addition, we excluded 5 participants who did not have least one follow-up visit after baseline (n=4) (i.e., follow-up time=0) or had an incomplete baseline MMSE (n=1). Thus, the final sample for analysis=342.

Cognitive status groupings at baseline

In our sample, the distribution of CDR-SB scores at baseline was broad, whether CDR-SB scores were based on clinician- or algorithm-based ratings. Individuals at the more impaired end of the spectrum (i.e., CDR-SB ≥ 2.0) appeared comparable to MCI patients recruited from clinic-based settings, based on similar likelihood of progression to a diagnosis of AD⁴. However, at the mild end of the spectrum (i.e., CDR-SB=0.5-1.5), many participants did not meet psychometric cut-offs commonly used to define amnesic MCI in epidemiologic studies and clinical trials^{14, 15}; such persons were considered to have very mild cognitive impairment^{5, 16}, or vMCI. Other groups have used similar terms – e.g., pre-MCI¹⁶ – to denote the presence of consistent and meaningful changes in subjective and/or informant-reported cognitive abilities without the presence of objective deficits meeting criteria for MCI; in current UDS³ terminology, the cognitive status of participants with vMCI would be categorized as “impaired, not MCI”³. The classification of baseline cognitive status was operationalized using a method detailed elsewhere⁵ and summarized here:

Normal: CDR global rating=0.0.

MCI: CDR global rating=0.5 and

- subjective memory complaint by subject and/or informant: 0.5 rating in Memory;
- evidence of preserved general cognitive function: MMSE ≥ 24 ;
- intact basic and instrumental activities of daily living (IADL): 0.5 ratings in community affairs, home-and-hobbies and personal care.
- objective memory deficit: 1 SD below the age-, gender-, and education-adjusted mean of normal participants' scores on the CVLT delayed free recall trial;

vMCI: those who otherwise met criteria for MCI (i.e., symptoms) but did not meet the cognitive testing requirement.

Thus, two baseline cognitive status classifications could be generated for each participant: one based on *clinicians' original* CDR ratings and the neuropsychological test results, and another generated from *algorithm-based* CDR ratings and the neuropsychological test results. Four participants who otherwise met criteria for MCI but had ratings ≥ 1.0 on community affairs, home-and-hobbies or personal care were excluded from analyses using these cognitive status classifications.

Procedures for diagnosis during follow-up

A consensus diagnosis was assigned to participants who developed significant cognitive and functional impairment at an annual follow-up visit; this consensus method incorporated clinical history, medical records, laboratory evaluation and neuroimaging studies, and is detailed elsewhere⁹. Individuals with dementia were classified as AD or another diagnosis (e.g. fronto-temporal dementia, vascular dementia) according to standard clinical research criteria¹⁷⁻¹⁹.

Statistical Analyses

Using prospective data on the development of AD among participants in both the development and replication cohorts, we evaluated how well the CDR-SB predicted AD diagnosis over follow-up. In survival analyses we examined CDR-SB as a predictor using the clinicians' scores and, separately, using the algorithm-based CDR-SB scores. Using traditional methods of survival analysis, we examined predictive validity in the full sample as well as in the three cognitive status sub-groups of normal, vMCI, and MCI defined according to clinician- or algorithm-based CDR ratings.

First, separate Kaplan-Meier curves were constructed to estimate the AD survival times since baseline evaluation in the normal, vMCI and MCI groups. In each condition (i.e., clinician-based or algorithm-based baseline cognitive status groupings), the log rank test was used to compare survival times among the three groups.

Second, we used the Cox proportional hazards method to allow for variable follow-up lengths, and estimated the hazard ratio (HR) of AD using CDR-SB score as the predictor. The primary focus of the analyses was time from baseline evaluation to diagnosis of AD or censoring; censoring events were death, loss-to-follow-up, or development of non-AD dementia. Two sets of models were created: one using the clinicians' CDR-SB scores and the other using the algorithm-based CDR-SB scores. In addition, Cox models were developed to evaluate prediction of AD within the full sample and separately among 156 participants jointly classified as vMCI by both clinician- and algorithm-based CDR ratings (note that 32 vMCI participants were excluded from these analyses: 18 who met vMCI criteria by clinician but not algorithm-based CDR ratings, and 14 who met criteria by algorithm but not clinician-based ratings; see results). All Cox models were adjusted for age (at baseline evaluation), gender and education; age and education were treated linearly because there was no evidence of a non-monotonic relationship between these variables and the log-hazard estimates of AD conversion. The proportionality assumption was confirmed by plotting the log-negative-log of the estimated survival distribution against the log of follow-up time. As the sets of models using either the clinicians' or the algorithm-based CDR-SB scores were non-nested, model fit could not be compared using likelihood ratio tests; instead, model fit was compared using Akaike information criteria (AIC)^{20, 21}, which is a commonly used measure for comparison of competing models (the model that produces

the minimum AIC is preferred)²¹. All traditional survival analyses were conducted using SAS version 9.1 (SAS Institute, Cary, NC).

