# Longitudinal standards for mid-life cognitive performance: Identifying abnormal within-person changes in the Wisconsin Registry for Alzheimer's Prevention

Rebecca L. Koscik[a], Erin M. Jonaitis[a], Lindsay R. Clark[b,a,c], Kimberly D. Mueller[a], Samantha L. Allison[a], Carey E. Gleason[b,c], Richard Chappell[c,d,f], Bruce P. Hermann[a,e], Sterling C. Johnson[b,a,c]

**Affiliations:**

[a]Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

[b]Geriatric Research Education and Clinical Center, William. S. Middleton Memorial Veterans Hospital, Madison WI, USA

[c]Alzheimer's Disease Research Center, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA

[d]Department of Biostatistics and Medical Informatics, University of Wisconsin School of Medicine and Public Health, Madison, WI USA

[e]Department of Neurology, University of Wisconsin School of Medicine and Public Health, Madison, WI 53705, USA

[f]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA

**Corresponding author:** Rebecca Koscik, 610 Walnut St. (9th Floor), Madison, WI, 53726

Office ph.: 608-262-6953; fax: 608-265-9122; e-mail: rekoscik@wisc.edu

**Manuscript body word count (max 5000): 5654**

**Abstract word count (max 250): 213**

**Abstract**

A major challenge in the field of cognitive aging is differentiating disease-related cognitive change that has not yet become overt impairment from the more gradual decline in performance expected with normal aging. Published normative reference values are nearly always for single time point performances rather than longitudinal change in performance. To gain insight into how we might tackle the problem of identifying worrisome trajectories, we borrow a method from anthropometry: the development of standards that are conditional on an individual's past measurements. We use quantile regression to create growth-curve-like models of performance on several common neuropsychological tests of memory and executive function while accounting for age, sex, education, estimated verbal ability, and past performance on the test, and then use these to estimate individuals' percentile ranks. Choosing the 7th percentile as a threshold (corresponding to approximately 1.5 standard deviations below the expected mean), we then explore relationships between subthreshold performance, clinical outcomes, and subjective impairment. Participants whose performance fell below the 7th percentile were more likely to be given an abnormal research diagnosis at the current visit, but not at later visits. Performance below this threshold was also linked to subjective and informant reports of worsening memory function. We discuss potential uses of this method in theoretical and applied research and clinical settings.

## Introduction

A major challenge in the field of cognitive aging is differentiating disease-related cognitive change from the more gradual decline in ability expected in normal aging. This is particularly difficult when a person's performance represents a decline from a prior baseline over serial testing, but may not be sufficiently impaired to warrant a diagnosis of mild cognitive impairment (MCI) or dementia. The importance of this challenge is increasing, as the 2018 diagnostic guidelines for Alzheimer's Disease (AD) identify a preclinical stage of cognitive decline called Clinical Stage 2 that represents significant within person change from a previous level of functioning, but not yet severe enough to be categorized as MCI (Jack et al., 2017). How such preclinical change should be operationalized is not yet established.

While most normed neuropsychological instruments intended for use in geriatric populations publish performance norms for several age bands (e.g., Steinberg, Bieliauskas, Smith, Ivnik, & Malec, 2005), with or without adjustments for relevant demographic features such as sex, education, or intelligence, they rarely account for a patient's prior cognitive performance on the instrument itself. The clinician or researcher must resort to using reliable change or deviation estimates that may not fully account for underlying age-associated change or practice effects. Norms that adjust for a test-taker's prior performance as well as demographic features would improve the validity of the interpretation in such circumstances. One approach to this problem has been the development of the regression-based approach to predicting change (Attix et al., 2009; Crawford, Garthwaite, Denham, & Chelune, 2012; Duff et al., 2005; B. P. Hermann et al.,

1996; Maassen, Bossema, & Brand, 2009). In this approach, participants' demographics and baseline scores are used to calculate expected scores on follow-up tests. The difference between predicted and observed values is then compared against an estimate of the standard error of prediction ($SE_p$); individual scores exceeding some threshold (e.g. $\pm 1.5 SE_p$) are considered evidence of reliable change. While influential, this approach assumes that the relationships between predictors and test scores are constant across all predictor strata. It is also a rather indirect way of understanding the quantity of greatest interest: we estimate the mean in order to understand the performance of individuals far from the mean.

Methodologies first developed for anthropometric indices (e.g., height and weight) provide an alternative approach. The first reference curves for height were published over a century ago using heights of Massachusetts schoolchildren (Bowditch, 1891; Cole, 2012). Similar curves are used through the present day to evaluate children's development (WHO Multicentre Growth Reference Study Group, 2006). Although reference curves are commonly termed growth curves, in reality they are often produced using cross-sectional, not longitudinal, anthropometric data, and have little to say about individual height trajectories (Tanner, Whitehouse, & Takaishi, 1966). When considering the growth of an individual child, the quantity of most interest is typically how usual or unusual their stature is now, given previous measurements. We refer to reference curves developed in this way as *conditional*, in contrast with the *unconditional* reference curves produced by cross-sectional information only. Conditional curves can be developed by regressing stature at time $t_j$ on stature at time $t_{j-1}$, controlling for the interval between observations (Berkey, Reed, & Valadian, 1983; Cameron, 1980; Cole, 1995; Healy, 1974). The earliest analyses of

this type used parametric methods, making them conceptually similar to the regression-based methods described by Maassen (2009); however, more recent work in this vein has instead used quantile regression, which requires fewer distributional assumptions (Wei, Pere, Koenker, & He, 2006) and gives directly relevant results. Through these methods, researchers can obtain curves describing the expected median cognitive trajectory, as well as trajectories for more extreme quantiles that denote surprising loss or gain, which standard regression models of means do not.

