Clinical Validation and Machine Learning Optimization of MyCog:

A Self-Administered Cognitive Screener

Stephanie Ruth Young, PhD*[1]; Yusuke Shono, MA, PhD[1,2]; Katherina Hauner, PhD[1,3]; Elizabeth M. Dworak, PhD[1]; Maxwell Mansolf[1]; Laura Curtis, MS[1]; Julia Yoshino Benavente, MPH[1]; Stephanie Batio, MS[1]; Richard C. Gershon, PhD[1]; Michael S. Wolf, MA, MPH, PhD[4]; Cindy J. Nowinski, MD, PhD[1]

1. Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL
2. School of Community and Global Health, Claremont Graduate University, Claremont, CA
3. Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL
4. Center for Applied Health Research on Aging, Northwestern University Feinberg School of Medicine, Chicago, IL

*Corresponding author email: stephanieruth.young@northwestern.edu

Word count: 3124

Abstract

Importance: Early detection of cognitive impairment in older adults is crucial for identifying reversible causes, managing symptoms, and enabling future planning; yet primary care settings face significant barriers to implementing effective cognitive screening.

Objective: To clinically validate an innovated self-administered, tablet-based cognitive screening tool ("MyCog", adapted from the NIH Toolbox Cognitive Battery) and to develop an optimized scoring algorithm for detecting cognitive impairment in older adults during primary care visits.

Design: Cross-sectional clinical validation study conducted between summer 2021 and fall 2024, comparing cognitive performance between impaired and non-impaired participants using multiple machine learning approaches to identify optimal predictors of impairment.

Setting: Academic medical centers and community-based general population recruitment.

Participants: 248 adults aged 65 and older, including 70 participants with clinically diagnosed cognitive impairment (MCI or dementia), 81 participants confirmed to be cognitively intact from clinical settings, and 97 participants from the general population. Participants were predominantly White (88%) with varied educational backgrounds and a mean age of 76.3 years (*SD* = 6.9).

Main Outcomes and Measures: Sensitivity, specificity, and area under the receiver operating characteristic curve (ROC AUC) were the primary metrics used for evaluating model fit in classifying cognitive impairment of known groups from the MyCog Picture Sequence Memory and Dimensional Change Card Sort tasks.

Results: Across nine machine learning model classes, Picture Sequence Memory exact match score and Dimensional Change Card Sort rate-correct score were identified as the most robust predictors of cognitive impairment. The final optimized model achieved a sensitivity of 82.5% (bootstrapped 95% CI [0.733, 0.913]) and a specificity of 76.8% (bootstrapped 95% CI [0.709, 0.828]) with an ROC AUC of 0.892, comparable to traditional, long form cognitive screening measures while requiring less than 7 minutes to self-administer.

Conclusions and Relevance: MyCog demonstrates strong diagnostic accuracy for detecting cognitive impairment in older adults with a parsimonious model focusing on episodic memory and executive function. As a brief, self-administered screening tool that can be integrated into electronic health records, MyCog addresses key barriers to cognitive screening in primary care settings, which may help increase screening and early detection rates.

Introduction

Cognitive impairment in older adulthood is a national and global health crisis[1]. Cognitive impairments may result from Alzheimer's disease (AD) and AD-related dementias (ADRD)[2-5], neurological disorders (e.g., Parkinson's disease[6], multiple sclerosis[7]), or reversible causes such as infections or medications[8]. Early detection of cognitive impairment can help identify reversible causes, address symptoms and comorbidities, determine appropriate caregiver involvement, and help families plan for the future[9,10]. Primary care visits offer an important opportunity for early detection of cognitive impairment in adults over age 65[11], especially for underserved or rural populations for whom specialist visits may not be feasible. However, primary care practices face barriers to screening, such as time and resource constraints[12], and may not have proper training on next steps following a positive results[13]. Clinicians lack standardized assessments that accurately detect cognitive impairment and are practical in real-world workflows, leading to frequent under-detection of cognitive impairment and dementia in primary care settings[14].

Self-administered, app-based cognitive screening systems may help overcome barriers in primary care. MyCog, a brief, self-administered cognitive screening tablet app was funded by the NIH to detect cognitive impairment in older adults during a primary care visit[15]. Adapted from the NIH Toolbox® for Assessment of Neurological and Behavioral Function Cognitive Battery (NIHTB-CB), MyCog consists of two cognitive tasks: (1) Dimensional Change Card Sort[16] (MyCog DCCS), which assess executive functioning and cognitive flexibility, and (2) Picture Sequence Memory[17] (MyCog PSM), which assesses episodic memory. Both measures have ample evidence of strong psychometric properties within the NIHTB[17-19]. The MyCog battery is designed to be brief (less than 7 minutes total) and easy to self-administer during the rooming process of a primary care visit. The MyCog measures have demonstrated strong evidence of reliability and construct validity, comparable to well-established, traditional cognitive assessments[20]. The two cognitive domains measured by MyCog—episodic memory (MyCog PSM) and executive function (MyCog DCCS)—align with clinical understanding of early changes associated with cognitive decline[21-23]. Episodic memory impairment is often among the earliest manifestations of

Alzheimer's pathology[21,22], while executive dysfunction measured by MyCog DCCS captures frontal-subcortical circuit disruption common in vascular cognitive impairment and other dementias[24]. MyCog DCCS measures both speed and accuracy, which is consistent with the subtle processing inefficiencies that often precede more obvious cognitive errors in early impairment[25,26].