Finally, in order to further evaluate predictive ability of clinician- vs. algorithm-based CDR-SB among all participants across the entire follow-up period, we applied a newer statistical approach to characterize the predictive accuracy of clinician- vs. algorithm-based CDR-SB: Receiver Operating Characteristic (ROC) curve methods for time-dependent outcomes (i.e., time-to-AD as a censored survival time)⁸. This time-dependent ROC method utilizes concepts of incident sensitivity (IS) and dynamic specificity (DS). IS at time t measures the expected fraction of participants with a “marker” value $> M$ among the sub-population of individuals who progress to AD at time t . DS at time t measures the expected fraction of subjects with a “marker” value $> M$ among the sub-population of individuals who have not progressed to AD by time t . The marker is a linear combination of estimated regression parameters and covariates (age, gender, years of education and either the clinician CDR-SB or the algorithm-based one) derived from the Cox model. In this context, M is a marker threshold. Thus, IS and DS are defined by dichotomizing the risk set at time t into those who progress and those who do not. The corresponding ROC curve at time t depicts IS versus DS at time t over a set of thresholds. A scalar summary for the ROC curve at time t is the area-under-the-ROC curve (AUC) at time t . Moreover, a global scalar that summarizes the sequence of time-dependent AUCs over the follow-up time is a weighted average of the AUCs, and has interpretation as the concordance probability⁸. This concordance probability, C , related to Kendall's tau measure of bivariate correlation, measures the predictive accuracy of a marker for an event that occurs at random times and may be right-censored. C reflects the probability that, for a pair of subjects, the person who progressed earlier to AD has a larger value of the marker: $C = P[M_j > M_k | T_j < T_k]$, where M_i and T_i denote the values of the marker and the survival time for the i^{th} individual, and where it is assumed that a higher marker value is predictive of poor prognosis. Therefore, C will reflect the probability that the predictions using this marker for a random pair of subjects are concordant with their outcomes – i.e., how well the marker predicts earlier progression to AD. In our study, M_1 and M_2 are marker values obtained from the Cox models using clinician CDR-SB and algorithm-based CDR-SB, respectively; C_1 and C_2 are the respective concordance summaries obtained using M_1 and M_2 . The 95% confidence intervals for the difference between the two concordance summaries were derived using the percentile bootstrap method²² and accounted for their correlation due to being calculated on the same subjects. Thus, using time-dependent ROC curve methodology, we were able to make a direct statistical comparison between original and algorithmic CDR-SB on predictive accuracy for AD over the entire range of follow-up time. These time-dependent ROC analyses were conducted using R (R Foundation for Statistical Computing, Vienna, Austria).

Results

Demographic and clinical characteristics of sample

Table 1 illustrates participants' demographic and clinical characteristics, by diagnostic grouping. Members of the normal group were several years younger, on average, than those in the MCI group and slightly younger than persons with vMCI. There were no group differences in gender or education. Not surprisingly, cognitively normal participants had better performance on all neuropsychological measures than the MCI group; however, their performance was comparable to that of participants with vMCI on most measures.

Classification of baseline cognitive status by clinician- and algorithm-based CDR ratings

Agreement of cognitive status classifications using clinicians' vs. algorithm-based CDR scores was high; the weighted kappa (κ) for agreement=0.84 (95% CI=0.78,0.91). Further,

when re-classifying to MCI the four participants who otherwise met criteria for MCI but who had either clinician- or algorithm-based CDR ratings ≥ 1.0 in community affairs, home and hobbies or personal care (i.e., “MCI-plus”), agreement was again high (weighted $\kappa=0.85$; 95% CI=0.79,0.91). (Data not shown in table).

Result from traditional survival analysis

Results from the adjusted Cox proportional hazards models demonstrated similar findings in models utilizing clinician vs. algorithm-based ratings (Table 2). The clinician-based model fit the data slightly better than the algorithm-based model (i.e., lower AIC); however, AD hazard ratios were the same: for a 1-point increment in baseline CDR-SB, there was a three-fold increase in the hazard. Within the vMCI group, the estimated two-fold increases in the hazard associated with each 1-point increase in CDR-SB were similar; model fit was identical. Finally, as would be expected given the similar CDR ratings, the Kaplan-Meier curves of the three cognitive status groups looked identical, whether the groups were based on clinician or algorithm-based CDR scores (Fig. 1).

