Received: 13 October 2021 | Revised: 20 April 2022 | Accepted: 8 May 2022 | Published online: 10 July 2022

DOI: 10.1002/dad2.12325

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Mayo normative studies: A conditional normative model for longitudinal change on the Auditory Verbal Learning Test and preliminary validation in preclinical Alzheimer's disease

Eva C. Alden[1] | Emily S. Lundt[2] | Erin L. Twohy[2] | Teresa J. Christianson[2] | Walter K. Kremers[2] | Mary M. Machulda[1] | Clifford R. Jack Jr.[3] | David S. Knopman[4] | Michelle M. Mielke[4,5] | Ronald C. Petersen[4] | Nikki H. Stricker[1]

[1]Division of Neurocognitive Disorders, Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, USA

[2]Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, USA

[3]Department of Radiology, Mayo Clinic, Rochester, Minnesota, USA

[4]Department of Neurology, Mayo Clinic, Rochester, Minnesota, USA

[5]Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, USA

Correspondence
Nikki H. Stricker, PhD, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.
E-mail: stricker.nikki@mayo.edu

Funding information
Rochester Epidemiology Project, Grant/Award Number: R01 AG034676; National Institutes of Health, Grant/Award Numbers: P50 AG016574, P30 AG062677, U01 AG006786, R37 AG011378, R01 AG041851, RF1 AG55151; Alzheimer's Association, Grant/Award Numbers: AARG-17-531322, Zenith Award

Abstract

Introduction: The aim of this study was to develop a conditional normative model for Rey's Auditory Verbal Learning Test (AVLT) that accounts for practice effects.

Methods: In our normative sample, robust conditional norms were derived from 1001 cognitively unimpaired (CU) adults ages 50 to 89 who completed the AVLT up to eight times. Linear mixed-effects models adjusted for baseline performance, prior test exposures, time, demographics, and interaction terms. In our preliminary validation, mean performance on conditional and typical normative scores across two to four completed follow-up tests in preclinical Alzheimer's disease participants at baseline with positive amyloid and tau positron emission ($n = 27$ CU amyloid [A]+tau[T]+) was compared to biomarker negative individuals ($n = 269$ CU A−T−).

Results: AVLT performance using typical norms did not differ across A+T+ and A−T− groups. Conditional norms z-scores were lower in the A+T+ relative to the A−T− group for 30-minute recall ($P = .033$) and sum of trials ($P = .030$).

Discussion: Conditional normative methods that account for practice effects show promise for identifying longitudinal cognitive decline.

KEYWORDS
Alzheimer's disease, amyloid, biomarker, memory, mild cognitive impairment, neuropsychology, practice effects, reliable change index (RCI), Rey Auditory Verbal Learning Test, robust normative data, standardized regression-based change scores (SRB), tau, transitional cognitive decline, validity

1 | INTRODUCTION

Distinguishing cognitive decline from normal aging in older adults based on a single neuropsychological assessment is challenging. It is increasingly recognized that serial testing may provide a better prognostic indicator for future conversion to mild cognitive impairment (MCI) and dementia.[1] Transitional cognitive decline in cognitively unimpaired (CU) individuals on the Alzheimer's disease (AD) continuum represents a decline from a previous level of functioning, with performance still in the non-impaired range.[2] Transitional

cognitive decline can be based on subjective report by the individual or informant, or on subtle decline on longitudinal cognitive testing.[2]

Evaluating change over time is confounded by many variables including practice effects, which obscure diagnostic accuracy and may negatively impact clinical decision -making.[3,4] Though evidence is mixed and can vary by follow-up interval,[5,6] prior studies show practice effects occur even in individuals with positive AD biomarkers and MCI.[7–9] One approach to account for practice effects is through conditional normative models, defined as models conditioned or based upon prior test exposure and performance.[10] A conditional normative approach can be contrasted with typical unconditional norms based only on cross-sectional data, though the normative score produced by conditional norms is interpreted similarly.[10] Conditional normative models use longitudinal data and evaluate the degree to which an observed follow-up score differs from expected performance based not just on demographic factors like typical norms, but also on baseline performance, time since baseline, and number of subsequent test exposures.[10,11] A conditional model is like standardized regression-based change (SRB) approaches, although the latter is frequently limited to two test sessions.[3,12] Conditional norms and SRB approaches both differ from simpler reliable-change index (RCI) scores, which indicate whether observed change is outside that expected from error alone[13] or error plus mean practice effect[14] in a binary fashion.

Our first aim was to develop a conditional normative prediction model using a robust sample of older adults who remain CU over time. The second aim was to assess clinical utility of the model by deriving conditional norms in a validation sample of individuals with preclinical AD at baseline based on positive amyloid and tau imaging biomarkers and then comparing scores obtained using the conditional normative model to typical normative data.[15] The primary hypothesis was that individuals with preclinical AD would have lower conditional normative z-scores than CU individuals without positive AD biomarkers, and that typical norms would be less sensitive to transitional cognitive decline than robust conditional norms.

RESEARCH IN CONTEXT

1. **Systematic Review**: The authors reviewed the literature through primary source materials. Serial cognitive testing is an important component of differential diagnosis and early detection of Alzheimer's disease (AD). There are no current clinical practice standards for evaluating cognitive change over multiple longitudinal follow-up visits. Moreover, there is limited normative data to account for practice effects across multiple time points on the Auditory Verbal Learning Test (AVLT).
2. **Interpretation**: Conditional normative models for the AVLT that account for demographic variables, baseline performance, and repeat test exposure are more sensitive to preclinical AD than conventional AVLT norms that do not account for practice effects.
3. **Future Directions**: These AVLT conditional normative models may help identify transitional cognitive decline in AD and may also be useful in other clinical populations in which cognitive change may be expected.

2 | METHODS

2.1 | Participants

The overall study design is shown in Figure 1. To address each aim, participants were selected from the Mayo Clinic Study of Aging (MCSA), an ongoing population-based study of cognitive aging among Olmsted County, Minnesota residents. Participants are randomly sampled using the Rochester Epidemiology Project medical records-linkage system. Enrollment follows an age- and sex-stratified random sampling design to ensure equal sex representation in each 10-year age strata; the complete study design and sampling procedures were previously published.[16] The study protocols were approved by the Mayo Clinic and Olmsted Medical Center Institutional Review Boards. All participants provided written informed consent.

Study visits included a physician examination, study coordinator interview, and neuropsychological testing by a trained psychometrist.[16] The physician examination included a medical history review, neurological examination, and the administration of the Short Test of Mental Status.[17] The study coordinator collected demographic information, medical history, and completed the Clinical Dementia Rating scale with the participant and informant.[18] See Roberts et al.[16] for details about the neuropsychological battery.

