Published in final edited form as:

Int Psychogeriatr. 2018 October ; 30(10): 1435–1445. doi:10.1017/S1041610217002666.

Does a cognitive stress test predict progression from mild cognitive impairment to dementia equally well in clinical versus population-based settings?

Joanne C. Beer¹, Beth E. Snitz², Chung-Chou H. Chang^{3,1}, David A. Loewenstein^{4,5}, and Mary Ganguli^{6,7,2}

¹Department of Biostatistics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA

²Department of Neurology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

³Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

⁴Departments of Psychiatry, Behavioral Sciences, and Neurology, University of Miami Miller School of Medicine, Miami, FL, USA

⁵Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center, Miami Beach, FL, USA

⁶Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

⁷Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA

Abstract

Background—Evidence suggests that semantic interference may be a sensitive indicator of early dementia. We examined the utility of the Semantic Interference Test (SIT), a cognitive stress memory paradigm which taps proactive and retroactive semantic interference, for predicting progression from mild cognitive impairment (MCI) to dementia in both a clinical and a population-based sample.

Methods—Participants with MCI in the clinical (n=184) and population-based (n=435) samples were followed for up to four years. We employed receiver operating characteristic (ROC) methods to establish optimal thresholds for four different SIT indices. Threshold performance was compared in the two samples using logistic and Cox proportional hazard regression models.

Results—Within 4 years, 42 (22.8%) MCI individuals in the clinical sample and 45 (10.3%) individuals in the population-based sample progressed to dementia. Overall classification accuracy of SIT thresholds ranged from 61.4 to 84.8%. Different subtests of the SIT had slightly different performance characteristics in the two samples. However, regression models showed that

Corresponding author: Joanne C. Beer, Department of Biostatistics, University of Pittsburgh, 7135 Parran Hall, 130 De Soto Street, Pittsburgh, PA 15261, Phone: +1 503 780 0305, joannebeer@gmail.com.

Conflict of Interest
None.

thresholds established in the clinical sample performed similarly in the population sample before and after adjusting for demographics and other baseline neuropsychological test scores.

Conclusions—Despite differences in demographic composition and progression rates, baseline SIT scores predicted progression from MCI to dementia similarly in both samples. Thresholds that best predicted progression were slightly below thresholds established for distinguishing between amnesic MCI and cognitively normal subjects in clinical practice. This confirms the utility of the SIT in both clinical and population-based samples and establishes thresholds most predictive of progression of individuals with MCI.

Keywords

Alzheimer's disease; mild cognitive impairment; dementia; memory; neuropsychological tests

Introduction

The paradigm of semantic interference has been described as a “cognitive stress test” that identifies subtle deficits characteristic of very early Alzheimer's disease (AD) (Loewenstein *et al.*, 2004; Loewenstein *et al.*, 2017). The Semantic Interference Test (SIT) is a modification of the three-trial Fuld Object Memory Evaluation (Fuld-OME) (Fuld, 1981), which consists of learning trials of 10 actual common objects. In the SIT modification, a second group of objects semantically related to the first group (e.g. matches versus lighter) is introduced, allowing the assessment of proactive and retroactive semantic interference. Proactive semantic interference (PSI) occurs when old learning interferes with new learning of a semantically similar list of different targets. Conversely, retroactive interference (RSI) refers to difficulties recalling the original targets as a function of learning the semantically related new targets.

The Fuld-OME and SIT have been shown to have little cultural or language bias and to have no or only weak associations with education level (Loewenstein, 1995; Snitz *et al.*, 2010). However, the SIT might function somewhat differently in different study settings for the following reasons: First, any given test with established sensitivity and specificity for a particular outcome can have different predictive values in different study populations depending upon the underlying distribution of the outcome in those populations. For example, positive predictive value will be high in a population with high prevalence of a disorder such as Alzheimer's disease (AD), especially one that has been enriched for that disease such as a specialized research clinic. Secondly, since semantic interference is associated with the particular type of cognitive decline seen in AD, the efficacy of SIT observed in AD research clinical samples may be greater than in more heterogeneous population-based samples.

Clinical thresholds for identification of AD have been established for the SIT which have yielded high sensitivity and specificity in clearly defined clinical samples with a high probability of underlying AD (Loewenstein *et al.*, 2007; Loewenstein *et al.*, 2004). Moreover, these thresholds have been strongly related to total and regional brain amyloid load in community-dwelling elderly adults who have otherwise normal scores on traditional neuropsychological tests (Loewenstein *et al.*, 2015). However, it cannot be assumed that

those thresholds that best distinguish between diagnostic groups are also the most appropriate to predict the likelihood of progression of mild cognitive impairment (MCI) to dementia over time. Also unknown is whether optimal thresholds for predicting progression might differ in specialty clinical versus community samples, between which the underlying prevalence of AD and other dementias would vary considerably.

In this investigation, we examined the data of a large population-based cohort as well as a sample of participants in an Alzheimer's Disease Research Center (ADRC) registry. We empirically derived thresholds that optimally distinguished between progressors and non-progressors from MCI to dementia, independently for each cohort, and for each of several specific SIT component measures (assessing initial recall, recall vulnerable to PSI, recall vulnerable to RSI, and the combined interference score). We explored how optimal thresholds might differ in the two samples and determined the extent to which these thresholds were useful for tracking progression.

Our hypotheses were that (1) Optimal thresholds predicting MCI progression to dementia would be lower (i.e. worse or more impaired) than optimal clinical thresholds previously established for distinguishing MCI from normal cognition. (2) Optimal thresholds would more clearly differentiate MCI progressors to dementia from non-progressors in the ADRC registry than in a population-based cohort in which the underlying prevalence of AD and other dementias is likely to be lower.