Work on conditional reference curves has recently been extended to the field of cognitive aging. Cheung and colleagues demonstrated the development of unconditional and conditional references using the Mini Mental State Exam (MMSE; Cheung et al., 2015). Using quantile regression, they established smooth, age-linked unconditional reference curves for several percentiles of interest adjusting for only baseline covariates. A second set of per-person conditional reference curves was then created, adjusting both for covariates and for previous MMSE performance. The conditional references provided a much narrower scope for normative performance than the unconditional references, as well as a clear visual aid for identifying potentially troublesome within-person change.

In this work we extend the conditional reference methodology of Cheung and colleagues to the Wisconsin Registry for Alzheimer's Prevention (WRAP) dataset. WRAP is a longitudinal study of a cohort of middle-aged to older adults who are followed with cognitive testing at regular intervals. This cohort is enriched with risk for Alzheimer's disease (AD) due to parental family history (S. C. Johnson et al., n.d.; Sager, Hermann, & La Rue, 2005). Their mean age at first visit was 54 and is now 65 at their most recent visit,

making WRAP a good population in which to examine preclinical cognitive decline. Our goals in this paper are: (1) to extend the unconditional and conditional reference methods using a wider range of tests, most of which are known to be sensitive to preclinical decline; (2) to develop a graphical tool for contextualizing individual performance over time; (3) to begin reviewing validity evidence for the method by examining how abnormal conditional performance (ACP) relates to concurrent and future clinical cognitive statuses, and to subjective measures of functioning; and (4) to explore the differences between individuals that are flagged for abnormal unconditional and conditional performance on select measures of interest. The overarching goal of the research program is to develop a procedure that can be used in this and other cohorts to identify Stage 2 decline.

## Methods

### Participants

The WRAP cohort currently includes neuropsychological data from 1561 participants who enrolled at midlife (~40-65 years of age) and were free of dementia at baseline. Follow-up visits are conducted at two- to four-year intervals. Participant retention to date is approximately 81%; the median length of follow-up is 9 years for active participants (S. C. Johnson et al., n.d.). The study is conducted in compliance with ethical principles for human subjects research laid out in the Declaration of Helsinki, including approval by the University of Wisconsin Institutional Review Board, and all subjects provided informed consent.

**Inclusion/exclusion criteria.** When constructing growth curves, whether conditional or unconditional, the question of whom to include in the sample is paramount (Cole, 2012; Corvalan, 2014). A reference curve aims to describe typical growth, whereas a standard uses a sample selected for optimal health (e.g., WHO Multicentre Growth Reference Study Group, 2006; J. Xu, Luntamo, Kulmala, Ashorn, & Cheung, 2014). The protracted timecourse and variable age of onset of AD-related dementia and other common dementias make it difficult to apply these strategies for sample selection directly, because it is not clear which members of a given sample are truly free of disease. Instead, many studies exclude only people already evincing clinically-significant impairment, such as those having a diagnosis of probable AD or Parkinson's disease (Kenny et al., 2013) and those with baseline scores on other neuropsychological tests suggestive of MCI (Cheung et al., 2015).

For this paper, we selected the subset of WRAP participants who were free of clinical MCI or dementia through their first two study visits and were free of neuropsychological conditions that could affect cognition. Specifically, participants were excluded from our analyses if they had not returned for at least three visits (N=376); received a clinical diagnosis of MCI or dementia at consensus conference before Visit 3 (N=14); were missing a research diagnosis for Visit 1 or 2, such that clinical status before Visit 3 could not be fully determined (N=2); were missing data on too many neuropsychological tests (N=7); were outside the desired age range (40-75) for one or more of the first three visits (N=14); were missing data on at least one model covariate (N=1); or reported a diagnosis of epilepsy, stroke, multiple sclerosis, or Parkinson's disease at or before visit 2 (N=58). Because our exclusion criteria included a

clinical/neurocognitive component, we refer to the curves we describe in this paper as standards.

## Measures

**Cognitive and clinical outcomes.** At each visit, participants complete a comprehensive neuropsychological battery. For these analyses, we created standards for the following tests and items: Rey Auditory Verbal Learning Test (AVLT; Schmidt, 1996), learning trials sum and delayed recall trial; Trail-Making Test (TMT; Reitan, 1958), part A and part B; WAIS-III Digit Span (Wechsler, 1997), forward and backward; Stroop Test (Trenerry, 1989), Color and Color-Word trials. These tests were selected because of the sensitivity of the episodic memory and executive function domains to early cognitive impairment (Hedden, Oh, Younger, & Patel, 2013) and because they had the most complete data of tests in the WRAP battery, as they were first administered at baseline. We also created standards for discrepancy scores calculated as follows: AVLT, delayed recall minus trial 5; Trail-Making Test, part A minus part B; Digit Span, backward minus forward; and Stroop, color-word minus color. These discrepancy scores were of interest to us based on earlier evidence from our sample that intraindividual cognitive variability may be indicative of cognitive deterioration (Koscik et al., 2016).