2.2 | Auditory Verbal Learning Test

As part of the neuropsychological test battery, all participants completed the Rey's Auditory Verbal Learning Test (AVLT), a widely used word list memory test.[19] Participants are read a list of 15 words (list A) and asked to recall as many as possible over five learning trials, after one intervening list and again 30 minutes later. This is followed by a written recognition test.[20] Participants were administered the same version (Form A) at all visits.

2.3 | Part 1: Developing robust conditional normative data (training sample)

2.3.1 | Sample

The robust sample consisted of 1001 CU adults ages 50 to 89 at baseline. All participants were test naïve at the time of their baseline AVLT. Participants completed as few as four and up to eight tests occurring

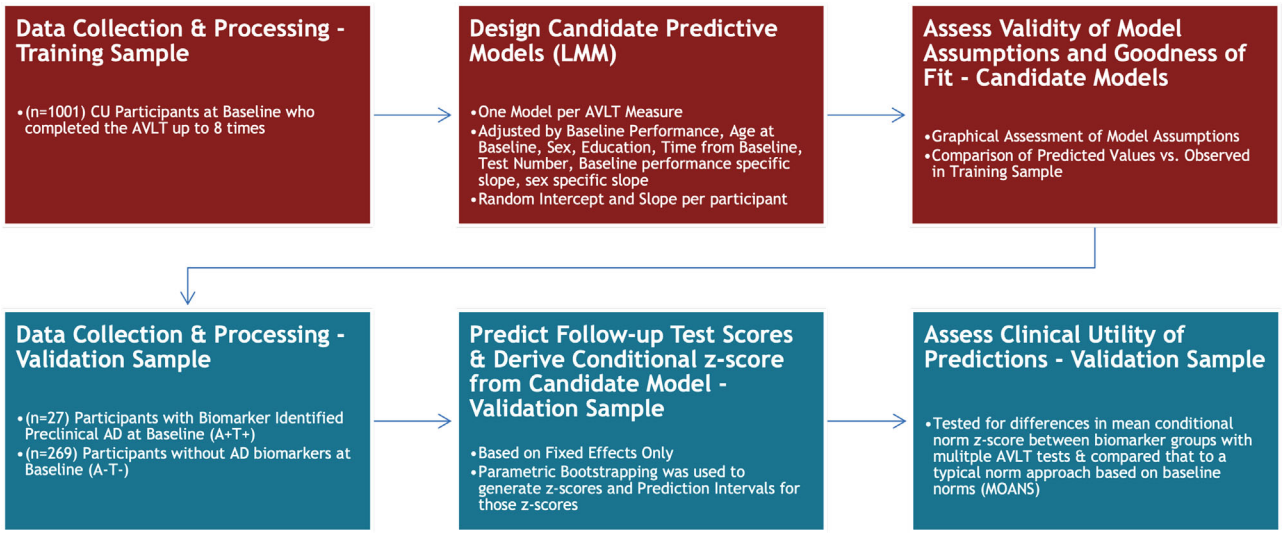


FIGURE 1 **Flowchart showing the overall study design and methods.** Part 1 (top row, red boxes) addresses study aim 1 to develop robust conditional normative data; this is our training sample. Part 2 (bottom row, blue boxes) addresses study aim 2 to assess clinical utility for preclinical Alzheimer's disease; this is our validation sample. A, amyloid; AD, Alzheimer's disease; AVLT, Auditory Verbal Learning Test; CU, cognitively unimpaired; LMM, linear mixed model; MOANS, Mayo's Older Americans Normative Studies; T, tau

approximately every 15 months. Fewer than 25% of participants had completed ≥9, thus tests 9+ were omitted.

To avoid circularity or incorporation bias, the neuropsychologist's diagnostic impression was not considered for inclusion criteria. Instead, a diagnosis of MCI or dementia was based on a consensus agreement between the interviewing study coordinator and examining physician.[21,22] To create a robust sample, participants were required to be CU at all available follow-ups. Other exclusion criteria included terminal illness or hospice.

2.3.2 | Analysis

Models were generated using a linear mixed effects (LME) regression-based approach that allows for multiple predictors and repeated measurements at multiple time points. Furthermore, prior work indicates similar regression-based approaches are more accurate predictors of follow-up performance trajectory, or expected change over time, compared to other methods.[23] Three LME models were fit, one for each AVLT measure including trials 1 through 5 total, 30-minute delayed recall, and sum of trials (trials 1–5 total + short delay recall + long delay recall). Within-subject variability in the model was specified by a random slope and intercept per participant (i.e., random effects). Predictions from these models were used to generate normative scores conditioned on practice effects for each AVLT measure of interest. Models were set up to predict from test number 2 onward by adjusting for the baseline value of the AVLT measure, age at baseline, sex, and education. Prior work has suggested these demographic factors have a significant impact on cognitive change over time,[24] and that baseline test performance is the strongest predictor of follow-up performance.[25] To account for practice effects, test number (2, 3, 4, 5–8) and time since baseline were also included as factors. Decade of age centered at 75 was used (age at baseline in years − 75)/10. To assess whether rates of decline over time differed by demographic factors, interaction terms were also evaluated, including sex by time, education by time, age at baseline by time, and baseline performance by time. Semi-sequential log likelihood ratio tests of education by time and age by time interactions were not significant at the $\alpha = 0.05$ level and were excluded from the models. Graphical representation and goodness of fit measures were used to assess assumptions of normality and homogeneity of variance.

After conditional models were finalized, prediction intervals for an AVLT score for each set of covariates in the robust sample were generated using demographic variables, baseline performance, time since baseline, and test number. The predictInterval R function from the merTools package was used to generate predicted mean scores at each follow-up. This function generates prediction intervals that account for residual variation, uncertainty of fixed effects, and uncertainty of variance parameters of the random effects. To have a normative calculation tool that can be applied to individuals not included in this study, random effects were not used to generate predictions. Prediction interval inclusion probabilities were set at 95% and taken as 2.5th and 97.5th percentiles of 1000 simulations generated by predictInterval. The predicted test score at each follow-up was taken as the median from the same simulations. To convert to a z-score scale more precisely, multiple simulations were run. Predicted test score and interval were converted to z-scores using the median of 5000 simulated standard deviations. Conditional normalized scores on the z-score scale for each follow-up test were then derived using the difference between the observed and predicted values, and corresponding standard deviation for that value.