Methods

Participants

Clinical sample—The clinical sample comprised 184 individuals aged 65 and older who were part of the Florida Alzheimer's Disease Research Center (ADRC) program beginning in 2005 and followed annually for up to four years. Some participants came to the ADRC as patients, others as spouses or normal volunteers, while some were referred from memory screening programs in the community. All of these individuals were diagnosed with MCI at baseline by an experienced clinician who completed a Clinical Dementia Rating (CDR) (Hughes *et al.*, 1982). Based on this interview and cognitive and functional history from the identified informant, the clinician diagnosed MCI if (a) there was a complaint about memory or other cognitive functions; (b) the clinician made a determination that based on informant report, there had been a decline in cognitive function; (c) the individual was able to independently manage ADLs and IADLs; and (d) there was no evidence of dementia by DSM-IV criteria after a diagnostic case conference. We excluded participants who were missing any SIT variables at baseline. Approximately 20% of participants evaluated at baseline were lost to follow up.

Population-based sample—The population cohort was drawn from the Monongahela-Youghiogheny Healthy Aging Team (MYHAT) study, a longitudinal study based in a region of southwestern Pennsylvania (Ganguli *et al.*, 2009). Starting in 2006, an age-stratified sample was randomly recruited from publicly available voter registration records. Eligibility criteria included age 65 or over, living within the designated area, and not living in a long-term care institution. Participants were excluded if they were too ill to participate, too

impaired in vision or hearing, decisionally incapacitated, or had substantial cognitive impairment defined as an age-education-corrected Mini-Mental State Exam (MMSE) score of less than 21 out of 30 (Folstein *et al.*, 1975; Mungas *et al.*, 1996). A total of 1982 participants met these criteria and underwent baseline and annual follow-up assessments for a maximum of seven total assessments. For the present study, we used CDR, which was not assigned on the basis of SIT or other neuropsychological test scores, to designate participants with MCI (CDR=0.5) and dementia (CDR $\leq$ 1) (Ganguli *et al.*, 2010a). Research interviewers underwent systematic training and certification in the use of the CDR in the population setting (Ganguli *et al.*, 2010b). We restricted present analyses to those with CDR of 0.5 at baseline (n=546) and full SIT data at baseline (n=529). Approximately 18% of participants were lost to follow up, resulting in a sample size of 435. We truncated follow-up time at a maximum of four years to match that of the clinical sample.

Memory Assessment: Fuld-OME and SIT

Participants completed the three-trial version of the Fuld-OME, in which they attempt to identify 10 common household objects placed into a cloth bag (Bag A), first by touch and then by sight. Interposed with distracter tasks, they try to recall, with selective reminders, the 10 items on three subsequent trials. Total Fuld Recall score is the sum of items correctly recalled in the three trials (maximum 30 points). The SIT introduces a second bag (Bag B) with 10 different but semantically related household objects (e.g. substituting bracelet for ring). After the three trials of Fuld-OME, participants identify the items in Bag B in the same manner as for Bag A. They are then asked to recall these items (Bag B Recall, susceptible to proactive interference, maximum 10 points). Then they are asked to recall the items in Bag A again (Bag A Recall, susceptible to retroactive interference, maximum 10 points). In our analyses we also considered the sum of Bag B Recall and Bag A Recall (Combined Interference Score).

Statistical Methods

Sample characteristics—We generated descriptive statistical summaries to characterize the clinical and population-based samples and performed significance testing to assess differences between the two samples as well as between non-progressors and progressors to dementia within both samples. Statistical significance was set at $p < 0.05$.

Assessing comparability of MCI and dementia diagnoses—Since different methodologies were used to assign MCI and dementia diagnoses in the two studies, we compared mean CDR Sum of Boxes and neuropsychological test scores for progressors to dementia in the two samples, using Wilcoxon rank sum test and Welch's t-test, respectively, both at baseline and at the time of dementia diagnosis. In addition to Fuld-OME and SIT, neuropsychological tests administered in both study settings included the Mini-Mental State Examination (Folstein *et al.*, 1975), WMS-R Logical Memory (immediate and delayed recall) (Wechsler, 1987), WMS-R Visual Reproduction (immediate and delayed recall) (Wechsler, 1987), and Trail Making Test B (Reitan, 1958).

Optimal thresholds—Treating progression to dementia (observed within a maximum of four years of follow-up) as a binary outcome, we used receiver operating characteristic

(ROC) analysis methods to determine optimal thresholds independently in the clinical and population samples for the four test scores of interest. We used two different criteria for establishing the optimal thresholds: (1) Youden's J statistic, which chooses the threshold that maximizes the sum of sensitivity and specificity; and (2) the threshold of maximum specificity that achieved a minimum sensitivity of 70%. We chose this minimum specificity criterion because the Youden's J statistic weights sensitivity and specificity equally, which could produce optimal thresholds that behave quite differently (e.g. a threshold with 20% sensitivity and 70% specificity would have the same J statistic as one with 70% sensitivity and 20% specificity). We wanted to compare optimal thresholds in the two samples in cases where we knew that at least 70% of progressors were identified. A sensitivity of 70% was chosen as the threshold because the results using the Youden's J statistic criterion yielded highest sensitivities close to 70%. We calculated the areas under the ROC curve (AUC), and for each optimal threshold we determined sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), classification accuracy, and Youden's J statistic. We repeated the analyses on 10,000 empirical bootstrap samples (stratified so as to maintain the same number of progressors and non-progressors) and used the 2.5th and 97.5th quantiles of the distributions of the statistics to form 95% confidence intervals for these quantities. Non-overlapping confidence intervals were considered to indicate a statistically significant difference.