Informant-based assessments of clinical status are also collected, including the Informant Questionnaire on Cognitive Decline in the Elderly, or IQCODE [Jorm & Jacomb (1989); score range = 16 to 80, with 48 representing no change and higher scores indicating worsening functioning] and the Quick Dementia Rating System, or QDRS [Galvin (2015); global score range = 0 to 3, with 0.5 indicating mild cognitive impairment]. A more

intensive assessment, the Clinical Dementia Rating Scale (CDR; Morris, 1997), is conducted for anyone with a score greater than 0 on the QDRS, plus a random sample of others (Berman et al., 2017). Scores on the QDRS are designed to parallel those on the CDR; therefore, we combine the scores from the two tests into one estimate of cognitive function, giving the CDR precedence when both scores are available. Subjective measures of functioning included two items representing participants' self-report of memory functioning: "Do you think you have a problem with your memory?" (0=no, 1=yes; Don't know coded to missing); and "Overall, how would you rate your memory in terms of the kinds of problems that you have?" (Likert scale range from 1="Major problems" to 7="No Problems") (taken from the Memory Functioning Questionnaire; Gilewski, Zelinski, & Schaie, 1990).

**Cognitive status determination.** Participant cognitive status is determined after each study visit via a consensus conference review of cognitive performance, medical history, and other factors (for details, see S. C. Johnson et al., n.d.; Koscik et al., 2016). This review results in one of the following research diagnoses: Cognitively Normal; Early MCI (i.e., scores that are low compared to our internal robust norms but which do not cross objective thresholds for MCI); MCI; Dementia; or Impaired Not MCI. This latter category is assigned to a small number of participants whose history suggests longstanding impairment (e.g., history of learning disability) rather than change from a midlife baseline. We included these participants in the standards development sample since they are representative of a certain subset of the population. They were removed, however, from analyses examining whether ACP was predictive of concurrent or subsequent AD-related cognitive outcomes.

## Statistical methods

**Software.** Data management tasks were done using R (R Core Team, 2017) and SAS software version 9.4. Analyses were conducted in R, and documented using RStudio (RStudio Team, 2016) and knitr (Xie, 2017).**Standards development.** For this paper, we created two sets of standards: unconditional standards, which summarize the distribution of each outcome within demographic strata, but do not consider previous measurements of that outcome; and conditional standards, which take into account both demographics and past performance on that test. To build both sets, we constructed regression quantiles using the R package quantreg (Koenker, 2017; R Core Team, 2017). We selected nine quantiles of interest to include the median, the 25th and 75th percentiles, and three quantiles in each tail which, for a normally-distributed outcome, correspond to approximately $\pm 1$, $\pm 1.5$, and $\pm 2$ standard deviations away from the mean (2%, 7%, 16%, 25%, 50%, 75%, 84%, 93%, 98%). For each outcome, preliminary unconditional models including only linear and quadratic age terms were constructed for the selected quantiles. If the quadratic term was nominally significant ($p<.05$) for at least two quantiles, it was retained in the model for all percentiles.

Following model selection, we constructed unconditional standards for each outcome using selected age terms plus three categorical covariates: sex (0=male, 1=female), college completion (0=no bachelor's degree, 1=has degree), and baseline WRAT reading score [Wilkinson (1993); here categorized as 0: 66-89; 1: 90-99; 2: 100-109; 3: 110-120]. We then modeled conditional standards by further controlling for an individual's baseline performance, as measured by their mean score on a given outcome at Visits 1 and

2, and their number of prior exposures to the battery. For both conditional and unconditional standards, we constructed regression quantiles for all percentiles from 1 to 99. Estimated subject-specific conditional and unconditional percentiles were then derived by comparing each true score to the 99 predicted regression quantiles and selecting the quantile with the minimum absolute error. In practice, selection of quantiles could be based upon their intended use and the amount of information available. Larger datasets will enable the estimation of more extreme quantiles than smaller ones. To control for the inclusion of multiple measurements per subject, cluster-robust bootstrap standard errors were used (Hagemann, 2017).

**Abnormal unconditional and conditional performance.** For any given visit and cognitive test, a participant's performance was referred to as abnormal unconditional performance (AUP) if the test score fell below the 7th unconditional percentile. For a normally-distributed outcome, this percentile corresponds to approximately -1.5 SD below the expected mean for that stratum, a cutoff we and others have used in previous work evaluating mild cognitive impairment (Clark et al., 2016; Cook & Marsiske, 2006; Farias, Mungas, & Jagust, 2005; Koscik et al., 2014). Similarly, for visits after baseline, participants whose score fell below the 7th conditional percentile were flagged as exhibiting abnormal conditional performance (ACP) at that visit. In this way, from Visit 3 onward, each test score was associated with two binary variables indicating whether it was potentially troubling compared to unconditional (0=normal, 1=AUP) or conditional (0=normal, 1=ACP) standards. A participant who is flagged for ACP or AUP at one visit for a given test may not be flagged on the same test at the next.

