Table S1. Optimal Semantic Interference Test (SIT) thresholds (yielding maximum specificity when sensitivity ≥ 70%) for predicting progression to dementia in the two samples (with 95% bootstrap confidence intervals)

	Optimal threshold	Sensitivity	Specificity	PPV	NPV	Accuracy	Youden's J
Total Fuld Recall							
Clinical sample	20.5 (18.5, 23.5)	71.4 (71.4, 83.3)	64.8 (35.9, 83.8)	37.5 (26.3, 57.7)	88.5 (83.9, 92.4)	66.3 (45.7, 81.5)	0.362 (0.131, 0.581)
Population sample	19.5 (16.5, 22.5)	73.3 (71.1, 82.2)	63.1 (32.3, 82.8)	18.6 (11.7, 33.0)	95.3 (92.3, 96.8)	64.1 (37.0, 81.8)	0.364 (0.098, 0.567)
SIT Bag B Recall							
Clinical sample	6.5 (5.5, 6.5)	83.3 (71.4, 90.5)	37.3 (23.9, 65.5)	28.2 (24.6, 39.7)	88.3 (80.7, 93.4)	47.8 (40.8, 68.5)	0.207 (0.058, 0.427)
Population sample	6.5 (2.5, 6.5)	88.9 (71.1, 93.3)	32.3 (6.4, 57.9)	13.2 (11.8, 34.3)	96.2 (91.8, 98.0)	38.2 (34.5, 86.9)	0.212 (-0.166, 0.374)
SIT Bag A Recall							
Clinical sample	3.5 (2.5, 4.5)	73.8 (71.4, 88.1)	61.3 (36.6, 71.8)	36.0 (26.8, 44.3)	88.8 (84.4, 93.9)	64.1 (46.2, 72.8)	0.351 (0.145, 0.500)
Population sample	4.5 (1.5, 5.5)	73.3 (71.1, 86.7)	50.5 (11.8, 67.7)	14.6 (11.5, 21.5)	94.3 (90.7, 96.7)	52.9 (37.0, 81.6)	0.238 (-0.119, 0.408)
SIT Combined Interference							
Clinical sample	8.5 (7.5, 10.5)	73.8 (71.4, 83.3)	69.0 (38.7, 83.8)	41.3 (26.1, 58.1)	89.9 (82.9, 93.7)	70.1 (46.7, 82.1)	0.428 (0.118, 0.619)
Population sample	9.5 (7.5, 10.5)	71.1 (71.1, 86.7)	58.2 (41.0, 79.2)	16.4 (12.7, 28.9)	94.6 (93.0, 96.8)	59.5 (45.0, 78.6)	0.293 (0.144, 0.526)

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.

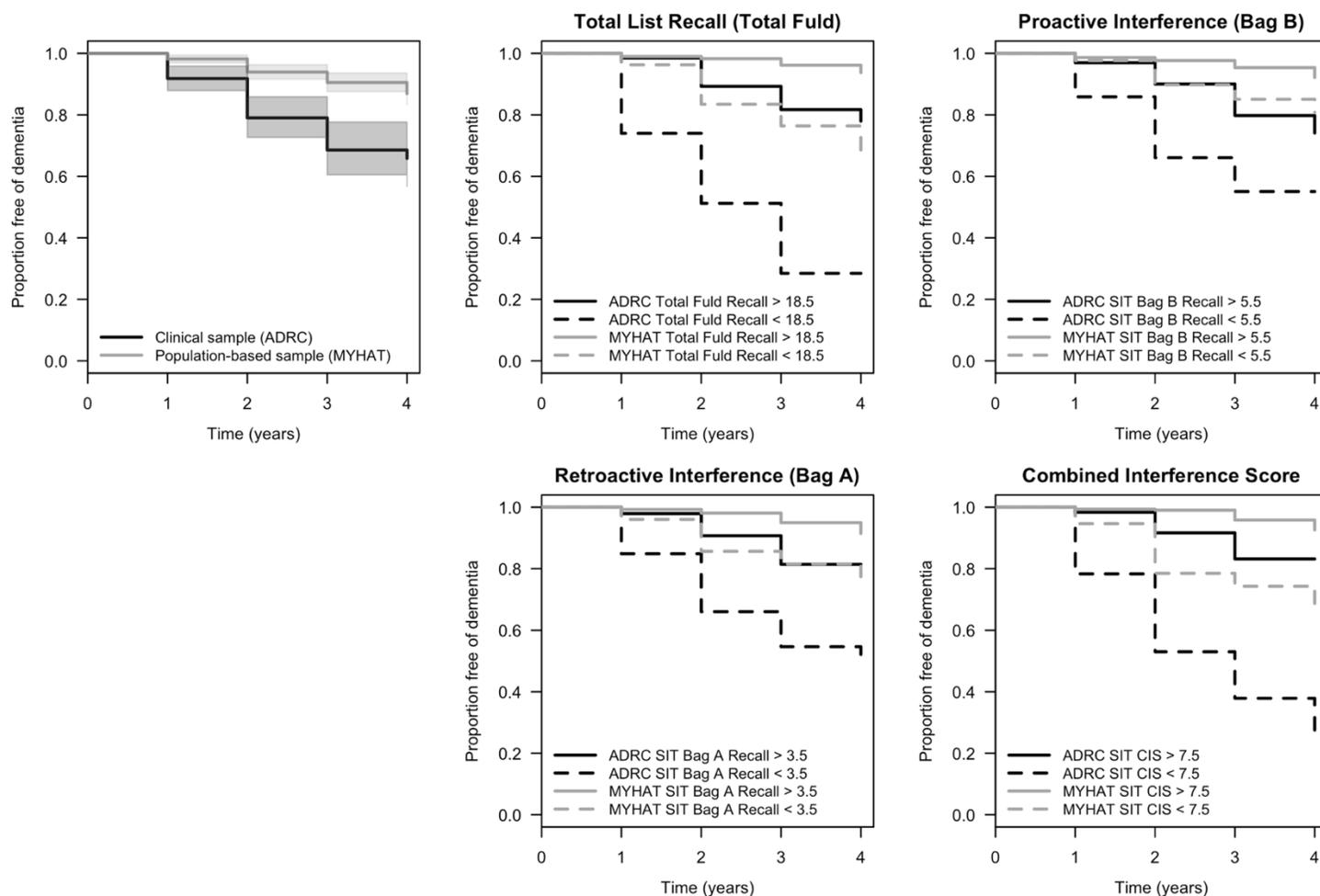Youden's J = sensitivity/100 + specificity/100 − 1