Adjusting for covariates—To further assess the behavior of the optimal thresholds in the two samples, we created binary variables using the set of optimal thresholds established for the clinical sample under the Youden's J criterion. For each SIT measure, we fitted an unadjusted logistic regression model with the binary test score variable, a sample indicator variable, and the interaction of the two as predictors. This allowed us to calculate odds ratios for both samples and to test whether the thresholds predicted progression differently across samples. To account for differences in baseline characteristics between the two samples, we then adjusted these models for age, sex, years of education, race/ethnicity, baseline CDR Sum of Boxes, and a baseline test composite combining scores from the other neuropsychological tests mentioned previously. To create the composite, test scores were each standardized using the baseline means and standard deviations from the pooled sample data, and then averaged together (negative Trail Making Test B scores were used since higher scores indicate worse performance). For the approximately 12% of participants who were missing one or more test scores, composites were formed using available tests. We also fitted unadjusted and adjusted Cox proportional hazards models to account for differences in follow up time and censoring. We employed log-log survival plots to check the proportional hazards assumption for the Cox models.

All statistical analyses were performed using R version 3.3.3 and the packages pROC (Robin *et al.*, 2011) and survival (Therneau, 2015).

Results

Sample characteristics

Comparing clinical and population samples—Sample characteristics are summarized in Table 1. Within four years of follow up, 42 (22.8%) individuals in the clinical sample and 45 (10.3%) individuals in the population sample progressed to dementia (Fisher's exact test $p < 0.001$). The mean (SD) times to dementia were 1.90 (0.82) years for the clinical sample and 2.51 (1.04) years for the population sample (Welch's $t(82.9) = 3.04$, $p = 0.003$); incidence rates for dementia were 105.0 and 33.1 cases per 1000 person-years, respectively. The two samples differed significantly in all demographic variables considered. The clinical sample was younger, had a greater proportion of men, had more years of education, and was more ethnically diverse with a greater proportion of Hispanic participants than the population sample. Median follow-up time was shorter in the clinical sample. While baseline CDR Sum of Boxes indicated slightly greater impairment in the clinical sample than the population sample, baseline test composite scores showed the opposite trend, although this difference was not significant.

Comparing progressors and non-progressors to dementia within samples—The clinical sample evidenced no significant differences in demographics or follow-up time between non-progressors and progressors. In the population-based sample, progressors were older than non-progressors, with no other significant differences in demographics or follow-up time. For both samples, Fuld, SIT, and other neuropsychological test score means at baseline all indicated significantly worse performance in the progressors than non-progressors.

Assessing comparability of MCI and dementia diagnoses—The clinical sample progressors to dementia had mean (SD) baseline CDR Sum of Boxes of 1.9 (1.0), slightly higher than the population sample progressors to dementia mean (SD) of 1.5 (0.9) ($p = 0.027$). However, at the time of dementia diagnosis the clinical sample progressors had a lower mean (SD) CDR Sum of Boxes, which was 4.5 (2.2) versus 5.2 (1.6) for the population sample progressors ($p = 0.002$). Neuropsychological test composite scores were higher for the clinical sample progressors at baseline, with mean (SD) -0.37 (0.60) versus -0.72 (0.66) for the population sample ($p = 0.011$), and at the time of dementia diagnosis, with mean (SD) -1.24 (0.81) versus -1.77 (0.85) for the population sample ($p = 0.004$). Within-subject change in CDR Sum of Boxes was smaller for clinical sample progressors (mean (SD) 2.6 (2.4)) than for the population sample progressors (mean (SD) 3.7 (1.8), $p = 0.013$). Within-subject change in test composite scores were not significantly different, with mean (SD) of -0.88 (0.98) for the clinical sample and -1.05 (1.00) for the population sample ($p = 0.425$). Results for individual neuropsychological tests are presented in Table S2.

Optimal thresholds

In both the clinical and population samples, AUC was highest for Total Fuld Recall, which yielded an AUC of approximately 0.75 for both (Table 2, Figure 1). According to the 95% bootstrap confidence intervals the AUC did not differ significantly between the samples for

any of the tests. Optimal threshold results for the Youden's J statistic criterion are summarized in Table 2. Bag A Recall threshold yielded a significantly higher PPV in the clinical sample; the other quantities were not significantly different. Optimal threshold results for the sensitivity 70% criterion are summarized in Table S1. Again we found a significantly higher PPV for Bag A Recall in the clinical sample and the other quantities were not significantly different. However, the point estimates for specificity, PPV, and accuracy were consistently higher in the clinical sample, and NPV was consistently higher in the population-based sample.

Adjusting for covariates—Logistic regression (Table 3) and Cox proportional hazards models (Table 4, Figure S1) demonstrated no significant interaction between the binary threshold variables and the clinical sample indicator in any of the unadjusted or adjusted models. The clinical sample indicator variable had an odds or hazard ratio significantly greater than one in most unadjusted and all of the adjusted models, consistent with a higher overall risk for progression in the clinical sample.