MyCog has been involved in three large-scale NIH-funded grants to date (5UH3NS105562, U01NS105562, AND R01AG069762), and is currently used by three large health systems around the country, including Oak Street Health, Access Community Health, and Northwestern Medicine. The app integrates with two large electronic health records, Epic and Greenway, and triggers customized clinical decision-making support following a positive result. MyCog has been self-administered by over 6,000 patients in 16 primary care clinics, with preliminary evidence supporting its sensitivity and specificity[18].

The present study builds upon this preliminary work via a robust clinical validation of MyCog in a combined sample of older adults with and without cognitive impairment. We leveraged nine modeling techniques combining frequentist regularization methods, Bayesian, and machine learning approaches to identify robust predictors of cognitive impairment, which were then used to create an optimized scoring algorithm for MyCog. The goal of this study was to determine the best derived scores for differentiating between individuals with and without cognitive impairment through a systematic investigation of multiple scoring methods.

Method

All methods were approved by Northwestern University's Institutional Review Board (STU00214921). Participants provided informed consent and were compensated for their participation. Participants with diagnoses of cognitive impairment were administered the Decision Making Assessment Tool (DMAT) adapted from the Brief Assessment of Capacity to Consent (UBACC) to determine their ability to provide informed consent at the time of enrollment.

Sample

Participants were recruited from two samples: a clinical sample, from Northwestern Medical Group (NMG) Geriatrics and General Internal Medicine and Regional Medical Group (RMG) Movement

Disorders and Neurodegenerative Diseases Center, as well as a general population sample recruited via a market research agency (Table 1). The clinical sample consisted of two subsamples of participants, including a group with cognitive impairment and a group without. Inclusion criteria for the cognitively impaired clinical subsample required that participants: (a) were 65 years or older; (b) spoke English; (c) had an upcoming appointment (within 4 weeks of study interview) or recent appointment (within 6 months of study interview) at NMG/RMG or Neurobehavior and Memory Clinic; and (d) had a documented diagnosis in their chart of impairment classification (MCI or dementia; Table 2) . Inclusionary criteria for the non-impaired clinical subsample required that participants were: (a) 65 years or older; (b) spoke English; and (c) obtained a Montreal Cognitive Assessment (MoCA) score ≥ 26 or Mini-Cog screener score ≥ 3, as administered at a recent appointment (within 6 months of validation interview) at NMG/RMG, or during the validation screening procedure. For both impaired and non-impaired clinical subsamples, exclusionary criteria included severe, uncorrected vision or hearing impairment.

The non-impaired clinical subsample was supplemented with a subsample of participants from the general population (Table 1) recruited by a market research agency as part of the MyCog validation study[20] conducted in parallel with a large-scale norming study of the NIH Toolbox Cognitive Battery (NIHTB-CB)[27-30]. Inclusion criteria for the general population sample required that participants were over the age of 65, spoke English, and provided informed consent. Participants in the general population sample were not screened for impairments prior to inclusion.

Procedure

All participants self-administered MyCog in dedicated research or clinical facilities alongside a larger battery of neuropsychological measures between the summer of 2021 and fall of 2024. For the clinical sample, diagnostic data were collected via a review of electronic medical records in the institution's Electronic Data Warehouse.

Measures

MyCog Dimensional Change Card Sort (MyCog DCCS) Task. Adapted from the NIH Toolbox Cognitive Battery (NIHTB-CB), the MyCog DCCS task is an executive function measure designed to assess cognitive flexibility[29]. This task requires respondents to sort images by color and shape. Each trial consists of two "reference" images differing in color and shape that are presented at the bottom of the screen. A third "target" image appears in the center of the screen, and respondents must match the target image to one of the reference images according to the prompted dimension (color or shape). Trials are subdivided into the following five blocks: (1) two practice trials that demonstrate matching by *shape*, (2) five live trials that require matching only by *shape*, (3) two practice trials that demonstrate matching only by *color*, (4) five live trials that require matching by *color*, (5) 30 (mixed) trials that require matching by shape or color in a pre-determined, pseudo-random order. Blocks that require matching only by shape are referred to as *pre-switch*, because a new dimension (color) has not yet been introduced. Blocks that require matching by shape are referred to as *post-switch*, because the dimension to be matched has switched from color to shape.

Inter-trial intervals (ITIs) are fixed at 1000 ms in the practice, pre-switch, and post-switch blocks. In the mixed trials, ITIs are fixed at either 300 ms or 1000 ms in pseudo-random order. Response times (RTs) are calculated as the number of milliseconds between target image onset and response delivery, with outliers defined as less than 100 ms[31]. Total testing time is approximately 2 to 3 min.