2.4 | Part 2: Preliminary validation of conditional normative data in preclinical AD (validation sample)

2.4.1 | Sample

Conditional norms were tested on a sample of preclinical AD participants with positive amyloid and tau imaging (CU amyloid [A]+ tau [T]+[29]) who were expected to have a higher rate of intra-individual cognitive decline relative to CU A−T− individuals.[26] All participants were required to be CU at baseline, were AVLT naïve at baseline, and >50 years of age. Participants were required to have at least one completed follow-up test and most (83.4%) had three follow-up tests available. Two participants (7%) in the A+T+ group were diagnosed with MCI during at least one of their follow-up visits (2–4), whereas <1% of A−T− participants were diagnosed with MCI. For validation, diagnosis followed typical MCSA procedures,[16] and was based on consensus among the neuropsychologist, study coordinator, and examining physician.

Positron emission tomography (PET)/computed tomography (CT) was performed using Pittsburgh compound B for amyloid and flortaucipir for tau, and individuals were considered A+ using a threshold of standardized uptake value ratio (SUVR) ≥ 1.48 (centiloid 22[27]) and T+ (using a temporal lobe meta region of interest [ROI]) at a threshold of SUVR ≥ 1.25.[28–30] Biomarker subgroups were based on the recently proposed research framework for a biological diagnosis of AD[2] and included participants with negative amyloid and tau PET biomarkers (A−T−) and participants with preclinical AD (A+T+).

2.4.2 | Analysis

Predictions and intervals of the mean score were generated by predictInterval for the validation sample and excluded variation from the random effects. Conditional z-scores were generated for individuals in the validation sample. To compare mean performance across CU A−T− and CU A+T+ groups, *t*-tests ($\alpha = 0.05$) were performed using both the conditional normative z-scores and typical age-corrected Mayo's Older American Normative Studies (MOANS) scores. To facilitate comparison of the two methods, the typical age-corrected MOANS scaled scores were converted to z-scores. Study hypotheses focused on results that averaged across follow-up visits 2 through 4.

3 | RESULTS

3.1 | Part 1: Robust conditional normative data

3.1.1 | Sample characteristics

Participants were predominantly White (see Table 1). The mean (standard deviation [SD]) age of the sample was 71.54 (7.24), and mean level of education was 15.08 (2.53). AVLT performance by follow-up test number is shown in Table 2. Figure 2 illustrates performance trajectories for test numbers 2 through 8 in males and females. Performance is shown separately for individuals at the 25th (Figure 1A), 50th (Figure 1B), and 75th (Figure 1C) percentiles.

TABLE 1 Robust normative sample characteristics

	No. of participants at baseline (%) (*N* = 1001)
Age, years, mean (range)	72 (51–89)
50–59	71 (7.1%)
60–69	268 (26.8%)
70–79	535 (53.4%)
80–89	127 (12.7%)
Sex (male)	496 (49%)
Education, years	
8–12	258 (26%)
13–15	247 (25%)
16	213 (21%)
17–20	283 (27%)
Race	
Non-White	12 (1%)
White	989 (99%)

3.1.2 | Model characteristics and distributional properties

Table 3 shows fixed effects estimates for predictors included in the model. Evaluation of model estimates indicates baseline performance, sex, and age at baseline are the strongest predictors of longitudinal performance (all $P < .001$). Non-linear practice effects were allowed within the model. There was little evidence of additional practice effect beyond five test exposures; therefore, the effect of test number 5 and beyond was collapsed into one model term. Essentially, the effect of practice with reference to test number 2 was captured by a set of four categorical terms within the model (test number 3, 4, 5+). The final interaction terms included in the model were baseline performance by time and sex by time, as prior work indicates baseline performance and sex modify cognitive performance and impact rates of cognitive change over time.[11,25] The interactions education by time and age at baseline by time did not meet criteria for inclusion in final models. Addition of non-linear effects for remaining variables, including age, did not improve overall model fit, and thus only linear age is represented in the model. Years of education, time from baseline, and test number were also significant predictors, though statistical significance varied across AVLT variables. Due to the limited number of participants with fewer than 11 years of education (*N* = 9), all individuals with 10 or fewer years of education were included as education = 11, with the remaining participants' level of education increasing linearly from 12 years on. As a result, the model may be less applicable to individuals with lower levels of education.