Results from time-dependent ROC analysis

The AUCs over follow-up were informative (Fig. 2). Concordance probabilities (global values for the weighted average of the AUCs over the follow-up period) were similar using original (concordance probability=0.78; 95% CI=0.74,0.82) and algorithmic (concordance probability=0.76; 95% CI=0.72,0.80) CDR-SBs. The estimated difference between concordance probabilities=0.02 (95% CI=-0.01,0.05). Thus, the CI for the difference contained zero – indicating that the data did not provide evidence of a statistical difference in concordance probabilities (which measure predictive accuracy of AD development over the entire range of follow-up). Furthermore, when restricting analysis to the replication cohort, concordance probabilities (95% CI) were 0.79 (0.73,0.85) using original CDR-SB and 0.78 (0.72,0.84) using algorithmic CDR-SB; the estimate for the difference between concordance probabilities contained zero: 0.015 (95% CI=-0.03,0.05). Interestingly, on visual inspection of Fig. 2, the original CDR-SB appeared to have a higher concordance probability for short-term predictions; however, for longer-term predictions, original clinician-based and algorithm-based CDR-SBs appeared comparable. However, we separately calculated the difference in the concordance probabilities for the first 2 and 3 years of follow-up, and there was no statistical evidence of a difference in predictive accuracy: estimated differences (comparing original vs. algorithmic CDR-SBs) were 0.017 (95% CI=-0.03,0.06) in the first 2 years and 0.013 (95% CI=-0.03,0.05) in the first 3 years. Similarly, when restricting analysis to the replication cohort, estimated differences were 0.011 (95% CI=-0.05,0.06) in the first 2 years and -0.006 (95% CI=-0.06,0.05) in the first 3 years. Finally, among the 156 participants jointly classified as vMCI, concordance probabilities (95% CI) were 0.72 (0.66,0.79) using original CDR-SB and 0.71 (0.66,0.78) using algorithmic CDR-SB; the estimated difference=0.005 (95% CI=-0.04,0.05) – demonstrating that, similar to findings from the traditional survival analyses, there was no evidence of a statistical difference in predictive accuracy between original interview-based and SIST-M algorithm-based CDR-SBs within the vMCI sub-group. (Data not shown in tables.)

Discussion

The SIST-M was recently developed as a time-efficient structured interview that can be used to generate CDR scores that are reliable and discriminate along the spectrum of mild cognitive deficits⁶; we previously identified high reliability and cross-sectional validity for this instrument. In the current report, we demonstrate that the instrument has high *predictive*

validity – identical to that of the lengthier expanded interview⁵ – even among those with the mildest cognitive symptoms.

Strengths of this study include development of the SIST-M within a well-characterized, community-based sample of older adults with a broad range of mild cognitive symptoms, established reliability of the SIST-M, and prospective design with lengthy follow-up (mean=7.4 years, SD=4.0, range=1.0-14.2). Furthermore, we utilized time-dependent ROC methodology to verify that high predictive accuracy of baseline CDR-SB scores was maintained over the length of the follow-up period, using either the shorter or longer interview format. A powerful advantage of this newer statistical method of time-dependent ROC analysis is that it facilitates a direct statistical comparison of the clinician-based and algorithm-based CDR-SB scores with respect to predictive accuracy over the entire range of follow-up time. In addition, the use of this method is attractive because of the familiarity and interpretability of the ROC concept for most clinical research audiences.

Limitations must also be recognized. First, prediction using the SIST-M was based on an algorithmic computer simulation of clinician judgment, not actual in-person interviews. Nevertheless, high concordance of clinician vs. algorithm-based CDR ratings and CDR-SB⁶ suggests little loss of information. Second, predictive validity was evaluated in a cohort of well-educated, mostly Caucasian elders who were aged 57-87 years at baseline; thus, generalizability of the SIST-M to less-educated, racial/ethnic minority or oldest-old populations is unclear. Nevertheless, the characteristics of our cohort are consistent with those observed nationally in many other ADC/ADRCs, and it is likely that the SIST-M would perform equally well at other sites with respect to prediction of clinical progression. Third, there is likely some inflation of the concordance probabilities in the time-dependent ROC curve analyses, due to our deriving the marker using the same data on which we assessed it. However, this is equally true for both the clinician- and algorithm-based CDR-SB, and so the conclusions regarding their comparison are valid. Finally, dementia diagnoses were not all autopsy-confirmed. Nevertheless, criteria-based diagnoses of probable AD, when made by skilled physicians, are known to be approximately 90% accurate, compared to pathologic diagnoses¹⁷.