Both response time and errors are captured on the MyCog DCCS measure. The MyCog DCCS accuracy score is calculated by dividing the number of correct responses by the total number of items to which the participant responded, with higher scores indicating better performance. The MyCog DCCS speed score is calculated as the median RT in milliseconds across the 30 mixed trials (with outliers removed), with lower speeds indicating faster reaction times. Finally, the MyCog DCCS integrated speed-accuracy score is computed using both the rate-correct score (RCS; Equation 1)[32] and linear integrated speed-accuracy score (LISAS; Equation 2)[33].

$$RCS = \frac{Accuracy}{Med(RT_c) + Med(RT_E)} \quad (1)$$

$Accuracy$ refers the number of correct responses and $Med(RT_c)$ and $Med(RT_E)$ are the median RT across correct and error trials, respectively. RCS is interpreted as the number of correct responses adjusted for the total time spent. Higher RCS's indicate better performance.

$$LISAS = Med(RT_c) + (1 - PE)\,(\frac{SD_{RT}}{SD_{PE)}}) \quad (2)$$

$Med(RT_c)$ is the median correct RT, PE is mean error proportion, SD_{RT} is the standard deviation of RT, and SD_{PE} is the standard deviation of the error proportion. LISAS is interpreted as the median RT adjusted for error. Lower LISAS indicate better performance.

MyCog Picture Sequence Memory (MyCog PSM) Task. Also adapted from NIHTB-CB, the MyCog PSM task is a measure of episodic memory that requires respondents to memorize a sequence of images. The task begins with a practice trial in which 4 images are presented sequentially in pseudo-random temporal order, each accompanied by a brief audio description. All images depict easily identifiable activities with no inherent order related to the theme of "A day with a friend." The images are presented in order with a simple audio description, then scrambled, and the patient is asked to arrange the pictures in the order they were shown. Total testing time is approximately 4 to 5 minutes. Scores considered in this study were based on the cumulative number of correct image placements (exact match) or the number of images correctly placed next to another correct image (adjacent pairs).

Analysis

Statistical analyses were performed using R version 4.2.0[34]. Multiple scores and their combinations were used to examine which best contributed to a model predicting cognitive impairment across different trial conditions (tables 3 and 4).

Prior to model construction, data underwent a preprocessing and preparation stage. We first compared the descriptives of the non-impaired clinical sample to the general population sample to ensure combining the samples was appropriate to serve as an overall control sample. After determining that the samples could be combined into one group (described in Results), a new outcome variable representing the diagnostic groups was created representing participants who were non-impaired (0) and cognitively impaired (1). After finalizing the outcome variable, all predictors were standardized prior to model fitting.

Prior to model building, we assessed the extent of missing data across all variables. Overall, missing data ranged from 0% to 3.21% (PSM exact match) across the included variables, with an average missingness of 0.66%. Missing data were handled using random forest imputation via the mice package[35], with 5 imputed datasets created using 10 classification trees per variable. The imputation model included all study variables (both the cognitive impairment outcome and the predictive variables) to preserve relationships between variables while estimating missing values.

Given that many of the scores we used were similar (e.g., different ways of scoring accuracy and speed), we employed multiple modeling approaches that handle potential collinearity. We implemented nine different supervised model classes for the prediction of cognitive impairment: Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and stepwise regression (each with both frequentist and Bayesian implementations), and Random Forest. Ridge regression and Elastic Net explicitly manage collinearity through their regularization approaches.

The Bayesian implementations used weakly informative priors that approximate the behavior of their frequentist counterparts - larger prior scales (2.5) for Bayesian Ridge allowing more parameter flexibility, smaller scales (1.0) for Bayesian LASSO encouraging more shrinkage, and intermediate scales (1.5) for Bayesian Elastic Net. These priors influence coefficient estimation in the presence of correlated predictors by controlling the degree of shrinkage toward zero, which helps stabilize estimates when predictors are highly correlated.

Multiple model classes were tested to enhance variable selection robustness and avoid model-dependent conclusions. Rather than selecting a final model based on performance metrics from models

constructed using these nine model classes, the developed models were used to identify the predictors that were consistently important across these models, increasing confidence that our parsimonious final model reflects true clinical signals rather than artifacts.

For regularized regression models (Ridge, LASSO, Elastic Net), we employed 5-fold cross-validation without repeats to select optimal hyperparameters. The primary hyperparameter tuned was lambda (penalty strength), using cv.glmnet's default sequence of 100 logarithmically-spaced values, selecting the value that minimized cross-validated error ('lambda.min'). For Elastic Net specifically, we used a fixed alpha value of 0.5 as the mixing parameter, balancing between L1 (LASSO) and L2 (Ridge) penalties. Since Ridge regression shrinks all coefficients rather than selecting variables, we applied a coefficient threshold of 0.1 to identify meaningful variable selection[36]. The random forest model used 500 trees with the default *mtry* parameter (equal to the square root of the total number of predictors) for feature sampling at each split, along with out-of-bag error estimation.

Models were evaluated using sensitivity, specificity, and area under the receiver operating characteristic curve (ROC AUC), using .7 as the threshold for acceptable sensitivity and specificity, and .8 for acceptable AUC[37,38]. Model stability was assessed through agreement analysis between different approaches, with calculation of pairwise agreement rates and Cohen's kappa statistics. Finally, we selected predictors consistently identified across all nine modeling approaches to use in logistic regression algorithm that could be programmed into the MyCog app. The logistic regression was optimized to balance sensitivity and specificity for a cognitive screener, with a minimum sensitivity of .8 and a minimum specificity of .75[37].