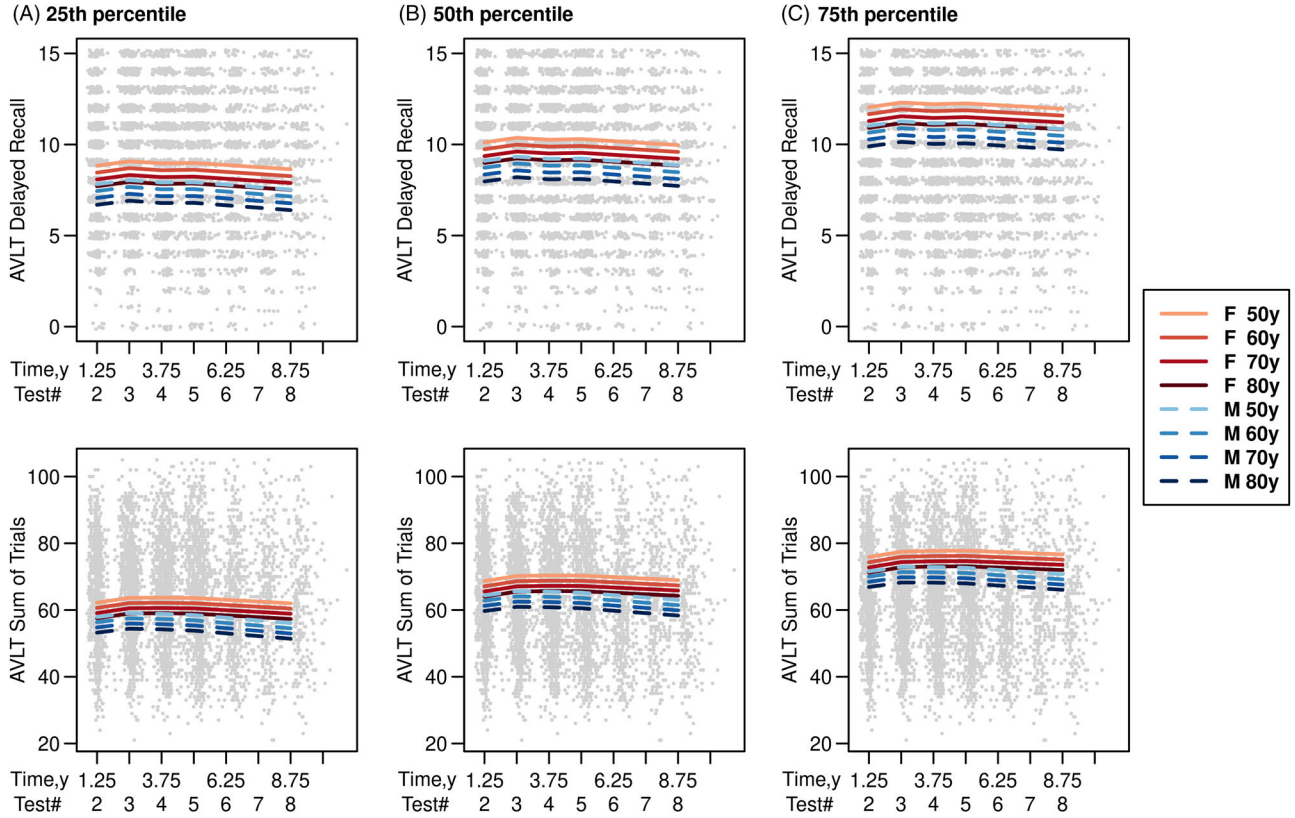


FIGURE 2 **Robust normative sample Auditory Verbal Learning Test (AVLT) raw scores.** Raw score estimates over test numbers 2 through 8 are displayed for males and females for the AVLT delayed recall and sum of trials. Test number 1 (baseline) is not depicted as it is one of the model predictors and not a model output. A, Performance trajectories at the 25th percentile, (B) at the 50th percentile, and (C) at the 75th percentile

3.2 | Part 2: Preliminary validation of conditional normative data in preclinical AD

3.2.1 | Sample characteristics

The preliminary validation sample consisted of 269 A−T− and 27 A+T+ participants who were CU at baseline (see Table 4). There is some overlap among participants in the validation and robust samples (7.69% of the total robust normative sample are in both, specifically 62 CU A−T− and 15 CU A+T+). Given the size of the robust normative sample, this overlap is unlikely to significantly impact results; further, no participant-specific random effects terms were used to derive conditional normative scores (e.g., were not included in predictions and z-score generation). The A+T+ group was older than the A−T− group (see Table 4). Groups were comparable on sex and level of education. Baseline AVLT performance did not differ across A+T+ and A−T− groups. PET imaging was obtained, on average, 2.68 (1.26) years after baseline AVLT.

3.2.2 | Group comparisons

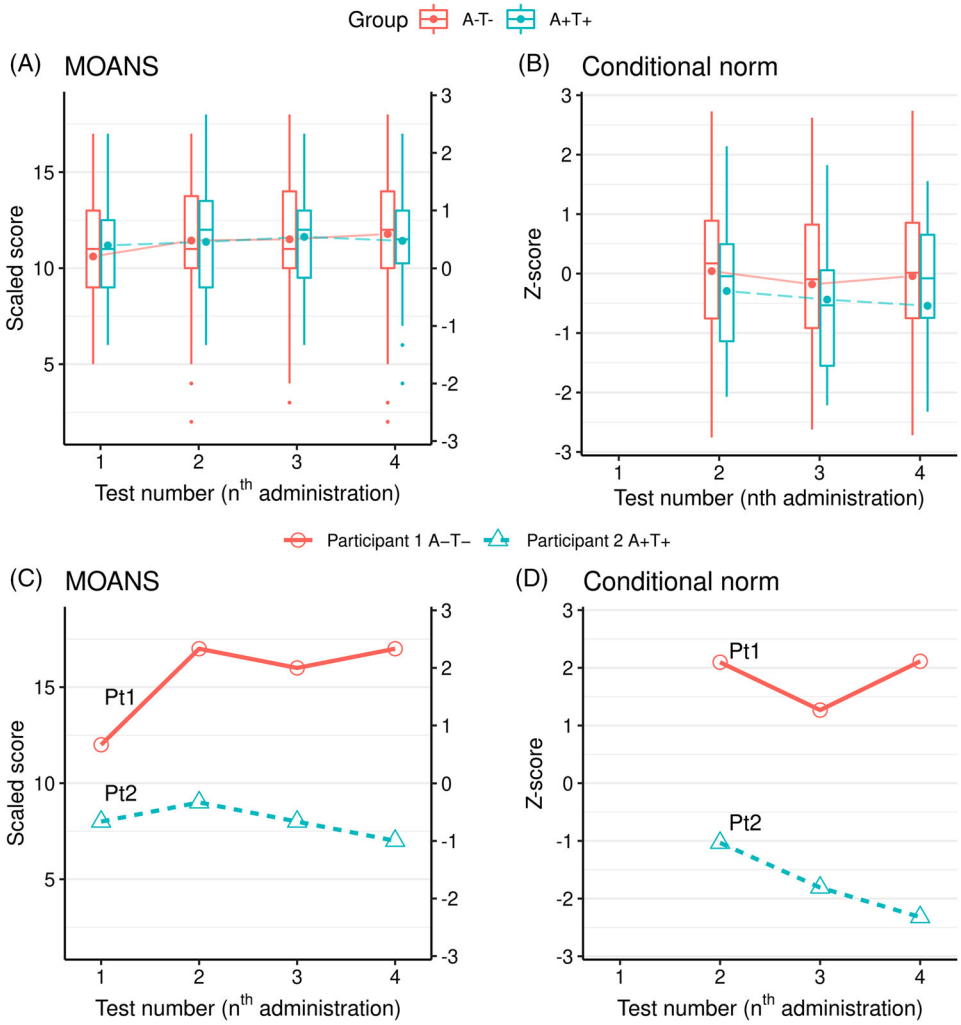
Across two to four follow-up tests, the typical MOANS delayed recall mean z-score did not differ between the A−T− and A+T+ groups (P = .791; see Table 4). However, the conditional normative z-scores that account for practice effects and typical trajectories of AVLT performance were significantly lower in the A+T+ group relative to the A−T− group for 30-minute delayed recall (P = .033) and sum of trials (P = .030) when averaging across all follow-up tests. Examining each separately, there were no group differences at test numbers 2 and 3. At test number 4, there were significant group differences on AVLT trials 1 through 5 (P = .044) and sum of trials (P = .017).

3.2.3 | Illustration of clinical utility

Figure 3 illustrates performance trajectories for the preliminary validation sample, as well as two individual participants.