Discussion

As dementia, including AD, becomes an increasingly urgent public health issue, it is important to identify where preventative efforts will have the most impact. Accordingly, many biomarkers and cognitive assessments have been studied for the purpose of identifying individuals with MCI who are most at risk of progression. The Semantic Interference Test (SIT), a cognitive stress test and an extension of the three-trial Fuld Object Memory Evaluation, has been developed as a sensitive measure that, unlike other verbal memory measures, shows minimal cultural and educational bias in both clinical and population-based studies (Loewenstein *et al.*, 2004; Snitz *et al.*, 2010). Previous research has demonstrated that thresholds of 19.5 or 20.5 (i.e. scores of less than 20 or 21 indicating likely impairment) for Total Fuld Recall, 4.5 for Bag B Recall (proactive semantic interference trial), and 8.5 for Combined Interference Score (total interference trial) have been optimal for differentiating clinical cases with amnesic MCI (aMCI) from cognitively normal older adults (Loewenstein *et al.*, 2004).

In the present study, we extended our work by comparing the utility of SIT for predicting progression of MCI to dementia within four years of follow up in two different samples: a well-characterized clinical sample and a well-defined population-based sample. We found that, despite differences in distributions of age, years of education, gender, race and ethnicity, and rate of progression to dementia, the SIT performed similarly for predicting progression in both samples. Specifically, the SIT subtests had comparable AUC across samples, the optimal thresholds were in many cases the same across samples (or differed by at most two points, depending on the test and optimality criterion), and for one particular set of thresholds there were no significant differences in the dichotomized test score effect across samples in both unadjusted and adjusted regression models.

Our original hypothesis was that optimal thresholds for identifying progressors in both the clinical and population-based samples would be more impaired (i.e. lower) than previously observed optimal thresholds for differentiating persons clinically diagnosed with aMCI from

cognitively normal older adults. We found that optimal thresholds determined in the present study under the same criterion (Youden's J, which yielded Total Fuld Recall thresholds of 17.5 and 18.5, Bag B Recall thresholds of 3.5 and 5.5, and Combined Interference Score threshold of 7.5 for both samples) are in fact slightly lower than the aforementioned clinical thresholds, with the exception of the threshold for Bag B Recall in the clinical sample. This suggests that those with MCI who are most at risk for progression may be more susceptible to semantic interference than MCI non-progressors, and further establishes the utility of SIT indices for clinical purposes. Unsurprisingly, the sensitivity, specificity, overall accuracy, and AUC tended to be lower for predicting MCI progression in the clinical sample than previously reported values for discrimination between clinically diagnosed aMCI and cognitively normal older adults in a clinical sample (Loewenstein *et al.*, 2004). However, overall accuracy for the population-based sample was higher than previously reported in the case of Bag B Recall (81.3% versus 84.8% in the present study) and Bag A Recall (76.2% vs 77.0% in the present study), possibly due to the high specificity of these thresholds resulting from the lower incidence of dementia in the population-based sample.

The finding that the optimal MCI progression threshold for Bag B Recall score (vulnerable to PSI) was one point higher in the clinical sample than the threshold found to best differentiate aMCI from cognitively normal adults contrasts with a previous study conducted by Loewenstein *et al.* (2007): In a sample of 76 predominantly clinically diagnosed MCI subjects, 35.5% progressed to dementia over an average 21 month period of follow up. Bag B Recall was the neuropsychological measure most predictive of progression, and a threshold of 4.5 on Bag B Recall of the SIT best distinguished those who progressed to dementia. The resulting (leave-one-out cross-validated) sensitivity was 70.4% while specificity was 73.5% (AUC = 0.775), which were higher than the present study's sensitivity of 66.7% and specificity of 59.9% (AUC = 0.666) for the clinical sample. It should be noted that over 85% of individuals in the Loewenstein *et al.* (2007) study were diagnosed with an underlying neurological condition (Alzheimer's disease, cerebral infarctions, diffuse Lewy body disease), and this group had lower mean Bag B Recall scores at baseline than the present study's clinical sample.

We also hypothesized that optimal SIT thresholds would discriminate between MCI progressors and non-progressors better in the clinical ADRC sample due to its likely greater proportion of AD patients than the population cohort. We did find that more than double the proportion of MCI participants progressed to dementia in the clinical versus the population sample over a four year period. However, there were no consistent differences in the discrimination performance of the SIT subtests as measured by AUC and overall classification accuracy (for thresholds found using Youden's J criterion). Moreover, the effect of SIT thresholds were similar in both cohorts, despite the differences between the two groups, and even after covariate adjustment for differences in initial demographic characteristics.

Yet, when sensitivity was fixed at a minimum of 70%, PPV for all optimal thresholds established in the clinical sample was greater than for the corresponding thresholds in the population-based cohort (which tended to have higher NPV values). Since predictive value is a function of underlying prevalence, the higher PPV and lower NPV in the clinical group

likely reflected a greater underlying prevalence of AD and other dementia-causing diseases. Specificity and accuracy were also consistently higher for the clinical sample under this minimum sensitivity criterion. Although it was revealing of differences between the two samples in the present study, we do not recommend setting sensitivity at a minimum of 70% to identify those at risk of MCI progression in the population-based sample, given the resulting low PPV.

Our aim of comparing the performance of a cognitive stress paradigm across clinical and population samples was primarily motivated by the inherent differences in the characteristics of participants in the two samples: patients and volunteers at a memory disorders clinic versus randomly selected older adults in the community at large. The former represents a sample enriched for AD, with its inherent selection bias, while the latter has a much lower prevalence of AD but is representative of and therefore generalizable to the population from which it is drawn. It is therefore particularly remarkable that we found comparable performance of SIT subtests in predicting progression from MCI to dementia in the two samples.