Results

The non-impaired clinical sample was significantly older (M = 76.07, SD = 5.91) than the general population sample (M = 74.01, SD = 6.57), t = 2.20, P = .029, which may have contributed to better performance for speed-accuracy measures on MyCog DCCS in the general population sample. Reaction time, RCSs, and intra-item variability scores (RISD measures) were all significantly better in the general

population sample across various conditions ($P < .01$ for RISD measures), though raw ISD and MSSD measures showed no significant differences (P values ranging from .678 to .996). One exception was the accuracy cost for non-dominant versus dominant trials, where the non-impaired clinical sample demonstrated a smaller performance decrement (0.01) compared to the general population sample (-0.03; $P = .05$). The non-impaired clinical sample also had faster median response times for correct non-dominant (color) trials (963.08 ms vs. 861.45 ms, $P = .03$). Given that the general population performed as well or better than the non-impaired clinical sample on the majority of tasks, we combined these two samples to serve as the control group in the subsequent analyses.

Model stability analysis showed strong agreement between different predictive approaches, with correlation coefficients between models ranging from 0.84 to 1.00, indicating high consistency. Agreement rates between prediction classifications ranged from 87.0% to 100.0%, with Cohen's kappa values between 0.64 and 1.00, suggesting substantial to excellent agreement. The nine models all demonstrated good sensitivity and specificity (Table 5), selecting 4 to 26 of the 33 scores tested. The MyCog PSM exact match score and MyCog DCCS overall rate correct score were the only two predictors that appeared in all 9 models.

The final logistic regression model (Equation 3), optimized to maintain a minimum sensitivity of 0.80 and a minimum specificity of 0.75, demonstrated strong classification performance with an AUC of 0.892. At the optimal cut point of 0.209, the model achieved a sensitivity of 82.5% (bootstrapped 95% CI [0.733, 0.913]), a specificity of 76.8% (bootstrapped 95% CI [0.709, 0.828]), and an overall accuracy of 78.3% (bootstrapped 95% CI [0.73, 0.83]).

$$P_{impaired} = \frac{1}{1 + e^{-(3.48 - 0.43(PSM{:}Exact\ Match) - 3.63(DCCS{:}\ Rate\ Correct\ Score))}} \quad (3)$$

Discussion

Our findings demonstrate that MyCog (a brief, self-administered cognitive screening tool adapted from the NIH Toolbox Cognitive Battery) can effectively identify cognitive impairment in older adults. The final parsimonious model incorporated just two metrics—the MyCog PSM exact match score and the MyCog DCCS overall rate correct score — and achieved excellent discrimination with an AUC of 0.892, with a sensitivity of 82.5% and specificity of 76.8%.

MyCog's level of accuracy is comparable to other commonly used screening tools, such as the Montreal Cognitive Assessment (MoCA), which typically demonstrates sensitivities of 81-84% and specificities of 75-86% for detecting MCI and dementia[39,40]. However, MyCog offers several distinct advantages over traditional tools. First, it eliminates the need for a trained examiner, reducing variability in scoring and interpretation, and triggers appropriate clinical decision-making support within the EHR, ensuring patients are offered proper follow-up. The self-administered nature of MyCog also helps overcome workforce constraints in busy primary care settings, where provider time is limited and specialized training in cognitive assessment may be unavailable.

Importantly, while we tested numerous derived metrics incorporating various aspects of cognitive performance (including complex intra-individual variability and condition-specific effects), our most robust model emerged from relatively straightforward measures. This simplicity enhances clinical utility and interpretability. The probabilistic formula we derived can be easily implemented within the app, enabling immediate risk stratification at the point of care.

The ability to effectively screen for cognitive impairment in primary care settings has significant implications for addressing healthcare disparities. Early detection allows for timely intervention, management of modifiable risk factors, and appropriate care planning[41]. Moreover, as new disease-modifying treatments emerge that are only effective during mild stages of ADRD, early detection is becoming more and more critical in easily accessible, primary care settings[42].

Limitations

This study has several limitations that should be considered when interpreting its results. First, to supplement the non-impaired sample, we combined a clinical sample that was assessed and determined not to have cognitive impairment with a general population sample that was not assessed, with unknown impairment status. Although performance suggests our general population sample was similar to the non-impaired sample, it is possible that participants in the general population sample had diagnoses related to cognitive impairment. Moreover, our samples were recruited from academic medical centers and research registries, which may limit generalizability to other settings, such as community-based primary care or rural health clinics. The sample was also predominantly White and highly educated, particularly in the non-impaired clinical group, which may not reflect the demographic diversity seen in many primary care clinics. However, MyCog's current use in several large and diverse healthcare systems suggests that it is suitable for patients across racial, ethnic, socioeconomic, and regional groups. It is also important to note the cross-sectional design cannot address the predictive validity of MyCog for detecting future cognitive decline or conversion to dementia. Longitudinal studies are needed to determine whether MyCog can identify individuals at risk for progression before symptoms impact daily functioning.

Finally, the optimization of the algorithm to favor specificity over sensitivity reflects a particular clinical perspective that may not be universally applicable. While reducing false positives is valuable in high-volume primary care settings, different clinical contexts (such as memory clinics or research studies) might benefit from alternative thresholds that prioritize sensitivity. Future implementations should consider providing adjustable thresholds with corresponding performance metrics to allow clinicians to tailor the tool to their specific practice needs.