**Graphical tool.** We developed a graphical tool to contextualize individual performance over time. Built on `ggplot2` (Wickham, 2009), this module plots a participant's individual test scores over time against a series of curves representing unconditional regression quantiles by age for a person of that participant's sex, education, and baseline literacy level. Test scores receiving ACP flags are circled. We illustrate use of the graphical tool with three participant cases.

**Assessing construct validity.** To assess whether ACP on a given test measured increased risk of abnormal cognitive status (one aspect of construct validity), we asked two questions: Do our quantile-regression-based abnormal conditional performance (ACP) indicators at a given visit correlate with research diagnoses at the same visit? And does ACP at an early visit, controlling for AUP, predict a person's research diagnosis at the last visit? To answer these questions, we constructed generalized linear models for each outcome, using data from Visit 3 onward (the first visit for which ACP information was available). For models linking ACP to concurrent diagnoses, we thus estimated the relationship using generalized estimating equations for ordinal outcomes (`multgee` package; Touloumis, 2015). For the question relating first-available ACP and AUP indicators to future diagnoses, we estimated the relationship using ordinal regression (`MASS` package, function `polr`; Venables & Ripley, 2002). For these analyses, we eliminated anyone whose research diagnosis was Impaired Not MCI.

Another means of establishing construct validity is to relate our new objective measure of decline to subjective measures. Evidence regarding the utility of subjective complaints in identifying cognitive change is mixed (Roberts, Clare, & Woods, 2009).

However, if complaints are indexing self-awareness about a drop in performance from a high baseline to a lower level that is still above conventional cross-sectional cutoffs, we would expect abnormal conditional performance to have some relationship with these subjective measures. To test this, we modeled ACP as a function of each of three subjective measures using generalized estimating equations for binary outcomes (geepack package; Højsgaard, Halekoh, & Yan, 2006).

Each of the above questions regarding cognitive status and subjective function involved constructing twelve models (one for each of eight test scores, plus one for each of four discrepancy scores), so for each set of twelve, we adjusted p-values for multiple comparisons using the Benjamini-Hochberg method for controlling the false discovery rate (Benjamini & Hochberg, 1995). This method assumes that outcomes are either independent or positively dependent, an assumption our ACP indicators generally met (minimum pairwise $r = -.07$).

**Joint distribution of ACP and AUP indicators.** For all outcomes, we present summaries of Visit 3 performance of participants whose scores place them in each of four categories: normal by both standards; AUP only; ACP only; and both ACP and AUP. For these outcomes, we report Ns in each cell. We expected that approximately 6-7% should meet criteria for each, but had no a priori hypothesis about the overlap between the two.

## Results

### Participant characteristics

After applying exclusion criteria, 1089 participants were included in the standards development sample. Baseline characteristics for this sample are shown in Table 1 overall and by sex.

**Unconditional standards**

Coefficients for unconditional regression quantiles at median and 7th-percentile performance are listed in Table 2 for all outcomes. The quadratic age term was retained for all models except AVLT Total, AVLT Delayed, and Stroop Discrepancy. Models were otherwise unselected, in that all coefficients were retained regardless of significance. Notably, after adding demographic terms, the coefficients for age were near zero (indicating minimal change per year) in several models, including Digit Span Forward (7th percentile), Digit Span Backward (median and 7th percentile), Digit Span Discrepancy (median and 7th percentile), and AVLT Discrepancy (median).

**Conditional standards**

Coefficients for conditional regression quantiles at median and 7th-percentile performance are listed in Table 3. Among these models, the linear age coefficient was near zero for Digit Span Backward (median) and Stroop Discrepancy (7th percentile). Coefficients for practice indicated benefit from previous exposures, except for Stroop Color-Word.

**Graphical display of example cases**

The R programming platform was used to develop a graphical display of an individual participant's performance superimposed on the unconditional standards with

scores falling below our conditional 7th percentile cut-off being demarcated by a red circle. In Figure 1, longitudinal performance of three participants are shown for AVLT Total, AVLT Delayed, and Stroop Color-Word.

Figure 1a illustrates the performance of a woman with a parental history of AD (PH+) who enrolled in WRAP at age 53 and has been followed for five visits over ten years. She has been involved with caregiving for family members since enrollment, including a parent with AD from Visits 1 to 3. Her educational attainment included some college but no degree; at enrollment she worked as a manager, and she retired between Visit 3 and Visit 4. She reported no significant history of mental health problems and CES-D scores were in the normal range for all visits. At her fifth visit she reported significant back pain that required a minor testing accommodation. At baseline, she had a full-scale IQ score of 107 (68th percentile), and a WRAT-3 reading standard score of 92. She was judged to be cognitively normal via consensus conference for four visits, and first given a diagnosis of MCI at Visit 5. However, she exhibited ACP on at least one test starting more than four years previously at Visit 3 (the first for which ACP information was available), and at all subsequent visits. The earliest test to show change was AVLT Total. By Visit 4, she was also exhibiting ACP on AVLT Delay, Trails Discrepancy, and Stroop Color. At Visit 5 (concurrent with her first clinical diagnosis), ACP emerged on AVLT Discrepancy.