Figure S1. Kaplan-Meier survival curves depicting rates of progression to dementia for the clinical and population-based samples overall (top left, with 95% confidence bands) and stratified by the optimal thresholds found using Youden's J criterion for each Semantic Interference Test subtest.
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.
Abbreviations: ADRC: Alzheimer's Disease Research Center, clinical sample; MYHAT: Monongahela-Youghiogheny Healthy Aging Team (MYHAT) study; population-based sample; SIT: Semantic Interference Test; CIS: Combined Interference Score.

Supplementary material for: Does a cognitive stress test predict progression from mild cognitive impairment to dementia equally well in clinical versus population-based settings? Beer et al. 2017

Table S2. Mean (SD) scores at baseline and at endpoint for progressors from mild cognitive impairment (CDR = 0.5) to dementia (CDR ≥ 1) in clinical and population-based samples

	Baseline (MCI)			Endpoint (Dementia)		
	Clinical sample N = 42	Population sample N = 45	P-value*	Clinical sample N = 42	Population sample N = 45	P-value*
CDR Sum of Boxes	1.9 (1.0)	1.5 (0.9)	0.027	4.5 (2.2)	5.2 (1.6)	0.002
Total Fuld Recall	17.3 (5.6)	15.3 (5.9)	0.104	12.1 (6.4)	8.1 (5.8)	0.012
SIT Bag B	4.5 (2.1)	3.8 (2.4)	0.162	2.1 (1.8)	1.5 (1.3)	0.143
SIT Bag A	2.4 (2.2)	2.9 (2.3)	0.268	1.6 (1.9)	1.3 (1.7)	0.462
SIT Bag B + Bag A	6.9 (3.5)	6.7 (4.1)	0.858	3.7 (3.1)	2.8 (2.7)	0.214
MMSE	26.6 (1.8)	25.1 (2.3)	0.001	23.2 (3.9)	21.1 (4.0)	0.025
Logical Memory Imm	8.1 (3.9)	6.2 (3.6)	0.025	5.6 (4.0)	3.3 (3.0)	0.007
Logical Memory Del	5.0 (3.8)	3.2 (3.6)	0.029	2.8 (3.6)	0.6 (1.5)	< 0.001
Visual Reproduction Imm	21.7 (7.5)	20.8 (8.6)	0.618	15.9 (4.6)	15.5 (9.2)	0.792
Visual Reproduction Del	8.7 (9.1)	6.4 (7.8)	0.216	3.9 (4.2)	2.0 (4.0)	0.098
Trails B Time (s)	176.7 (80.7)	170.3 (54.1)	0.688	231.1 (87.5)	217.3 (37.2)	0.462
Test Composite	-0.37 (0.60)	-0.72 (0.66)	0.011	-1.24 (0.81)	-1.77 (0.85)	0.004
Within-subject change in CDR Sum of Boxes				2.6 (2.4)	3.7 (1.8)	0.013
Within-subject change in Test Composite				-0.88 (0.98)	-1.05 (1.00)	0.425

* Welch's t-test for all variables except CDR Sum of Boxes which used Wilcoxon rank sum test

Notes: Test Composite is the mean of the 6 test scores directly above, after normalization 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: CDR: Clinical Dementia Rating; MCI: mild cognitive impairment; SIT: Semantic Interference Test; MMSE: Mini-Mental State Examination; Imm: Immediate; Del: Delayed; Trails B: Trail Making Test B

Clinical sample baseline missing: 1 Logical Memory Del

Population sample baseline missing: 1 Logical Memory Imm, 2 Logical Memory Del, 3 Visual Reproduction Imm, 3 Visual Reproduction Del, 14 Trails B Time

Clinical sample endpoint missing: 1 CDR Sum of Boxes, 11 Total Fuld OME, 8 SIT Bag B, 8 SIT Bag A, 8 SIT Bag B + Bag A, 1 MMSE, 1 Logical Memory Imm, 1 Logical Memory Del, 13 Visual Reproduction Imm, 14 Visual Reproduction Del, 7 Trails B Time

Population sample endpoint missing: 14 Total Fuld OME, 14 SIT Bag B, 14 SIT Bag A, 14 SIT Bag B + Bag A, 14 Logical Memory Imm, 14 Logical Memory Del, 15 Visual Reproduction Imm, 16 Visual Reproduction Del, 34 Trails B Time

Supplementary material for: Does a cognitive stress test predict progression from mild cognitive impairment to dementia equally well in clinical versus population-based settings? Beer et al. 2017