Conclusion

MyCog demonstrates strong diagnostic accuracy for detecting cognitive impairment in older adults through a parsimonious model focusing on episodic memory and executive function. By addressing key implementation barriers in primary care settings—requiring only minutes to self-administer during the rooming process—MyCog has potential to increase screening rates and facilitate earlier detection of cognitive impairment. As a validated tool that integrates with electronic health records and triggers

appropriate clinical decision support, MyCog represents a practical solution to improve cognitive health monitoring in primary care.

Acknowledgments

This work was funded by a grant from the NIH (R01AG074245-01). The funding agency played no role in the study design, collection of data, analysis, or interpretation of data.

References

1. Time to confront the global dementia crisis. *Lancet Neurol*. Sep 2008;7(9):761. doi:10.1016/S1474-4422(08)70175-3
2. 2021 Alzheimer's disease facts and figures. *Alzheimers Dement*. Mar 2021;17(3):327-406. doi:10.1002/alz.12328
3. Bature F, Guinn BA, Pang D, Pappas Y. Signs and symptoms preceding the diagnosis of Alzheimer's disease: a systematic scoping review of literature from 1937 to 2016. *BMJ Open*. Aug 28 2017;7(8):e015746. doi:10.1136/bmjopen-2016-015746
4. Corey-Bloom J. The ABC of Alzheimer's disease: cognitive changes and their management in Alzheimer's disease and related dementias. *Int Psychogeriatr*. 2002;14 Suppl 1:51-75. doi:10.1017/s1041610203008664
5. Lindeboom J, Weinstein H. Neuropsychology of cognitive ageing, minimal cognitive impairment, Alzheimer's disease, and vascular cognitive impairment. *Eur J Pharmacol*. Apr 19 2004;490(1-3):83-6. doi:10.1016/j.ejphar.2004.02.046
6. Nagano-Saito A, Bellec P, Hanganu A, et al. Why Is Aging a Risk Factor for Cognitive Impairment in Parkinson's Disease?-A Resting State fMRI Study. *Front Neurol*. 2019;10:267. doi:10.3389/fneur.2019.00267
7. Jakimovski D, Weinstock-Guttman B, Roy S, et al. Cognitive Profiles of Aging in Multiple Sclerosis. *Front Aging Neurosci*. 2019;11:105. doi:10.3389/fnagi.2019.00105
8. Chari D, Ali R, Gupta R. Reversible dementia in elderly: Really uncommon? *J Geriatr Ment Health*. 2015;2:30-37.
9. Chang CY, Silverman DH. Accuracy of early diagnosis and its impact on the management and course of Alzheimer's disease. *Expert Rev Mol Diagn*. Jan 2004;4(1):63-9. doi:10.1586/14737159.4.1.63
10. Rasmussen J, Langerman H. Alzheimer's Disease - Why We Need Early Diagnosis. *Degener Neurol Neuromuscul Dis*. 2019;9:123-130. doi:10.2147/DNND.S228939
11. Brayne C, Fox C, Boustani M. Dementia screening in primary care: is it time? *JAMA*. Nov 28 2007;298(20):2409-11. doi:10.1001/jama.298.20.2409
12. Day H, Eckstrom E, Lee S, Wald H, Counsell S, Rich E. Optimizing health for complex adults in primary care: current challenges and a way forward. *J Gen Intern Med*. Jun 2014;29(6):911-4. doi:10.1007/s11606-013-2749-x
13. Eggenberger E, Heimerl K, Bennett MI. Communication skills training in dementia care: a systematic review of effectiveness, training content, and didactic methods in different care settings. *Int Psychogeriatr*. Mar 2013;25(3):345-58. doi:10.1017/S1041610212001664
14. Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis Assoc Disord*. Oct-Dec 2009;23(4):306-14. doi:10.1097/WAD.0b013e3181a6bebc
15. Curtis L, Opsasnick L, Benavente JY, et al. Preliminary Results of MyCog, a Brief Assessment for the Detection of Cognitive Impairment in Primary Care. *Innovation in Aging*. 2020;16(4)(1):259–60.
16. Zelazo PD. The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nat Protoc*. 2006;1(1):297-301. doi:10.1038/nprot.2006.46
17. Dikmen SS, Bauer PJ, Weintraub S, et al. Measuring episodic memory across the lifespan: NIH Toolbox Picture Sequence Memory Test. *J Int Neuropsychol Soc*. Jul 2014;20(6):611-9. doi:10.1017/S1355617714000460
18. Curtis LM, Batio S, Benavente JY, et al. Pilot Testing of the MyCog Assessment: Rapid Detection of Cognitive Impairment in Everyday Clinical Settings. *Gerontol Geriatr Med*. Jan-Dec 2023;9:23337214231179895. doi:10.1177/23337214231179895
19. Shields RH, Kaat AJ, McKenzie FJ, et al. Validation of the NIH Toolbox Cognitive Battery in intellectual disability. *Neurology*. Mar 24 2020;94(12):e1229-e1240. doi:10.1212/WNL.0000000000009131