4 | DISCUSSION

This study highlights the importance of developing and using robust conditional normative data that can be applied to longitudinal neuropsychological assessments. While most prior normative studies that are designed to evaluate significant change over time for the AVLT provide normative models for a single follow-up assessment, this study presents conditional normative data for up to seven follow-up



(E)

	Delayed recall raw				MOANS scaled score				Conditional norm z-score		
Test#	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**2**	**3**	**4**
Pt1 A–T–	11	15	14	15	12	17	16	17	2.1	1.3	2.1
Pt2 A+T+	5	6	5	4	8	9	8	7	-1	-1.8	-2.3

FIGURE 3 **Comparative performance trajectories in validation sample participants for 30-minute recall.** Example of performance trajectories across four tests for AVLT 30-minute delayed recall. A,B, Performance trajectories for the entire validation sample depicted by biomarker group using the traditional cross-sectional MOANS norms applied at each follow-up (A) and the conditional normative model at second through fourth follow-up (B). C,D, Two individual participants, with corresponding raw and numeric normative scores reported in (E). The first participant is a 64-year-old female with 16 years of education who is A–T– and remains CU across all follow-up tests (orange solid line). Conditional norms suggest this individual demonstrated a higher than typical improvement in performance at test number 2 and the trajectory remained above average at test numbers 3 and 4. The second participant is a 65-year-old female with 14 years of education who is A+T+ and is CU at tests 1 and 2, but is diagnosed with MCI at tests 3 and 4 (blue dashed line) per consensus conference. Despite a subtle improvement in raw score at test number 2, the conditional norm suggests that this individual's ability to benefit from practice was subtly low relative to similar peers. At test number 3, despite an identical raw score as test number 1, the conditional norms indicate this individual's trajectory over time is deviating from what is typical and this corresponds with the consensus diagnosis of MCI at this test number. By test number 4, performance is clearly abnormal per conditional norms that consider the number of test exposures (failure to benefit from practice), baseline performance, age, education, sex, and time since baseline despite a MOANS score that is only subtly low and in a range many label as within normal limits (equivalent to a z of −1) and a raw score only 1 point lower than the baseline test number. A, amyloid; AVLT, Auditory Verbal Learning Test; CU, cognitively unimpaired; MCI, mild cognitive impairment; MOANS, Mayo's Older Americans Normative Studies; T, tau

TABLE 2 Robust normative sample AVLT performance characteristics and practice effects at each test number. Reported statistics are of the form mean (SD) or estimate (95% CI) unless otherwise specified

	1 (*N* = 1001)	2 (*N* = 991)	3 (*N* = 988)	4 (*N* = 994)	5 (*N* = 986)	6 (*N* = 493)	7 (*N* = 343)	8 (*N* = 250)
Years since baseline		1.26 (0.16)	2.58 (0.23)	3.89 (0.29)	5.18 (0.34)	6.64 (0.37)	7.97 (0.35)	9.21 (0.36)
AVLT trials 1–5	43.46 (8.89)	45.57 (9.40)	46.56 (9.61)	46.71(10.26)	46.54 (10.74)	44.68 (10.72)	45.15 (10.83)	45.10 (11.02)
AVLT delayed recall	8.42 (3.24)	9.12 (3.17)	9.39 (3.25)	9.26 (3.35)	9.30 (3.46)	8.72 (3.49)	8.79 (3.44)	8.90 (3.53)
AVLT sum of trials	60.62 (14.23)	64.02 (14.60)	65.54 (14.93)	65.47 (15.79)	65.26 (16.54)	62.35 (16.49)	62.91 (16.47)	63.08 (16.88)
Cohen's d (relative to prior)								
AVLT trials 1–5		0.23 (0.19, 0.28)	0.09 (0.05, 0.14)	0.02 (−0.02, 0.06)	−0.02 (−0.05, 0.02)	−0.01 (−0.07, 0.04)	−0.03 (−0.10, 0.04)	−0.06 (−0.14, 0.01)
AVLT delayed recall		0.23 (0.18, 0.27)	0.07 (0.03, 0.12)	−0.03 (−0.07, 0.01)	0.01 (−0.03, 0.05)	−0.05 (−0.11, 0.01)	−0.02 (−0.10, 0.05)	−0.03 (−0.11, 0.04)
AVLT sum of trials		0.24 (0.20, 0.28)	0.09 (0.05, 0.13)	0 (−0.04, 0.04)	−0.01 (−0.05, 0.02)	−0.02 (−0.07, 0.03)	−0.03 (−0.09, 0.03)	−0.05 (−0.12, 0.01)
Cohen's d (relative to 1st)								
AVLT Trials 1–5		0.23 (0.19, 0.28)	0.33 (0.28, 0.38)	0.33 (0.28, 0.38)	0.30 (0.25, 0.35)	0.31 (0.23, 0.39)	0.35 (0.26, 0.44)	0.30 (0.20, 0.40)
AVLT delayed recall		0.23 (0.18, 0.27)	0.30 (0.25, 0.34)	0.25 (0.20, 0.30)	0.26 (0.21, 0.30)	0.25 (0.18, 0.33)	0.26 (0.17, 0.35)	0.26 (0.16, 0.36)
AVLT sum of trials		0.24 (0.20, 0.28)	0.33 (0.29, 0.38)	0.32 (0.27, 0.36)	0.29 (0.24, 0.34)	0.30 (0.23, 0.38)	0.33 (0.25, 0.42)	0.30 (0.21, 0.39)

Notes: Average raw scores are shown across the entire training sample not stratified by age, sex, and education. These data are for illustration purposes only, and not intended to be used for calculating reliable or normative change scores. The effect size measure Cohen's d with pooled variance is reported in two ways: relative to prior (2 to 1, 3 to 2, 4 to 3, …) and relative to first baseline AVLT test (2 to 1, 3 to 1, 4 to 1, …).
Abbreviations: AVLT, Auditory Verbal Learning Test; CI, confidence interval; SD, standard deviation.

assessments. This flexible model can be applied after at least two and up to seven follow-up assessments. By using a linear mixed-effects normative model, we not only include baseline performance and number of test exposures to account for practice effects, but also relevant demographic factors that can impact trajectories. In a preliminary validation sample, results show that the z-scores derived from the conditional norms model were more sensitive to transitional cognitive decline in preclinical AD than traditional baseline MOANS norms applied repeatedly across test sessions. Overall, these findings suggest that conditional norms may be a powerful tool for clinicians and researchers for detecting and monitoring early cognitive changes in AD, and likely for determining trajectories of change over time in other neurological disorders as well. We used the publicly available Shiny package from RStudio[31] to develop a web-based application that provides a calculation tool to allow others to apply these conditional norms for non-commercial clinical and research use, which is available at https://rtools.mayo.edu/avlt_conditional_norms/.

At present there is no established standard for defining when a clinically meaningful change in test performance has occurred. Unlike reliable change indices that focus on whether the amount of change is significantly different from chance,[32] the conditional model indicates the magnitude of observed performance deviation from the expected longitudinal cognitive trajectory of a CU individual at a specified follow-up in the form of a continuous z-score. This model cannot be used for the first exposure to the AVLT (baseline assessment). Our updated conventional norms for a single time point assessment are recommended for a baseline administration of the AVLT.[33] These conditional norms can be applied for any follow-up (up to 7) when baseline and relevant demographic data are available. Applying this model can help identify when individuals do not show the expected benefit from practice on follow-up testing.[34,35] This provides an easily interpretable metric that does not rely on a change over any single time interval and can help determine whether and to what extent an individual's trajectory is atypical. This approach can be used in the same way as conventional normative scores and helps remove subjective judgement about whether a predefined or specific amount of change is clinically significant.