Figure 1b illustrates the performance of a man with no parental history of AD (PH-) who enrolled in WRAP at age 63 and has been followed for four visits over eight years. His educational achievement included some graduate school; at the time he enrolled, he was retired from a 30-year career in teaching, and worked part time as a coach. He has a history

of depression and anxiety, for which he began taking buproprion between Visit 3 and Visit 4; his most recent CES-D score was 34. He also reported hearing difficulty, which was evident in testing at Visit 4. At baseline he received a full-scale IQ score of 113 (88th percentile), and a WRAT-3 reading standard score of 103. He was judged to be cognitively normal for three visits, and was assigned a diagnosis of Early MCI at Visit 4. However, he exhibited ACP on AVLT Total and Stroop Color-Word two years previously at Visit 3, and on several tests, including AVLT Total, AVLT Delay, TMT Part A, TMT Part B, Stroop Color, and Stroop Color-Word, at Visit 4.

Figure 1c illustrates the performance of another man (PH-) who enrolled in WRAP at age 51 and has been followed for three visits over seven years. His educational attainment included some graduate school; at enrollment he had worked for 19 years as a management consultant, but by Visit 2 he was working only part-time. He reported no significant mental health history, and CES-D scores were normal at all visits. However, he experienced a stroke between Visits 2 and 3. At baseline, he scored well above-average, receiving a full-scale IQ score of 133 (99th percentile) and a WRAT-3 reading standard score of 111. He has been judged to be cognitively normal at all visits, but at Visit 3 (his most recent), he exhibited ACP on AVLT Total, AVLT Delay, TMT Part B, Stroop Color, and Stroop Color-Word.

**Validity of ACP**

**Concurrent clinical status.** Although the study is still ongoing with only few conversions to MCI and dementia, a validity question we can ask is how much overlap exists between the set of participants flagged on a given outcome as exhibiting abnormal

conditional performance and the set identified with an abnormal research diagnosis, based on test scores from the same visit. Figure 2 depicts results from an ordinal GEE regression linking participants' ACP flags (0=normal, 1=ACP) to their concurrent diagnoses (0=Cognitively Normal; 1=Early MCI; 2=MCI/Dementia). The confidence intervals shown represent the odds ratios of receiving a worse research diagnosis in those with an ACP flag on that test compared to those without; odds ratios greater than 1 indicate increased risk of impairment. Abnormal conditional performance on AVLT Total, AVLT Delayed, AVLT Discrepancy, TMT Part A, TMT Part B, Stroop Color, and Stroop Color-Word were each associated with a higher predicted probability of a concurrent research diagnosis indicating impairment (Early MCI or MCI/Dementia). No relationships were observed with any flag based on Digit Span. Confidence intervals for most discrepancy scores overlapped zero; in the one instance in which this was not true, AVLT Discrepancy, the observed difference was weaker than that for each of the component scores.

**Prospective clinical status.** A second question we can begin to ask is whether ACP on a given test at an early visit adds anything to our understanding of later cognitive outcomes, after controlling for AUP on the same test. Among the subset that was cognitively normal at Visit 3 and had a final research diagnosis at Visit 4 or later (N=709), we examined whether first available ACP (at Visit 3) was predictive of last available research diagnosis (at visit 4, 5, or 6; 0=Cog Normal (N=651), 1=EarlyMCI (N=53), 1=MCI/Dementia (N=5)), controlling for AUP. In no analysis was this predictive relationship significant. Abnormal unconditional performance itself was predictive of final diagnosis for AVLT Delayed and TMT part B. Results are summarized in Table 4.

In a secondary analysis on the larger subset whose research diagnosis was unimpaired (inclusive of normal and early MCI at visit 3; N=830), we examined whether first available ACP and concurrent AUP on a given test were significant predictors of the last available clinical status (0=Cog Normal or EarlyMCI (N=815), 1=MCI/Dementia (N=15)). Similarly to the ordinal results, clinical status was not related to Visit 3 ACP for any of the neuropsychological measures, and only related to Visit 3 AUP for AVLT Delayed.

**Subjective memory complaints and informant reports.** After adjustment for multiple comparisons, Likert-scale self-ratings of memory showed significant relationships with abnormal conditional performance on AVLT Total and AVLT Delayed, and with abnormal *unconditional* performance on AVLT Delayed and TMT part B (Supplemental Table 1; Figure 3). Similarly, participants' responses to the binary question, "Do you think you have a problem with your memory?" were significantly related to abnormal conditional performance on AVLT Total and AVLT Delayed, but not to ACP in other domains (Supplemental Table 2). Finally, models using informant scores on the IQCODE to predict concurrent ACP are summarized in Table 5. Whereas subjective memory related most strongly to ACP on memory measures, IQCODE related instead to ACP on the more difficult conditions from three different tests: AVLT Delayed, TMT part B, and Stroop Color-Word.

**Summary statistics: numbers flagged under each method**

Table 6 contains numbers of participants identified at Visit 3 as exhibiting ACP, AUP, or both. If regression quantiles are well-constructed, approximately 6-7% of measurements should be flagged as either ACP or ACP+AUP; similarly, 6-7% of measurements should be flagged as either AUP or ACP+AUP. Conditional models performed as expected, flagging 6-

7% of observations at Visit 3 for all neuropsychological outcomes. Unconditional models were a bit more variable, generally flagging 5-8% of observations, but more for Digit Span Forward (10%) and Digit Span Backward (10.1%). We see substantial overlap between the categories, with disproportionate numbers receiving both flags; independence of the two variables would lead us to expect $0.07 \times 0.07 \approx 0.005\%$ in the far right column.