However, the validity of our conclusions depends upon the comparability of the MCI and dementia diagnoses in both study settings. Even though the CDR was used to classify participants in both studies, clinic participants and their required informants underwent in-depth assessments by experienced clinicians, while population study participants, mostly without informants, had systematic protocol assessments by trained research interviewers. To assess the comparability of diagnoses across studies, we examined CDR Sum of Boxes and neuropsychological test scores for progressors to dementia in both samples at baseline and at the time of dementia diagnosis. Progressors in the clinical sample were slightly more impaired than those in the population sample at baseline based on CDR Sum of Boxes, but less impaired according to the neuropsychological test composite score. At the time of dementia diagnosis, the population sample progressors were more impaired according to both measures, and showed a significantly greater increase in CDR Sum of Boxes. While these between-sample differences were significant, their magnitudes were not large and were smaller than the magnitude of within-subject change from baseline to dementia diagnosis, which itself was not significantly different according to what may be considered the more objective measure of the two – the neuropsychological test composite. Nonetheless, the incorporation of neuropsychological measures explicitly into the MCI and dementia diagnoses may be a better approach to control for diagnostic methodological differences.

While the SIT-paradigm has advantages of being easy to administer, is more culture and language-fair, has minimum educational bias, and translates well from clinic versus population-based studies, it is only one measure, and future studies should determine if incremental predictive power in both population and clinical samples can be obtained by combining multiple cognitive measures tapping different domains. Newer cognitive stress tests such as the Loewenstein-Acevedo Scales of Semantic Interference and Learning (LASSI-L; analogous to an exercise electrocardiogram in cardiology) may be more sensitive to early disease and correlates more strongly with biological markers (Loewenstein *et al.*, 2017; Loewenstein *et al.*, 2016).

The results of this investigation demonstrates the importance of establishing the utility of both established and novel cognitive instruments in diverse samples. The target population and underlying prevalence of expected disease will make an important difference in establishing ideal thresholds. Different research questions will determine the need to opt for different balances between sensitivity and specificity, particularly as they relate to different study populations. Our study also highlights the varied uses of neuropsychological assessment: (a) differentiating normal from abnormal performance; (b) differentiating between different diagnostic groups; and (c) delineating those individuals at most risk for disease progression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Duara Ranjan and the Wien Center for Alzheimer's Disease and Memory Disorders at Mount Sinai Medical Center for collecting the clinical sample data. The work reported here was supported in part by the following grants: R01AG023651, K07AG044395, 1P50AG025711-05, and R01AG047649-01A1 from the National Institute on Aging, United States Department of Health and Human Services.

Description of Authors' Roles

J. C. Beer was responsible for statistical design and analyses and contributed to writing the manuscript including writing the first draft. B. E. Snitz and C.-C. H. Chang contributed to writing the manuscript. D. A. Loewenstein was responsible for design of the clinical study, is the developer of the Semantic Interference Test, and contributed to writing the manuscript. M. Ganguli was responsible for design of the population-based study and contributed to writing the manuscript.

References

- 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]
- Fuld P. The Fuld Object Memory Test. The Stoelting Instrument Company, Chicago, IL. 1981
- Ganguli M, et al. Cognitive test performance predicts change in functional status at the population level: the MYHAT Project. *J Int Neuropsychol Soc.* 2010a; 16:761–770. [PubMed: 20609270]
- Ganguli M, Chang C-CH, Snitz BE, Saxton JA, Vanderbilt J, Lee C-W. Prevalence of mild cognitive impairment by multiple classifications: the Monongahela-Youghiogheny Healthy Aging Team (MYHAT) project. *The American Journal of Geriatric Psychiatry.* 2010b; 18:674–683. [PubMed: 20220597]
- Ganguli M, Snitz B, Vander Bilt J, Chang CC. How much do depressive symptoms affect cognition at the population level? The Monongahela-Youghiogheny Healthy Aging Team (MYHAT) study. *Int J Geriatr Psychiatry.* 2009; 24:1277–1284. [PubMed: 19340894]
- Hughes CP, Berg L, Danziger WL, Coben LA, Martin R. A new clinical scale for the staging of dementia. *The British journal of psychiatry.* 1982; 140:566–572. [PubMed: 7104545]
- Loewenstein DA, Acevedo A, Agron J, Duara R. Vulnerability to proactive semantic interference and progression to dementia among older adults with mild cognitive impairment. *Dement Geriatr Cogn Disord.* 2007; 24:363–368. [PubMed: 17911981]
- Loewenstein DA, Acevedo A, Luis C, Crum T, Barker WW, Duara R. Semantic interference deficits and the detection of mild Alzheimer's disease and mild cognitive impairment without dementia. *Journal of the International Neuropsychological Society.* 2004; 10:91–100. [PubMed: 14751011]
- Loewenstein DA, Curiel RE, Duara R, Buschke H. Novel Cognitive Paradigms for the Detection of Memory Impairment in Preclinical Alzheimer's Disease. Assessment. 2017 1073191117691608.