Previous work demonstrated that conditional models outperform unconditional models (i.e., typical cross-sectional norms) that do not account for practice in their ability to identify individuals with cognitive impairment.[10] While similar, our study differs from this prior work due to the robust nature of our sample, which was comprised of individuals who remained CU for the entire length of time they were followed in the MCSA, even beyond the number of AVLT administrations included in the normative models. Many prior studies that provide methods for defining clinically meaningful cognitive change based on convenience samples may unintentionally include asymptomatic individuals who later develop incident MCI.[36] Although such samples may be representative of the broader population,[37] research suggests robust normative samples decrease performance variability, raise mean performance estimates, and increase sensitivity to MCI and dementia.[38,39] Therefore, robust normative data can likely improve diagnostic

TABLE 3 Linear mixed effects regression model parameter estimates

	AVLT trials 1–5		AVLT delayed recall		AVLT sum of trials	
	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value
Fixed effects						
Baseline performance	0.662 (0.613, 0.710)	<.001	0.636 (0.591, 0.681)	<.001	0.711 (0.668, 0.754)	<.001
Age at baseline	−1.261 (−1.744, −0.778)	<.001	−0.376 (−0.534, −0.219)	<.001	−1.559 (−2.259, −0.858)	<.001
Sex (1 = male, 0 = female)	−2.969 (−3.795, −2.142)	<.001	−0.999 (−1.284, −0.714)	<.001	−4.026 (−5.205, −2.848)	<.001
Education	0.361 (0.227, 0.496)	<.001	0.061 (0.016, 0.105)	.007	0.428 (0.233, 0.623)	<.001
Time from baseline	−0.441 (−0.892, 0.010)	.06	−0.112 (−0.210, −0.015)	.02	−0.782 (−1.357, −0.207)	.008
AVLT number 3	1.289 (0.823, 1.755)	<.001	0.359 (0.199, 0.519)	<.001	1.986 (1.344, 2.629)	<.001
AVLT number 4	1.871 (1.266, 2.477)	<.001	0.362 (0.155. 0.570)	<.001	2.595 (1.756, 3.434)	<.001
AVLT number 5+	2.119 (1.230, 3.008)	<.001	0.506 (0.202, 0.810)	.001	3.019 (1.787, 4.252)	<.001
Baseline performance × time	0.006 (−0.004, 0.015)	.25	0.003 (−0.006, 0.012)	.48	0.007 (−0.001, 0.015)	.11
Sex × time	−0.202 (−0.366, −0.038)	.02	−0.013 (−0.068, 0.042)	.64	−0.217 (−0.451, 0.017)	.07
Random effects						
SD of intercept	3.87		1.42		5.76	
SD of slope	0.61		0.20		0.94	
SD of residual	4.63		1.59		6.35	
Intercept × slope correlation	0.28		0.02		0.23	

Notes: Baseline performance is the baseline score for each psychometric test of interest; Age is in decades from 75 years represented in the model as (Age – 75) / 10; sex is 1 for males and 0 is for females; education is in total years with all individuals 11 or less coded as 11; AVLT number 3, 4, 5+ are dichotomous variables to represent the effect each follow-up test relative to test number 2 (the reference group).
Abbreviations: AVLT, Auditory Verbal Learning Test; CI, confidence interval; SD, standard deviation.

TABLE 4 Characteristics of validation sample by group with means and standard deviations of AVLT performance combined across test numbers 2–4, and separately beginning at baseline

	CU A−T− (*n* = 269) M (SD)	CU A+T+ (n = 27) M (SD)	*P*-value
Mean age at baseline (SD)	62 (8)	69 (4)	<.001
Age range	50-87	61-76	
Mean level of education (SD)	15 (2)	15 (3)	.349
Mean time of PET from baseline in years (SD)	2.64 (1.26)	3.02 (1.24)	.136
Sex (% male)	151 (56%)	16 (59%)	.755
N with follow-up data available by cycle			
AVLT number 2	258	27	
AVLT number 3	249	27	
AVLT number 4	221	26	
Baseline AVLT performance			
30-minute Recall observed raw score	9.20 (3.22)	8.63 (3.27)	.385
MOANS 30-minute Recall scaled score	10.61 (2.59)	11.19 (2.80)	.276
AVLT performance across test numbers 2–4			
30-minute recall observed raw score	10.02 (3.39)	8.69 (3.52)	.002
MOANS 30-minute recall z*	0.52 (0.94)	0.49 (1.04)	.791
30-minute recall conditional norm z	−0.06 (1.45)	−0.42 (1.45)	.033
Trials 1–5 conditional norm z	−0.10 (1.39)	−0.34 (1.47)	.143
Sum of trials conditional norm z	−0.11 (1.45)	−0.49 (1.59)	.030
AVLT performance at test number 2			
30-minute recall observed raw score	9.98 (3.32)	8.81 (3.66)	.086
MOANS 30-minute recall z*	0.48 (0.92)	0.46 (1.10)	.906
30-minute recall conditional norm z	0.04 (1.38)	−0.29 (1.49)	.235
Trials 1–5 conditional norm 5	−0.06 (1.34)	−0.39 (1.34)	.223
Sum of trials conditional norm z	−0.06 (1.38)	−0.45 (1.51)	.170
AVLT performance at test number 3			
30-minute recall observed raw score	10.00 (3.42)	8.81 (3.43)	.088
MOANS 30-minute recall z*	0.50 (0.95)	0.54 (1.05)	.832
30-minute recall conditional norm z	−0.18 (1.47)	−0.44 (1.40)	.387
Trials 1–5 conditional norm z	−0.17 (1.41)	0.02 (1.64)	.516
Sum of trials conditional norm z	−0.21 (1.48)	−0.21 (1.69)	.993
AVLT performance at test number 4			
30-minute recall observed raw score	10.09 (3.46)	8.42 (3.58)	.022
MOANS 30-minute recall z*	0.59 (0.96)	0.47 (1.01)	.555
30-minute recall conditional norm z	−0.04 (1.49)	−0.54 (1.50)	.106
Trials 1–5 conditional norm z	−0.07 (1.42)	−0.67 (1.38)	.044
Sum of trials conditional norm z	−0.07 (1.50)	−0.82 (1.56)	.017

Notes: *P*-values represent *t*-test for mean comparisons or Pearson's Chi-squared test for frequency comparisons. z, z-score; sum of trials, total of trials 1–5 + 6 + 30-minute recall.

*MOANS scaled scores were converted to z-scores for ease of interpretation.

Abbreviations: A, amyloid; AVLT, Auditory Verbal Learning Test; CU, cognitively unimpaired; MOANS, Mayo's Older Americans Normative Studies; SD, standard deviation; T, tau.

accuracy relative to use of traditional norms that may unintentionally underestimate rates of cognitive impairment.