## Discussion

This paper describes the development of unconditional and conditional standards for cognitive change across the ages of 40-75 using a cognitively healthy, AD risk-enriched sample. Both sets of standards incorporate sex, literacy, age, and education in development of equations representing percentiles spanning from low to high functioning. In the low range, percentiles 2, 7, and 16 were plotted since they correspond to approximately -2, -1.5, and -1 for variables following a standard normal distribution: cut-offs commonly used to define "abnormal performance" (Clark et al., 2016; Koscik et al., 2014). We depict longitudinal performance of three WRAP participants, superimposing their performance on the unconditional percentiles and highlighting scores that show potentially troubling within-person change according to conditional norms. In these cases, abnormal conditional performance by our measure was seen at visits prior to an abnormal research diagnosis, although in the last case it emerged after the participant experienced a stroke that did not leave him with impairment significant enough to warrant a clinical diagnosis of MCI or dementia.

The application of quantile regression methods to these tests represents a significant extension of the development by Cheung and colleagues (2015) of unconditional

and conditional standards for MMSE, a measure that is less sensitive to subtle cognitive change (refs). Their sample was older (minimum baseline age=55; mean age=65; max age=84) and had less follow-up data (a maximum of two follow-up visits). By extending the method to situations more conducive to identifying early cognitive change, we can begin to evaluate its utility in populations most likely to benefit from prevention efforts.

The development of conditional standards directly addresses new criteria for identifying the earliest stages of decline. In the forthcoming model of the Alzheimer's pathophysiological continuum, currently in development by Jack and colleagues (2017), Stage 2 corresponds to a drop in cognitive performance that has not yet crossed objective thresholds for cognitive impairment. We propose the method described in this paper as one way to operationalize such decline. Critically, current clinical practice does not have an agreed-upon method for identifying worrisome change. The relatively weak relationships we observed between ACP on individual tests and research diagnoses may be a result of this gap: the feature of interest here is preclinical and does not correspond neatly to any extant diagnosis, and our young sample has relatively few clinical endpoints to date. Follow-up analyses will be done when more of the sample has converted to dementia, and when biomarkers are available on enough of the sample to evaluate the method in the context of Jack and colleagues' (2017) ATN framework.

Nevertheless, our data suggest that the conditional standards may be sensitive to subjective changes in memory, a construct whose utility has been debated (Roberts et al., 2009). The relationship we observed is in no sense diagnostic, as most subjects who reported poor memory were not flagged by our algorithm as exhibiting ACP. However, the

existence of a relationship suggests that our approach has face validity, as both measures attempt to estimate cognitive change from some presumably healthy baseline. It is possible that individuals are sensitive enough to their own change that a threshold closer to the median would better match their judgments. It is also possible that our conditionally-poor performers represent a mix of people who are aware of their deficit, and people whose decline is already advanced enough that they are experiencing anosognosia (Roberts et al., 2009; Vannini et al., 2017). Future analyses will need to tease apart these explanations.

We included discrepancy scores in our data because they are thought by some clinicians to be especially sensitive measures (Hayden et al., 2014; Jacobson, Delis, Bondi, & Salmon, 2002; though see also Smith, Ivnik, & Lucas, 2008). However, in our dataset we saw little evidence that widening gaps between the test pairs we selected were related to any other variables of interest, and what evidence we did see did not improve on what we could infer from component scores (Figure 2). By their nature, the reliability of discrepancy scores is bounded by the reliability of the two component scores and their correlation (Crawford, Sutherland, & Garthwaite, 2008), which is a clear limitation in the case of developing longitudinally-informed norms. That said, it may be the case that different pairs of scores would create contrasts more useful than the ones presented here.

Our method is similar in some ways to earlier regression-based methods of identifying reliable change (Crawford et al., 2012; Maassen et al., 2009). The major difference in the quantile regression approach is that it more naturally handles tests whose distributions cannot be expected to be normal, such as the MMSE (e.g. as in Cheung et al., 2015), and situations in which relationships between predictors and outcomes varies by

quantile, as may well be the case for age trajectories on cognitive tests. Future work will explore the relationship between these two methods on the same dataset to better understand the degree of overlap between them.

**Limitations and future directions**

While promising, this work has limitations. First, these standards reflect the performance of a single, non-population-based sample of highly educated, mostly White individuals. More representative standards could be constructed by incorporating data from more diverse sources. With a larger dataset, it would be instructive to partition the data into a training set and a holdout set, in order to better evaluate the nominal quantiles estimated and their relationships with other variables of interest. With such a dataset, it could also be useful to compare robust longitudinal standards, obtained by removing individuals with extremely unusual longitudinal performance from the standards development sample, to the conventional ones described in this paper (Clark et al., 2016; De Santi et al., 2008; cf. Koscik et al., 2014).

Second, the 7th percentile threshold we specified for abnormality, while consistent with past conventional approaches, may not be the most useful for all contexts. Future work will address this by exploring the sensitivity and specificity of multiple thresholds. In addition, there may be information in either the relationship between concurrent ACP flags on different domains, or in the temporal sequencing of flags within a domain, which we have not explored in these data.