- Loewenstein DA, et al. A Novel Cognitive Stress Test for the Detection of Preclinical Alzheimer Disease: Discriminative Properties and Relation to Amyloid Load. *Am J Geriatr Psychiatry*. 2016; 24:804–813. [PubMed: 27160985]
- Loewenstein DA, et al. Proactive semantic interference is associated with total and regional abnormal amyloid load in non-demented community-dwelling elders: A preliminary study. *The American Journal of Geriatric Psychiatry*. 2015; 23:1276–1279. [PubMed: 26525994]
- Lowenstein D. The Fuld OME as a more culture-fair memory test in the elderly. *Clinical Gerontologist: The Journal of Aging and Mental Health*. 1995
- Mungas D, Marshall SC, Weldon M, Haan M, Reed BR. Age and education correction of Mini-Mental State Examination for English and Spanish-speaking elderly. *Neurology*. 1996; 46:700–706. [PubMed: 8618670]
- Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and motor skills*. 1958; 8:271–276.
- Robin X, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*. 2011; 12:77. [PubMed: 21414208]
- Snitz BE, et al. A novel approach to assessing memory at the population level: vulnerability to semantic interference. *Int Psychogeriatr*. 2010; 22:785–794. [PubMed: 20105350]
- Therneau T. A Package for Survival Analysis in S. R package version 2.37–7. 2015. 2014. *URL*<https://cran.r-project.org/package=survival>
- Wechsler D. Wechsler memory scale-revised (WMS-R). Psychological Corporation; 1987.

Figure 1.

Receiver operating characteristic (ROC) curves depicting sensitivity and specificity for thresholds on baseline Semantic Interference Test subtests for predicting progression to dementia within 4 years of follow-up in the clinical (ADRC) and population-based (MYHAT) samples. Circles indicate optimal thresholds derived using Youden's J criterion; triangles indicate thresholds of maximum specificity yielding at least 70% sensitivity. Solid lines and plotting points are used for clinical sample and dashed lines and open plotting points are used for population-based sample.

Table 1

Participant Characteristics

	Memory Disorders Clinic (ADRC)				Population Study (MYHAT)				ADRC All vs MYHAT All P-value *
	NonProg N = 142	Prog N = 42	NonProg vs Prog P-value *	All N = 184	NonProg N = 390	Prog N = 45	NonProg vs Prog P-value *	All N = 435	
Age at baseline (years) Mean (SD)	75.0 (5.4)	76.5 (6.3)	0.151	75.3 (5.7)	78.8 (7.3)	82.8 (5.2)	< 0.001	79.2 (7.3)	< 0.001
Female N (%)	62 (43.7)	23 (54.8)	0.222	85 (46.2)	222 (56.9)	24 (53.3)	0.751	246 (56.6)	0.022
Years education Mean (SD)	13.4 (3.3)	13.9 (3.8)	0.450	13.5 (3.4)	12.6 (2.5)	12.0 (2.0)	0.061	12.5 (2.5)	< 0.001
Race / Ethnicity N (%)									
White Non-Hispanic	91 (64.1)	29 (69.0)	0.414	120 (65.2)	359 (92.1)	39 (86.7)	0.391	398 (91.5)	< 0.001
Black Non-Hispanic	14 (9.9)	1 (2.4)		15 (8.2)	27 (6.9)	6 (13.3)		33 (7.6)	
White Hispanic	35 (24.6)	11 (26.2)		46 (25.0)	3 (0.8)	0 (0.0)		3 (0.7)	
Other	2 (1.4)	1 (2.4)		3 (1.6)	1 (0.3)	0 (0.0)		1 (0.2)	
Follow-up time (years) Median (Min, Max)	2 (1, 4)	2 (1, 4)	0.091	2 (1, 4)	4 (1, 4)	4 (1, 4)	0.919	4 (1, 4)	< 0.001
CDR Sum of Boxes at baseline Mean (SD)	1.0 (0.8)	1.9 (1.0)	< 0.001	1.2 (0.9)	0.8 (0.5)	1.5 (0.9)	< 0.001	0.9 (0.6)	< 0.001
Fuld-OME/SIT Baseline Test Scores Mean (SD)									
Total Fuld Recall	21.8 (4.5)	17.3 (5.6)	< 0.001	20.8 (5.1)	20.4 (4.2)	15.3 (5.9)	< 0.001	19.9 (4.7)	0.034
SIT Bag B Recall	5.7 (2.0)	4.5 (2.1)	0.001	5.4 (2.1)	5.6 (1.7)	3.8 (2.4)	< 0.001	5.4 (1.9)	0.880
SIT Bag A Recall	4.0 (2.2)	2.4 (2.2)	< 0.001	3.6 (2.3)	4.4 (2.3)	2.9 (2.3)	< 0.001	4.3 (2.3)	0.002
SIT Combined Interference	9.7 (3.3)	6.9 (3.5)	< 0.001	9.1 (3.6)	10.0 (3.2)	6.7 (4.1)	< 0.001	9.7 (3.5)	0.051
Other Baseline Test Scores Mean (SD)									
MMSE	27.6 (1.8)	26.6 (1.8)	0.003	27.4 (1.8)	27.0 (2.1)	25.1 (2.3)	< 0.001	26.8 (2.2)	0.002
Logical Memory Imm	9.5 (4.2)	8.1 (3.9)	0.048	9.2 (4.2)	9.1 (4.0)	6.2 (3.6)	< 0.001	8.8 (4.1)	0.321
Logical Memory Del	7.9 (4.4)	5.0 (3.8)	< 0.001	7.3 (4.4)	6.3 (4.0)	3.2 (3.6)	< 0.001	6.0 (4.1)	0.001
Visual Reproduction Imm	23.9 (7.5)	21.7 (7.5)	0.093	23.4 (7.6)	25.9 (7.2)	20.8 (8.6)	< 0.001	25.4 (7.5)	0.003
Visual Reproduction Del	15.9 (9.5)	8.7 (9.1)	< 0.001	14.2 (9.9)	14.8 (10.3)	6.4 (7.8)	< 0.001	14.0 (10.4)	0.744
Trails B Time (s)	142 (74)	177 (81)	0.016	150 (77)	135 (54)	170 (54)	0.001	138 (54)	0.059
Test Composite	0.14 (0.74)	-0.37 (0.60)	< 0.001	0.02 (0.74)	0.04 (0.69)	-0.72 (0.66)	< 0.001	-0.04 (0.73)	0.351