Sample composition may also impact the degree of practice effects observed. Prior results from our group suggest practice effects vary depending on the presence or absence of neurodegeneration,[9] which may explain differences in the magnitude of observed practice effects across studies, as other studies report relatively negligible practice effects.[40] The present findings show the largest practice effect occurs between the first and second test (Cohen's d's = 0.23 to 0.24). Although small in magnitude these results resemble those previously reported for a comparable follow-up interval and suggest different test characteristics may play a role in degree of practice effect results across cohorts.[7,40]

Longitudinal neuropsychological assessment is increasingly recognized as integral to detecting early and transitional cognitive decline in older adults. For example, rate of cognitive decline in preclinical AD can predict future risk for conversion to MCI.[41] The present study found that participants with preclinical AD at baseline had significantly lower conditional z-scores over all follow-up compared to those that were biomarker negative. This suggests the conditional normative approach may be more sensitive to transitional cognitive decline in preclinical AD and may help facilitate early diagnosis when applied to serial assessments. Similarly, other recent studies indicate regression-based change models developed for use in longitudinal neuropsychological assessment are sensitive to subtle cognitive decline, may improve diagnostic accuracy, and can aid in preclinical AD staging.[40,42] Collectively, these studies underscore the importance of using normative models specifically developed for use in repeat neuropsychological assessment, which can improve our ability to detect transitional cognitive decline. Furthermore, longitudinal normative data may play an important role in enrichment of clinical trials, wherein repeated cognitive assessment can help identify individuals most at risk for future cognitive decline.

This preliminary validation of the conditional normative model focused on preclinical AD. However, this approach is likely applicable to other clinical populations in which evaluating cognitive change over time is an important aspect of patient care. For instance, it is common for patients to be evaluated pre- and post-treatment/surgery in other neurological populations, such as epilepsy, neuro-oncology, essential tremor, and Parkinson's disease.[43,44] While it may be anticipated that some individuals show improved performance as opposed to decline, the model would be sensitive to whether individuals remained stable (i.e., do not show expected practice effects) or showed added benefit from practice beyond what is expected in a typical CU individual, and could therefore be used in the same fashion. In addition, because the conditional model is capable of being applied at multiple time points, it would be useful in clinical populations who undergo long-term surveillance.

There are also some limitations to this work. First, there is limited racial and ethnic diversity among Olmsted County residents, and conditional normative data in a more diverse sample is needed. Second, as with other normative studies, the model may be less applicable to individuals at the extremes of the sample distribution, such as level of education less than 11 and age less than 50. Last, even robust normative samples may include individuals who remain CU over time but who have subclinical neurodegenerative and cerebrovascular changes, or positive AD biomarkers that could subtly influence robust normative models. For example, 15 CU individuals in the robust normative sample were A+T+. In addition, there are fewer participants having completed five or more tests, and some caution is needed when applying this model for clinical interpretation at later time points. However, the number of participants and length of time they were followed are considerable strengths of our study, as we provide data for 1001 individuals, 250 of whom completed up to eight serial neuropsychological assessments over 9 years. To the best of our knowledge, this is one of the largest robust samples that provides longitudinal data for the AVLT. Future directions include evaluating application of these conditional norms in other populations and validation in an independent sample. Developing conditional norms for additional neuropsychological tests and combining them into a composite score in the future may also improve early detection in individuals with more widespread deficits.

In summary, we present a valuable clinical and research tool that can be used to objectively quantify an individual's degree of departure from a typical cognitive trajectory. Application of these conditional norms to a sample with preclinical AD suggests that the conditional normative method has promise for identifying transitional cognitive decline.

ACKNOWLEDGMENTS

The authors wish to thank the participants and staff at the Mayo Clinic Study of Aging. This work was supported by the Rochester Epidemiology Project (R01 AG034676), the National Institutes of Health (grant numbers P50 AG016574, P30 AG062677, U01 AG006786, and R37 AG011378, R01 AG041851, RF1 AG55151), a grant from the Alzheimer's Association (AARG-17-531322), Zenith Award from the Alzheimer's Association, the Robert Wood Johnson Foundation, The Elsie and Marvin Dekelboum Family Foundation, Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic, Liston Award, Schuler Foundation, GHR Foundation, AVID Radiopharmaceuticals, and the Mayo Foundation for Education and Research. We would like to greatly thank AVID Radiopharmaceuticals, Inc., for their support in supplying AV-1451 precursor, chemistry production advice and oversight, and FDA regulatory cross-filing permission and documentation needed for this work. We would like to thank Sabrina M. Albertson for building the Rshiny application that allows use of the conditional norms.

CONFLICTS OF INTEREST

NHS & MMMi have served as consultants to Biogen and Lundbeck. D.S.K. serves on a Data Safety Monitoring Board for the DIAN-TU study and is an investigator in clinical trials sponsored by Lilly Pharmaceuticals, Biogen, and the University of Southern California. R.C.P. has served as a consultant for Hoffman-La Roche Inc., Merk Inc., Genentech Inc., Biogen Inc., Eisai, Inc., and GE Healthcare. WKK has received research funding from Biogen, Roche, and AstraZeneca. The authors report no conflicts of interest. All other authors declare no conflicts of interest.