Third, although these standards take into account multiple administrations of the same test, the question of how to combine information across tests remains unresolved and

is of high clinical interest. Future analyses will explore whether measures such as the number of flags received at a single visit, or the number of domains with at least one flagged test, are more robust indicators of cognitive change or dementia risk.

Fourth, our technique shares with regression-based norms of the past the need for a large longitudinal dataset on which standards for a given test can be developed. In addition, our use of a multiple-visit baseline limits the immediate clinical application of this technique, as ACP as we have defined it cannot be determined before a patient's third clinic visit. The benefit of this approach is that, by using a baseline that is presumably closer to the true score, it reduces spurious identification of patients whose apparent change represents regression toward the mean. In future work we will examine the usefulness and stability of a "provisional" ACP flag for the second visit.

Finally, we caution that the conditional and unconditional standards presented here only capture test performance and thus should not be used in isolation. As always, clinical judgement is central to making any clinical or research diagnoses based on the totality of information available.

**Conclusions**

In this paper, we have presented application of quantile regression to development of unconditional and conditional (longitudinal) standards for performance on a variety of cognitive measures that are sensitive to cognitive decline associated with Alzheimer's dementia. The graphical interface offers a clinically intuitive visualization of an individual's cognitive performance across measures and over time and clearly identifies tests

demonstrating potentially troubling within person change from baseline. Such methods are needed in order to operationalize preclinical declines in AD and other dementias.

Table 1.

Baseline characteristics of sample, overall and separated by sex.

| Variable | Overall | Male | Female |
|---|---|---|---|
| N | 1089 | 339 | 750 |
| Baseline age, mean(SD) | 54.3 (6.32) | 54.8 (6.24) | 54.1 (6.35) |
| Years of follow-up, mean (SD) | 9.9 (2.25) | 9.9 (2.23) | 9.9 (2.26) |
| WRAT-III Reading, mean (SD) | 105.9 (9.3) | 105.9 (9.92) | 105.9 (9.01) |
| College degree, N (%) | 675 (62%) | 229 (68%) | 446 (59%) |
| Race, white, N (%) | 1036 (95%) | 323 (95%) | 713 (95%) |
| Parental history of AD, N (%) | 793 (73%) | 234 (69%) | 559 (75%) |
| APOE-e4 positive, N (%) | 423 (39%) | 125 (37%) | 298 (40%) |
| AVLT Total | 51.4 (7.88) | 47.9 (8.08) | 53 (7.24) |
| AVLT Delayed | 10.6 (2.77) | 9.4 (2.9) | 11.1 (2.55) |
| AVLT Discrepancy | -1.9 (1.9) | -2.3 (1.94) | -1.7 (1.85) |
| TMT Part A | 26.6 (8.34) | 27.9 (8.48) | 26 (8.22) |
| TMT Part B | 61.6 (22.86) | 65.9 (25.71) | 59.7 (21.19) |
| TMT Discrepancy | 35.1 (20.31) | 38 (23.12) | 33.7 (18.76) |
| Stroop Color | 235.1 (42.23) | 233.8 (45.44) | 235.7 (40.72) |
| Stroop Color-Word | 108.3 (20.57) | 104.7 (22.63) | 110 (19.38) |
| Stroop Discrepancy | -126.8 (38.61) | -129 (39.71) | -125.8 (38.1) |
| Digit Span Forward | 10.6 (2.15) | 10.8 (2.28) | 10.5 (2.09) |
| Digit Span Backward | 7.1 (2.21) | 7.3 (2.33) | 7 (2.16) |
| Digit Span Discrepancy | -3.4 (2.04) | -3.5 (2.08) | -3.4 (2.02) |

Table 2.

Regression coefficients for median and 7th-percentile unconditional models.