20. Young SR, McManus Dworak E, Byrne GJ, et al. Protocol for a construct and clinical validation study of MyCog Mobile: a remote smartphone-based cognitive screener for older adults. *BMJ Open*. Apr 2 2024;14(4):e083612. doi:10.1136/bmjopen-2023-083612
21. Baudic S, Barba GD, Thibaudet MC, Smagghe A, Remy P, Traykov L. Executive function deficits in early Alzheimer's disease and their relations with episodic memory. *Arch Clin Neuropsychol*. Jan 2006;21(1):15-21. doi:10.1016/j.acn.2005.07.002
22. Gallagher M, Koh MT. Episodic memory on the path to Alzheimer's disease. *Curr Opin Neurobiol*. Dec 2011;21(6):929-34. doi:10.1016/j.conb.2011.10.021
23. Seo EH, Kim H, Lee KH, Choo IH. Altered Executive Function in Pre-Mild Cognitive Impairment. *J Alzheimers Dis*. Oct 4 2016;54(3):933-940. doi:10.3233/JAD-160052
24. Sudo FK, Amado P, Alves GS, Laks J, Engelhardt E. A continuum of executive function deficits in early subcortical vascular cognitive impairment: A systematic review and meta-analysis. *Dement Neuropsychol*. Oct-Dec 2017;11(4):371-380. doi:10.1590/1980-57642016dn11-040006
25. Harvey PD, Cosentino S, Curiel R, et al. Performance-based and Observational Assessments in Clinical Trials Across the Alzheimer's Disease Spectrum. *Innov Clin Neurosci*. Jan-Feb 2017;14(1-2):30-39.
26. Marson D. Investigating Functional Impairment in Preclinical Alzheimer's Disease. *J Prev Alzheimers Dis*. Mar 2015;2(1):4-6. doi:10.14283/jpad.2015.44
27. Casaletto KB, Umlauf A, Beaumont J, et al. Demographically Corrected Normative Standards for the English Version of the NIH Toolbox Cognition Battery. *J Int Neuropsychol Soc*. May 2015;21(5):378-91. doi:10.1017/S1355617715000351
28. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. *Neurology*. Mar 12 2013;80(11 Suppl 3):S2-6. doi:10.1212/WNL.0b013e3182872e5f
29. Weintraub S, Bauer PJ, Zelazo PD, et al. I. NIH Toolbox Cognition Battery (CB): introduction and pediatric data. *Monogr Soc Res Child Dev*. Aug 2013;78(4):1-15. doi:10.1111/mono.12031
30. Weintraub S, Dikmen SS, Heaton RK, et al. The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: validation in an adult sample. *J Int Neuropsychol Soc*. Jul 2014;20(6):567-78. doi:10.1017/S1355617714000320
31. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 2013;49(4):764-766.
32. Woltz DJ, Was CA. Availability of related long-term memory during and after attention focus in working memory. *Mem Cognit*. Apr 2006;34(3):668-84. doi:10.3758/bf03193587
33. Vandierendonck A. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behav Res Methods*. Apr 2017;49(2):653-673. doi:10.3758/s13428-016-0721-5
34. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2023. https://www.R-project.org/
35. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(3):1-67. doi:10.18637/jss.v045.i03
36. Tibshirani R. Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1996;58(1):267-288.
37. Bujang MA, Adnan TH. Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis. *J Clin Diagn Res*. Oct 2016;10(10):YE01-YE06. doi:10.7860/JCDR/2016/18129.8744
38. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2 ed. John Wiley and Sons; 2000.
39. De Roeck EE, De Deyn PP, Dierckx E, Engelborghs S. Brief cognitive screening instruments for early detection of Alzheimer's disease: a systematic review. *Alzheimers Res Ther*. Feb 28 2019;11(1):21. doi:10.1186/s13195-019-0474-3

40. Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J Psychiatr Res*. Jan 2009;43(4):411-31. doi:10.1016/j.jpsychires.2008.04.014
41. National Plan to Address Alzheimer's Disease. ASPE. Accessed March 26, 2025. https://aspe.hhs.gov/collaborations-committees-advisory-groups/napa/napa-documents/napa-national-plan.
42. Pardo-Moreno T, Gonzalez-Acedo A, Rivas-Dominguez A, et al. Therapeutic Approach to Alzheimer's Disease: Current Treatments and New Perspectives. *Pharmaceutics*. May 24 2022;14(6)doi:10.3390/pharmaceutics14061117

Table 1. Demographic information for participant samples.