REFERENCES

1. Petersen RC, Lopez O, Armstrong MJ, et al. Practice guideline update summary: mild cognitive impairment. *Neurology*. 2018;90(3):126-135.
2. Jack CR, Bennett DA, Blennow K, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14(4):535-562.
3. Heilbronner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, Hart RP. Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: the utility and challenges of repeat test administrations in clinical and forensic contexts. *Clin Neuropsychol*. 2010;24(8):1267-1278.
4. Elman JA, Jak AJ, Panizzon MS, et al. Underdiagnosis of mild cognitive impairment: a consequence of ignoring practice effects. *Alzheimers Dement (Amst)*. 2018;10:372-381.
5. Duff K, Hammers DB, Dalley BCA, et al., Short-term practice effects and amyloid deposition: providing information above and beyond baseline cognition. *J Prev Alzheimers Dis*. 2017;4(2):87-92.
6. Hammers DB, Suhrie KR, Porter SM, Dixon AM, Duff K. Validation of one-year reliable change in the RBANS for community-dwelling older adults with amnestic mild cognitive impairment. *Clin Neuropsychol*. 2020;20:1-24. Online ahead of print.
7. Gavett BE, Gurnani AS, Saurman JL, et al. Practice effects on story memory and list learning tests in the neuropsychological assessment of older adults. *PLoS One*. 2016;11(10):e0164492.
8. Machulda MM, Pankratz VS, Christianson TJ, et al. Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *Clin Neuropsychol*. 2013;27(8):1247-1264.
9. Machulda MM, Hagen CE, Wiste HJ, et al. Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. *Clin Neuropsychol*. 2017;31(1):99-117.
10. Koscik RL, Jonaitis EM, Clark LR, et al. Longitudinal standards for mid-life cognitive performance: identifying abnormal within-person changes in the Wisconsin Registry for Alzheimer's Prevention. *J Int Neuropsychol Soc*. 2018;25(1):1-14.
11. Gavett BE, Ashendorf L, Gurnani AS. Reliable change on neuropsychological tests in the uniform data set. *J Int Neuropsychol Soc*. 2015;21(7):558-567.
12. Mcsweeny AJ, Naugle RI, Chelune GJ, Lüders H. "T scores for change": an illustration of a regression approach to depicting change in clinical neuropsychology. *Clin Neuropsychol*. 1993;7:300-312.
13. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991;59(1):12-19.
14. Chelune GJ, Naugle RI, Lüders H, et al. Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology*. 1993;7:41-52.
15. Ivnik RJ, Malec JF, Smith GE, et al. Mayo's Older Americans Normative Studies: updated AVLT norms for ages 56 to 97. *Clin Neuropsychol*. 1992;6:83-104.
16. Roberts RO, Geda YE, Knopman DS, et al. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. *Neuroepidemiology*. 2008;30(1):58-69.
17. Kokmen E. The short test of mental status: correlations with standardized psychometric testing. *Arch Neurol*. 1991;48(7):725-728.
18. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*. 1993;43(11):2412-2414.
19. Rey A. *L'examen clinique en psychologie*. Presses Universitaires de France; 1964.
20. Ferman TJ, Lucas JA, Ivnik RJ, et al. Mayo's Older African American Normative Studies: auditory verbal learning test norms for African American elders. *Clin Neuropsychol*. 2005;19(2):214-228.
21. Petersen RC. Mild cognitive impairment as a diagnostic entity. *J Int Med*. 2004;256(3):183-194.
22. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. American Psychiatric Association. 2000.
23. Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: a comparison of four models. *J Int Neuropsychol Soc*. 1999;5:357-369.
24. Duff K, Schoenberg M, Patton D, et al. Regression-based formulas for predicting change in RBANS subtests with older adults. *Arch Clin Neuropsychol*. 2005;20(3):281-290.
25. Attix DK, Story TJ, Chelune GJ, et al. The prediction of change: normative neuropsychological trajectories. *Clin Neuropsychol*. 2009;23(1):21-38.
26. Jack CR, Wiste HJ, Therneau TM, et al. Associations of amyloid, tau, and neurodegeneration biomarker profiles with rates of memory decline among individuals without dementia. *JAMA*. 2019;321(23):2316-2325.
27. Klunk WE, Koeppe RA, Price JC, et al. The centiloid project: standardizing quantitative amyloid plaque estimation by PET. *Alzheimers Dement*. 2015;11(1):1-15.
28. Jack CR, Lowe VJ, Senjem ML, et al. 11C PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment. *Brain*. 2008;131(Pt 3):665-680.
29. Jack CR, Wiste HJ, Weigand SD, et al. Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement*. 2017;13(3):205-216.
30. Vemuri P, Lowe VJ, Knopman DS, et al. Tau-PET uptake: regional variation in average SUVR and impact of amyloid deposition. *Alzheimers Dement (Amst)*. 2017;6:21-30.
31. RStudio. *Shiny from R Studio*. 2020. Accessed April 15, 2022. Available from: https://shiny.rstudio.com/
32. Ivnik RJ, Smith GE, Lucas JA, et al. Testing normal older people three or four times at 1- to 2-year intervals: defining normal variance. *Neuropsychology*. 1999;12(1):121-127.
33. Stricker NH, Christianson TJ, Lundt ES, et al., Mayo normative studies: regression-based normative data for the auditory verbal learning test for ages 30-91 years and the importance of adjusting for sex. *J Int Neuropsychol Soc*. 2020;27(3):1-16.
34. Machulda MM, Pankratz VS, Christianson TJ, et al. Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *Clin Neuropsychol*. 2013;27(8):1247-1264.
35. Machulda MM, Hagen CE, Wiste HJ, et al. Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. *Clin Neuropsychol*. 2016;31(1):99-117.
36. Holtzer R, Goldin Y, Zimmerman M, Katz M, Buschke H, Lipton R. Robust norms for selected neuropsychological tests in older adults. *Arch Clin Neuropsychol*. 2008;23(5):531-541.
37. Casaletto KB, Heaton RK. Neuropsychological assessment: past and future. *J Int Neuropsychol Soc*. 2017;23(9-10):778-790.
38. Sliwinski M, Lipton RB, Buschke H, Stewart W. The effects of preclinical dementia on estimates of normal cognitive functioning in aging. *J Gerontol*. 1996;51B(4):P217-P225.
39. Grober E, Mowrey W, Katz M, Derby C, Lipton RB. Conventional and robust norming in identifying preclinical dementia. *J Clin Exp Neuropsychol*. 2015;27(10):1098-1106.
40. Kiselica AM, Kaser AN, Webber TA, Small BJ, Benge JF. Development and preliminary validation of standardized regression-based change scores as measures of transitional cognitive decline. *Arch Clin Neuropsychol*. 2020:acaa042. https://doi.org/10.1093/arclin/acaa042. Online ahead of print.
41. Papp KV, et al., Clinical meaningfulness of subtle cognitive decline on longitudinal testing in preclinical AD. *Alzheimer Dementia*. 2019;10(3):552-560.
42. Nation DA, Ho JK, Dutt S, Han SD, Lai MHC. Neuropsychological decline improves prediction of dementia beyond Alzheimer's disease

biomarker and mild cognitive impairment diagnoses. *J Alzheimers Dis*. 2019;69(4):1171-1182.
43. Baxendale S, Wilson SJ, Baker GA, et al., Indications and expectations for neuropsychological assessment in epilepsy surgery in children and adults: Executive summary of the report of the ILAE Neuropsychology Task Force Diagnostic Methods Commission: 2017-2021. *Epilepsia*. 2019;90(9):1794-1796.
44. Noll KR, Weinberg JS, Ziu M, Benveniste RJ, Suki D, Wefel JS. Neurocogniteive changes associated with surgical resection of left and right temporal lobe glioma. *Neurosurgery*. 2015;77(5):777-785.

How to cite this article: Alden EC, Lundt ES, Twohy EL, et al. Mayo normative studies: A conditional normative model for longitudinal change on the Auditory Verbal Learning Test and preliminary validation in preclinical Alzheimer's disease. *Alzheimer's Dement*. 2022;14:e12325.
https://doi.org/10.1002/dad2.12325