In summary, the validity of the SIST-M in generating CDR-SB scores is supported by its high predictive validity for the development of AD. Furthermore, these results support the value of the CDR-SB in grading subtle deficits among cognitively impaired persons, even in the setting of vMCI. Thus, the SIST-M can have particular value in efficiently generating CDR scores where research interest is targeted on recruiting and characterizing samples of very mildly impaired individuals for prevention, early intervention or disease-modifying trials, or where clinical attention is focused on assessment of primary care patients and community-based individuals who might benefit from early intervention of novel therapeutic agents.

Acknowledgments

This study was supported by the Harvard NeuroDiscovery Center and National Institute on Aging (P01-AG004953, P50-AG005134).

The authors thank Jeanette Gunther and Kelly A. Hennigan for assistance with participant recruitment, retention and visit coordination; Laura E. Carroll, Sheela Chandrashekar and Michelle Schamberg for assistance with data collection, entry and quality checking; and Mary Hyde for assistance with data management. We express special appreciation to all of our study participants.

Funding: This study was supported by the Harvard NeuroDiscovery Center and the National Institute on Aging of the National Institutes of Health (P01-AG004953, P50-AG005134). This study is not industry-sponsored. The copyrights in the two instruments referred to in this manuscript, the SIST-M and the SIST-M-IR, belong to The

General Hospital Corporation d/b/a Massachusetts General Hospital. Additional information and downloadable .pdfs are freely available at <http://madrc.mgh.harvard.edu/structured-interview-scoring-tool-massachusetts-adrc-sist-m>.

Dr. Okereke receives funding from the NIH and the Alzheimer's Association. She serves on the Board of Directors of the Alzheimer's Association, Massachusetts/New Hampshire Chapter.

Dr. Hyman receives funding from the NIH, the Alzheimer's Association, and Fidelity Biosciences. He reports consulting with pharmaceutical and biotechnology companies: EMD Serrano, Janssen, Takeda, BMS, Neurophage, Pfizer, Quanterix, foldrx, Elan, and Link. He holds no stock options. He reports no conflicts pertaining to this manuscript.

Dr. Albert reports serving as a consultant for Genentech and Eli Lilly, and receiving grants to her institution from GE Healthcare.

Dr. Blacker receives funding from the NIH, the Alzheimer's Association, and Fidelity Foundation. She serves on the Board of Directors of the Alzheimer's Association, Massachusetts/New Hampshire Chapter.

References

1. Hughes CP, Berg L, Danzinger WL, et al. A new clinical scale for the staging of dementia. *Br J Psychiatry*. 1982; 140:566–572. [PubMed: 7104545]
2. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*. 1993; 43:2412–2414. [PubMed: 8232972]
3. The National Alzheimer's Coordinating Center (NACC). <http://www.alz.washington.edu>
4. Daly E, Zaitchik D, Copeland M, Schmahmann J, Gunther J, Albert M. Predicting conversion to Alzheimer disease using standardized clinical information. *Arch Neurol*. 2000; 57:675–680. [PubMed: 10815133]
5. Dickerson BC, Sperling RA, Hyman BT, Albert MS, Blacker D. Clinical prediction of Alzheimer disease dementia across the spectrum of mild cognitive impairment. *Arch Gen Psychiatry*. 2007; 64(12):1443–1450. [PubMed: 18056553]
6. Okereke OI, Copeland M, Hyman BT, Wanggaard T, Albert MS, Blacker D. The SIST-M: Development, reliability and cross-sectional validation of a brief structured Clinical Dementia Rating interview. *Arch Neurol*. 2011 in press.
7. Schafer KA, Tractenberg RE, Sano M, et al. Reliability of monitoring the clinical dementia rating in multicenter clinical trials. *Alzheimer Dis Assoc Disord*. 2004; 18(4):219–222. [PubMed: 15592134]
8. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005; 61:92–105. [PubMed: 15737082]
9. Albert MS, Moss MB, Tanzi R, Jones K. Preclinical prediction of AD using neuropsychological tests. *J Int Neuropsychol Soc*. 2001; 7:631–639. [PubMed: 11459114]
10. Delis, DC.; Kramer, JH.; Kaplan, E.; Ober, BA. California Verbal Learning Test, Research Edition, Manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich; 1987.
11. Grober E, Buschke H. Genuine memory deficits in dementia. *Dev Neuropsychol*. 1987; 3(1):13–36.
12. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975; 12:189–198. [PubMed: 1202204]
13. Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. *Percept Mot Skills*. 1958; 8:271–276.
14. Davis HS, Rockwood K. Conceptualization of mild cognitive impairment: a review. *Int J Geriatr Psychiatry*. 2004; 19:313–319. [PubMed: 15065223]
15. Petersen RC, Thomas RG, Grundman M, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. *N Engl J Med*. 2005; 352:2379–2388. [PubMed: 15829527]
16. Storandt M, Grant EA, Miller JP, Morris JC. Longitudinal course and neuropathologic outcomes in original vs revised MCI and in pre-MCI. *Neurology*. 2006; 67:467–473. [PubMed: 16894109]

17. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlam M. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology*. 1984; 34:939–944. [PubMed: 6610841]
18. Roman GC, Tatemichi TK, Erkinjuntti T, et al. Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop. *Neurology*. 1993; 43:250–260. [PubMed: 8094895]
19. Neary D, Snowden JS, Gustafson L, et al. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. *Neurology*. 1998; 51(6):1546–1554. [PubMed: 9855500]
20. Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, BN.; Csaki, F., editors. *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado; 1973. p. 267-281.
21. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19(6):716–723.
22. Davison, AC.; Hinkley, DV. *Bootstrap Methods and Their Application* (Cambridge Series in Statistical and Probabilistic Mathematics, No 1). Cambridge, UK: Cambridge University Press; 1997.

Abbreviations

SIST-M	Structured Interview and Scoring Tool - Massachusetts ADRC
SIST-M-IR	SIST-M-Informant Report

Figure 1. Kaplan-Meier Curves of Clinician-based and Algorithm-based Baseline Cognitive Status Groupings

vMCI = very mild cognitive impairment; MCI = mild cognitive impairment

Figure 2. Predictive Accuracy of CDR-SB in Alzheimer disease: AUC by Follow-up Time
CDR-SB = Clinical Dementia Rating-Sum-of-Boxes. AUC = area-under-the-curve.

Table 1
Characteristics at study entry, by baseline cognitive status groups*

Demographic, Clinical and Genetic	Normal	vMCI[†]	MCI[†]
Number of participants (n)	112	174	56
Mean (SD) length of follow-up (years)	7.0 (3.8)	7.8 (4.1)	7.0 (3.7)
Mean (SD) age	71.4 (4.8)	72.6 (5.7)	74.9 (6.1)
Mean (SD) years of education	15.6 (2.9)	15.5 (2.9)	15.4 (2.9)
Female (%)	60.7	54.6	57.1
Non-white race (%)	7.1	10.3	7.1
Marital status (%)			
- married	53.6	65.5	62.5
- widowed	20.5	23.0	17.9
- separated/divorced	12.5	8.6	14.3
- never married	13.4	2.9	5.4
History of hypertension (%)	39.3	40.8	35.7
History of diabetes (%)	2.7	4.6	7.1
Current or past smoking (%)	60.8	55.7	69.1
APOE ε4 status (%)	26.1	29.2	33.9
Neuropsychological Testing, Mean (SD)	Normal	vMCI[†]	MCI[†]
MMSE	29.5 (0.7)	29.3 (1.0)	28.6 (1.5)
CVLT Total Score	52.0 (10.5)	50.7 (9.5)	34.7 (9.6)
CVLT Percent Retention	84.4 (17.8)	88.5 (14.3)	53.9 (25.4)
Time to Complete Trails B (seconds)	92.8 (35.2)	103.7 (57.1)	127.4 (85.6)
Phonemic Fluency (total of F, A, S)	46.5 (12.6)	44.8 (13.1)	39.5 (13.1)
Digit Span Backwards	5.6 (1.4)	5.2 (1.5)	5.1 (1.4)

* For the purposes of Table 1, baseline cognitive status groups were determined using clinicians' Clinical Dementia Rating scores, rather than the algorithm-based ones. SD = standard deviation.

[†] vMCI = very mild cognitive impairment; MCI = mild cognitive impairment (see text under Methods for operationalized definitions of these terms).

Table 2
Adjusted* Hazard Ratio (95% CI) of Alzheimer disease, by Clinicians' and Computer
Algorithm-based CDR Sum-of-Boxes

	Clinicians' Ratings	Algorithm-based Ratings
Full sample (n=342)	3.1 (2.5-3.9)	2.8 (2.2-3.5)
AIC[†]	848	859
vMCI sub-group (n=156)	2.2 (1.6-3.2)	2.1 (1.5-3.0)
AIC[†]	446	446

* Models were adjusted for age, gender, and education. Hazard ratios represent relative hazard of each 1-point increment in baseline CDR-SB (Clinical Dementia Rating-Sum-of-Boxes).

[†] AIC = Akaike Information Criteria; smaller number indicates better model fit.