	Clinical Sample		Sample from General Population*
	Cognitively impaired	Non-impaired	
Total Sample N	70	81	97
Mean Age (SD)	79.8 (6.9)	76.0 (5.9)	74.0 (6.6)
[Min, Max]	[67, 94]	[67, 93]	[65, 89]
	n (%)		
Gender			
Female	39 (55.7)	55 (67.9)	63 (64.9)
Male	31 (44.3)	26 (32.1)	34 (35.1)
Racial Identity			
Asian	2 (2.9)	2 (2.5)	0 (0)
Black or African American	8 (11.4)	8 (9.9)	5 (5.2)
Middle Eastern or North African	0 (0)	0 (0)	0 (0)
Native Hawaiian or Pacific Islander	0 (0)	0 (0)	0 (0)
White	58 (82.9)	69 (85.2)	92 (94.8)
Other Identity	1 (1.4)	0 (0.0)	0 (0.0)
Prefer Not to Answer	1 (1.4)	2 (2.5)	0 (0.0)
Ethnic Identity			
Hispanic	1 (1.4)	2 (2.5)	2 (2.1)
Not Hispanic	69 (98.6)	79 (97.5)	95 (97.9)
Education			
Some High School	0 (0)	0 (0)	1 (1.0)
High School Diploma or GED	14 (20.0)	1 (1.2)	43 (44.3)
Some College or 2-Year Degree	14 (20.0)	13 (16.0)	23 (23.7)
4-Year College Degree	18 (25.7)	17 (21.0)	14 (14.4)
Graduate Degree	24 (34.3)	50 (61.7)	16 (16.5)

Table 2. Pertinent ICD codes for cognitive deficits and impairment.

	Criteria (diagnosis with either ICD code)	
	ICD-9	ICD-10
Dementia	290.X, 291.0-291.9, 292.82, 294.10, 294.11, 294.20, 294.21	F01.50, F01.51, F10.27, F10.97, F19.97, F19.27, F02.80, F02.81, F03.90, F03.913
MCI	331.83	G31. 84
Cognitive deficits	438.0	I69.91
Other memory loss	780.93	R41.3

ICD, International Classification of Diseases; MCI, mild cognitive impairment.

Table 3. Description and interpretation of MyCog DCCS scores and trial conditions.

Score	Description	Interpretation
Total score	Raw sum of correct responses across all trials	Overall performance
New MyCog DCCS	Composite score combining color and shape IRT scores	Overall performance
IRT scores (color, shape)	IRT-based ability estimates for color and shape dimensions	Ability on respective dimension
Rate-correct score (RCS)	Number of correct responses divided by mean response time (in seconds)	Speed-accuracy trade-off
Intraindividual standard deviation (ISD)	SD of response time for correct trials	Performance variability
Residualized ISD (RISD)	ISD after controlling for mean response time, practice effects, and age	Performance variability
Mean square successive difference (MSSD)	Average of squared differences between successive response times	Moment-to-moment variability
Linear integrated speed-accuracy score (LISAS)	Combined speed and accuracy measure weighted by SD	Performance error (in terms of combined speed and accuracy)
Proportion correct	Proportion of correct responses	Accuracy
Response time (median, sum)	Median or sum of response times for correct trials (in milliseconds)	Response delay
Difference scores (DIFF)	Differences between conditions (e.g., dominant vs non-dominant, switch vs repeat)	Performance error for harder conditions

Trial Condition	
Total	All 30 trials, encompassing switch, repeated, dominant, and non-dominant conditions
Switch	Trials that prompt matching with a different dimension than the proceeding trial (e.g., switches from shape to color or vice versa)
Non-Dominant	Trials that prompt matching with the color dimension

Table 4. Comparison of clinical samples and general population across all study variables (*M [SD], minimum – maximum*).

Variables	Clinical Sample		General Population	Non-impaired vs. Gen. Pop. (*P*)*
	Impaired	Non-impaired		
N	70	81	97	
Age	79.90 (6.97), 67.00–94.00	76.07 (5.91), 67.00–93.00	74.01 (6.57), 65.00–89.00	0.03
MyCog Picture Sequence Memory (PSM)				
Exact Match	2.60 (2.54), 0.00–12.00	6.24 (2.99), 2.00–12.00	6.19 (2.46), 1.00–12.00	0.90
Correct Adjacent Pairs	0.83 (1.90), 0.00–11.00	3.62 (3.39), 0.00–11.00	3.33 (2.66), 0.00–11.00	0.52
MyCog Dimensional Change Card Sorting (DCCS)				
Total Raw Score	24.23 (5.76), 9.00–30.00	28.68 (2.27), 20.00–30.00	28.54 (1.84), 21.00–30.00	0.65
IRT Composite Score	-0.91 (1.04), -3.28–0.82	0.21 (0.81), -2.21–0.82	0.00 (0.76), -2.04–0.82	0.08
Rate Correct Score (RCS)				
All Trials	0.53 (0.27), 0.04–1.08	0.92 (0.32), 0.19–1.96	1.02 (0.38), 0.17–2.16	0.06
Switch Trials	0.55 (0.29), 0.05–1.17	0.90 (0.31), 0.17–1.83	0.98 (0.37), 0.10–2.21	0.12
Non-Dominant Trials	0.68 (0.40), 0.03–1.62	1.01 (0.35), 0.11–2.26	1.10 (0.45), 0.00–2.90	0.13
Intraindividual Standard Deviation (ISD)				
All Trials	985.27 (553.42), 228.93–2656.00	532.19 (332.34), 160.13–1786.26	541.14 (361.44), 127.50–2060.75	0.86
Switch Trials	903.37 (650.87), 181.64–2814.34	449.22 (274.57), 156.62–1667.53	464.61 (320.79), 97.90–1819.19	0.73
Non-Dominant Trials	681.78 (699.97), 17.71–4130.55	339.24 (299.50), 72.82–2439.48	338.98 (327.84), 35.01–2296.03	0.99
Residualized Intraindividual Standard Deviation (RISD)				
All Trials	12.33 (6.96), 2.96–31.30	6.59 (4.00), 2.08–22.86	8.69 (5.80), 2.14–34.51	0.001
Switch Trials	12.13 (8.77), 2.64–39.08	6.15 (3.80), 1.63–23.09	8.83 (5.85), 1.95–31.43	< 0.0001
Non-Dominant Trials	11.29 (12.86), 1.06–76.08	5.33 (4.37), 0.56–32.97	8.07 (7.27), 0.13–43.02	< 0.0001
Mean Square Successive Difference				
All Trials	2260697.99 (2618029.18), 71283.85–13361662.20	679596.26 (1070400.53), 56814.65–6684108.62	650423.59 (858306.90), 22063.33–5226704.68	0.84
Switch Trials	2476427.29	567050.99	593389.34	0.85