* Welch's t-test for continuous variables, Fisher's exact test for categorical variables, Wilcoxon rank sum test for follow-up time and CDR Sum of Boxes

Notes: Total Fuld Recall assesses total list recall; SIT Bag B Recall assesses proactive semantic interference; SIT Bag A Recall assesses retroactive semantic interference, and SIT Combined Interference is the sum of the latter two scores. Test Composite is the mean of the 6 test scores directly above, after standardization of the scores using means and standard deviations calculated from the pooled clinic and population study baseline data (negative Trails B Time was used since higher scores indicate worse performance).

Abbreviations: ADRC: Alzheimer's Disease Research Center; MYHAT: Monongahela-Youghiogheny Healthy Aging Team; Prog: Progressors to dementia during follow-up; NonProg: Non-Progressors to dementia during follow-up; CDR: Clinical Dementia Rating; Fuld-OME: Fuld Object Memory Evaluation; SIT: Semantic Interference Test; MMSE: Mini-Mental State Examination; Imm: Immediate; Del: Delayed; Trails B: Trail Making Test B

ADRC missing: 1 Logical Memory Delayed, 1 Visual Reproduction Immediate, 1 Visual Reproduction Delayed, 4 Trails B Time

MYHAT missing: 10 Logical Memory Immediate, 18 Logical Memory Delayed, 18 Visual Reproduction Immediate, 26 Visual Reproduction Delayed, 49 Trails B Time

Table 2

Optimal Semantic Interference Test (SIT) thresholds (yielding maximum sensitivity + specificity) for predicting progression to dementia in the two samples (with 95% bootstrap confidence intervals)

	Optimal threshold	Sensitivity	Specificity	PPV	NPV	Accuracy	Youden's J	AUC
Total Fuld Recall								
Clinical sample	18.5 (17.5, 21.5)	59.5 (42.9, 88.1)	82.4 (55.6, 93.7)	50.0 (34.6, 69.8)	87.3 (84.0, 94.2)	77.2 (61.4, 84.8)	0.419 (0.287, 0.586)	0.751 (0.658, 0.837)
Population sample	17.5 (12.5, 19.5)	64.4 (40.0, 82.2)	78.2 (62.8, 97.2)	25.4 (19.0, 64.7)	95.0 (93.1, 97.1)	76.8 (64.4, 92.0)	0.426 (0.305, 0.579)	0.757 (0.670, 0.837)
SIT Bag B Recall								
Clinical sample	5.5 (3.5, 6.5)	66.7 (31.0, 95.2)	59.9 (25.4, 90.8)	32.9 (26.8, 54.8)	85.9 (80.7, 94.3)	61.4 (44.6, 78.8)	0.265 (0.132, 0.427)	0.666 (0.565, 0.751)
Population sample	3.5 (2.5, 5.5)	51.1 (35.6, 86.7)	88.7 (36.9, 93.3)	34.3 (15.4, 44.4)	94.0 (89.7, 95.9)	84.8 (58.6, 89.2)	0.398 (0.019, 0.545)	0.714 (0.373, 0.798)
SIT Bag A Recall								
Clinical sample	3.5 (1.5, 4.5)	73.8 (42.9, 90.5)	61.3 (42.3, 89.4)	36.0 (30.4, 59.1)	88.8 (83.1, 95.0)	64.1 (52.7, 81.0)	0.351 (0.227, 0.510)	0.706 (0.610, 0.793)
Population sample	2.5 (-0.5, 5.5)	48.9 (37.8, 100.0)	80.3 (0.0, 86.9)	22.2 (13.5, 29.4)	93.2 (89.7, 96.9)	77.0 (42.3, 89.7)	0.291 (0.000, 0.445)	0.677 (0.385, 0.758)
SIT Combined Interference								
Clinical sample	7.5 (7.5, 9.5)	69.0 (57.1, 85.7)	78.2 (59.2, 85.2)	48.3 (36.7, 58.6)	89.5 (85.7, 94.6)	76.1 (64.1, 82.1)	0.472 (0.325, 0.625)	0.736 (0.640, 0.824)
Population sample	7.5 (5.5, 10.5)	60.0 (37.8, 86.7)	78.2 (47.9, 93.3)	24.1 (15.9, 46.9)	94.4 (92.7, 97.0)	76.3 (51.7, 89.0)	0.382 (0.262, 0.538)	0.729 (0.639, 0.812)

Notes: Total Fuld Recall assesses total list recall; SIT Bag B Recall assesses proactive semantic interference; SIT Bag A Recall assesses retroactive semantic interference, and SIT Combined Interference is the sum of the latter two scores. Youden's J = sensitivity/100 + specificity/100 - 1

Table 3

Unadjusted and adjusted logistic regression odds ratios for progression to dementia during follow-up