	(3816165.79), 72272.21–16408903.55	(946739.14), 24124.17–6679232.02	(871047.61), 8802.30–5725661.27	
Non-Dominant Trials	2125515.13 (6604375.31), 627.30–51055212.11	327249.25 (889638.52), 2729.19–7503457.81	384365.65 (934532.87), 2443.11–7177565.43	0.68
Linear Integrated Speed-Accuracy Score (LISAS)				
All Trials	1853.58 (1497.41), 709.17–10000.00	1027.11 (347.05), 423.33–2035.93	986.04 (596.91), 412.18–4559.15	0.57
Switch Trials	1845.34 (1574.26), 689.84–10000.00	1074.62 (358.89), 401.30–2595.57	1029.65 (617.78), 420.61–5308.67	0.55
Non-Dominant Trials	1761.13 (1661.33), 602.14–10000.00	1020.59 (465.21), 384.96–4224.65	987.73 (997.17), 330.57–10000.00	0.77
Accuracy (Proportion Correct)				
All Trials	0.81 (0.19), 0.30–1.00	0.96 (0.08), 0.67–1.00	0.95 (0.06), 0.70–1.00	0.65
Switch Trials	0.80 (0.19), 0.31–1.00	0.95 (0.09), 0.62–1.00	0.93 (0.09), 0.46–1.00	0.20
Non-Dominant Trials	0.84 (0.22), 0.14–1.00	0.96 (0.10), 0.43–1.00	0.93 (0.14), 0.00–1.00	0.07
Median Response Time Correct Trials				
All Trials	1443.82 (699.61), 651.35–4261.34	951.28 (268.75), 423.33–1626.01	910.13 (516.42), 412.18–3762.33	0.50
Switch Trials	1419.16 (661.79), 676.89–4270.49	1012.58 (280.16), 401.30–1696.00	969.62 (566.34), 420.61–4514.23	0.51
Non-Dominant Trials	1371.15 (761.24), 602.14–5561.56	963.08 (263.50), 384.96–1617.83	861.45 (354.89), 330.57–2635.66	0.03
Rate Correct Difference Scores				
Switch – Repeated	0.00 (0.25), -0.89–0.68	-0.06 (0.25), -0.93–0.65	-0.09 (0.26), -0.74–0.59	0.38
Dominant – Non-Dominant	-0.17 (0.35), -1.03–1.17	-0.10 (0.26), -0.93–0.65	-0.09 (0.32), -0.89–0.90	0.79
Linear				
Switch – Repeated	-52.87 (1148.16), -5783.25–6427.27	84.89 (285.71), -1137.29–791.98	112.70 (325.55), -1195.13–2283.76	0.55
Dominant – Non-Dominant	-188.95 (1218.93), -5437.27–6325.49	-20.47 (399.99), -993.23–2411.29	7.56 (701.89), -1477.79–6237.67	0.74
Median Reaction Time Differences				
Switch – Repeated	-111.65 (537.52), -2102.88–1163.32	91.06 (195.75), -321.54–765.90	90.49 (240.13), -792.51–1489.32	0.99
Dominant – Non-Dominant	-163.54 (646.03), -2200.79–2380.22	6.60 (215.20), -543.99–806.53	-28.66 (266.24), -1422.54–599.46	0.33

Accuracy (Proportion Correct) Differences				
Switch – Repeated	-0.01 (0.17), -0.46–0.57	-0.01 (0.09), -0.31–0.41	-0.03 (0.10), -0.54–0.22	0.11
Dominant – Non-Dominant	0.05 (0.28), -0.86–0.87	0.01 (0.09), -0.31–0.30	-0.03 (0.15), -1.00–0.26	0.05

*All comparisons (far right column) were derived from two-sample Welch's t-tests.

Table 5. Comparison of diagnostic accuracy across nine models.

Model	ROC AUC	Sensitivity	Specificity	Predictors (N)
Ridge	0.90	0.84	0.88	25
LASSO	0.90	0.88	0.85	23
Elastic Net	0.89	0.83	0.88	27
Stepwise	0.87	0.82	0.85	13
Bayes Ridge	0.90	0.64	0.90	29
Bayes LASSO	0.89	0.65	0.92	21
Bayes Elastic Net	0.90	0.66	0.90	25
Bayes Stepwise	0.90	0.64	0.90	29
Random Forest	0.87	0.82	0.82	3