	Unadjusted			Adjusted*		
	OR (95% CI)	z-value	P-value	OR (95% CI)	z-value	P-value
Total Fluid Recall						
Total Fluid < 18.5	5.50 (2.89, 10.89)	5.07	<0.001	2.33 (1.11, 4.98)	2.22	0.027
Clinical sample	2.77 (1.34, 5.79)	2.75	0.006	2.28 (0.89, 5.72)	1.75	0.079
(Total Fluid < 18.5)*(Clinical sample)	1.25 (0.46, 3.41)	0.44	0.660	1.41 (0.46, 4.41)	0.60	0.547
Bag B Recall						
Bag B < 5.5	2.53 (1.33, 5.04)	2.75	0.006	1.02 (0.48, 2.21)	0.05	0.960
Clinical sample	2.45 (1.11, 5.39)	2.24	0.025	2.06 (0.78, 5.38)	1.47	0.141
(Bag B < 5.5)*(Clinical sample)	1.18 (0.44, 3.16)	0.33	0.743	1.45 (0.48, 4.45)	0.66	0.509
Bag A Recall						
Bag A < 3.5	2.87 (1.54, 5.44)	3.28	0.001	1.36 (0.67, 2.77)	0.85	0.397
Clinical sample	1.76 (0.78, 3.78)	1.41	0.157	1.89 (0.72, 4.74)	1.34	0.180
(Bag A < 3.5)*(Clinical sample)	1.55 (0.58, 4.27)	0.87	0.383	1.64 (0.54, 5.09)	0.87	0.387
Combined Interference Score (CIS)						
CIS < 7.5	5.38 (2.85, 10.39)	5.13	<0.001	2.13 (1.01, 4.51)	1.99	0.047
Clinical sample	1.98 (0.92, 4.16)	1.80	0.072	2.12 (0.86, 5.08)	1.66	0.096
(CIS < 7.5)*(Clinical sample)	1.48 (0.55, 4.09)	0.77	0.439	1.43 (0.48, 4.33)	0.64	0.522

* Adjusted for age, sex, years of education, race/ethnicity, baseline CDR sum of boxes, and baseline neuropsychological test composite score.

Notes: Total Fluid Recall assesses total list recall; SIT Bag B Recall assesses proactive semantic interference; SIT Bag A Recall assesses retroactive semantic interference, and SIT Combined Interference is the sum of the latter two scores.

The reference level for the regression models is a test score above the noted threshold and membership in the population-based sample. Notation such as (Total Fluid < 18.5)*(Clinical sample) denotes the interaction between the binary variable encoding test score (equals 1 if score is below threshold, zero otherwise) and the binary variable indicating sample (equals 1 if participant is from the clinical sample, zero if from the population-based sample).

Abbreviations: SIT: Semantic Interference Test, CIS: Combined Interference Score, OR: odds ratio, CI: confidence interval

Table 4

Unadjusted and adjusted Cox proportional hazard ratios for time to dementia

	Unadjusted			Adjusted*		
	HR (95% CI)	z-value	P-value	HR (95% CI)	z-value	P-value
Total Fluid Recall						
Total Fluid < 18.5	6.06 (3.25, 11.28)	5.68	<0.001	2.82 (1.44, 5.49)	3.04	0.002
Clinical sample	4.03 (2.00, 8.13)	3.90	<0.001	4.25 (1.86, 9.69)	3.44	0.001
(Total Fluid < 18.5)*(Clinical sample)	1.08 (0.45, 2.60)	0.18	0.858	0.78 (0.31, 1.94)	-0.54	0.592
Bag B Recall						
Bag B < 5.5	2.73 (1.45, 5.13)	3.11	0.002	1.24 (0.64, 2.42)	0.63	0.529
Clinical sample	3.48 (1.65, 7.34)	3.27	0.001	3.70 (1.58, 8.69)	3.01	0.003
(Bag B < 5.5)*(Clinical sample)	1.07 (0.43, 2.62)	0.14	0.890	0.96 (0.38, 2.43)	-0.08	0.933
Bag A Recall						
Bag A < 3.5	3.28 (1.81, 5.93)	3.92	<0.001	1.79 (0.96, 3.31)	1.84	0.066
Clinical sample	2.67 (1.26, 5.64)	2.57	0.010	3.91 (1.72, 8.87)	3.26	0.001
(Bag A < 3.5)*(Clinical sample)	1.17 (0.47, 2.90)	0.34	0.733	0.81 (0.32, 2.05)	-0.45	0.650
Combined Interference Score (CIS)						
CIS < 7.5	5.89 (3.24, 10.71)	5.81	<0.001	2.69 (1.41, 5.14)	3.00	0.003
Clinical sample	2.93 (1.43, 6.01)	2.93	0.003	3.87 (1.75, 8.57)	3.34	0.001
(CIS < 7.5)*(Clinical sample)	1.21 (0.50, 2.93)	0.42	0.676	0.82 (0.33, 2.02)	-0.44	0.659

* Adjusted for age, sex, years of education, race/ethnicity, baseline CDR sum of boxes, and baseline neuropsychological test composite score.

Notes: Total Fluid Recall assesses total list recall; SIT Bag B Recall assesses proactive semantic interference, SIT Bag A Recall assesses retroactive semantic interference, and SIT Combined Interference is the sum of the latter two scores.

The reference level for the regression models is a test score above the noted threshold and membership in the population-based sample. Notation such as (Total Fluid < 18.5)*(Clinical sample) denotes the interaction between the binary variable encoding test score (equals 1 if score is below threshold, zero otherwise) and the binary variable indicating sample (equals 1 if participant is from the clinical sample, zero if from the population-based sample).

Abbreviations: SIT: Semantic Interference Test, CIS: Combined Interference Score, HR: hazard ratio, CI: confidence interval