| Outcome | %ile | Int | Age | Age^2 | Male | No BA | WRAT <90 | WRAT 90-99 | WRAT 100-109 |
|---|---|---|---|---|---|---|---|---|---|
| AVLT Total | 50 | 56.6 | -0.256 | -- | -5.97 | -1.57 | -5.53 | -3.8 | -2.44 |
| | 7 | 44.8 | -0.265 | -- | -6.22 | -2.84 | -4.81 | -2.61 | -1.63 |
| AVLT Delayed | 50 | 12.5 | -0.0679 | -- | -1.96 | -0.596 | -1.7 | -1.22 | -0.846 |
| | 7 | 7.93 | -0.0814 | -- | -2.26 | -0.862 | -0.726 | -0.621 | -0.278 |
| AVLT Discrepancy | 50 | -1 | -2.41e-17 | 1.93e-18 | -1 | 5.32e-17 | -1 | -1 | -3.9e-16 |
| | 7 | -4.22 | -0.0303 | -0.000656 | -0.777 | -0.227 | -0.501 | 0.145 | -0.167 |
| TMT Part A | 50 | 23.2 | 0.218 | -- | 2.02 | 0.483 | 1.55 | -1.29 | -0.395 |
| | 7 | 16.2 | 0.128 | -- | 0.641 | 0.1 | 1.66 | -0.983 | -0.168 |
| TMT Part B | 50 | 49.1 | 0.752 | 0.0145 | 5.03 | 0.696 | 18.9 | 5.41 | 2.77 |
| | 7 | 33.9 | 0.439 | 0.00333 | 2.96 | 0.528 | 8.84 | 1.69 | 1.07 |
| TMT Discrepancy | 50 | 25.6 | 0.535 | 0.00856 | 2.35 | 1.16 | 16 | 4.96 | 2.84 |
| | 7 | 10.9 | 0.314 | 0.00695 | 1.8 | 0.119 | 9.96 | 4.83 | 2.58 |
| Stroop Color | 50 | 248 | -0.604 | -0.0412 | -2.22 | 1.77 | -47.2 | -19.9 | -9.62 |
| | 7 | 185 | -0.0577 | -0.0645 | -5.47 | 1.45 | -43.6 | -14.2 | -12 |
| Stroop Color-Word | 50 | 117 | -0.706 | -0.0244 | -3.63 | -0.0426 | -23.6 | -6.77 | -3.66 |
| | 7 | 88.1 | -0.8 | -0.0285 | -6.41 | 1.2 | -20.2 | -8.3 | -2.1 |
| Stroop Discrepancy | 50 | -131 | -0.0572 | -- | -2.44 | -0.614 | 26 | 12.4 | 4.46 |
| | 7 | -183 | -0.169 | -- | -3.39 | -0.227 | 34.2 | 15.5 | 9.81 |
| Digit Span Forward | 50 | 11.3 | -0.0332 | -- | 0.426 | 0.144 | -3.23 | -1.62 | -1.05 |
| | 7 | 8 | 4.66e-17 | -- | -4.19e-16 | -5.35e-16 | -2 | -1 | -1.32e-15 |
| Digit Span Backward | 50 | 8 | -6.54e-17 | -8.79e-18 | 5.26e-16 | -3.85e-16 | -3 | -2 | -1 |
| | 7 | 5 | -7.01e-18 | -3.25e-18 | -1.08e-17 | 5.95e-17 | -2 | -1 | -1 |
| Digit Span Discrepancy | 50 | -3 | 2.96e-17 | -6.14e-20 | 1.09e-16 | 1.9e-16 | -2.48e-16 | 3.6e-16 | 2.47e-16 |
| | 7 | -6 | 1.48e-16 | 9.12e-18 | -1 | -2.66e-15 | 7.02e-15 | 8.09e-15 | 5.33e-15 |

Table 3.

Regression coefficients for median and 7th-percentile conditional models.

| Outcome | %ile | Int | Age | Age^2 | Male | No BA | WRAT <90 | WRAT 90-99 | WRAT 100-109 | Practice | Baseline |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AVLT Total | 50 | 12.2 | -0.137 | -- | -1.79 | -0.795 | -1.35 | -1.14 | -1.07 | 0.273 | 0.792 |
| | 7 | 0.445 | -0.188 | -- | -1.9 | -0.45 | -1.15 | -1.21 | -0.723 | 0.202 | 0.866 |
| AVLT Delayed | 50 | 1.94 | -0.0253 | -- | -0.558 | -0.216 | -0.41 | -0.119 | -0.0206 | 0.238 | 0.816 |
| | 7 | -2.23 | -0.011 | -- | -0.152 | -0.38 | 0.51 | 0.085 | 0.182 | 0.232 | 0.882 |
| AVLT Discrepancy | 50 | -1.09 | -0.0269 | 0.000751 | -0.482 | 0.0728 | -0.132 | -0.136 | -0.0964 | 0.256 | 0.485 |
| | 7 | -3.73 | -0.0423 | 0.000879 | -0.383 | -0.315 | 0.696 | 0.538 | 0.179 | 0.13 | 0.561 |
| TMT Part A | 50 | 10.3 | 0.178 | -- | 0.13 | -0.354 | 1.97 | 0.473 | 0.363 | -1.36 | 0.636 |
| | 7 | 9.85 | 0.145 | -- | -0.221 | -0.23 | 1.73 | 0.157 | 0.469 | -0.986 | 0.382 |
| TMT Part B | 50 | 14.3 | 0.292 | 0.00872 | 1.59 | 0.203 | 2.81 | 2.06 | -0.494 | -1.04 | 0.727 |
| | 7 | 16.5 | 0.155 | 0.02 | 1.01 | -0.424 | 4.54 | 1.18 | 0.989 | -1.32 | 0.45 |
| TMT Discrepancy | 50 | 9.71 | 0.157 | 0.0121 | 1.83 | 1.11 | 2.2 | 1.65 | 0.155 | -0.366 | 0.629 |
| | 7 | 2.26 | 0.0554 | 0.0171 | 1.87 | -1.35 | 5.89 | 2.26 | 1.68 | 0.717 | 0.331 |
| Stroop Color | 50 | 25.7 | -0.496 | -0.00945 | -2.79 | 0.813 | -8.35 | -7.13 | -1.53 | 1.53 | 0.884 |
| | 7 | -20.2 | -0.914 | 0.00632 | 6.34 | 7.87 | -16 | -10 | 1.01 | 3.73 | 0.87 |
| Stroop Color-Word | 50 | 11.5 | -0.166 | -0.0186 | -1.03 | -0.0388 | -1.76 | 1.04 | 0.573 | -0.315 | 0.913 |
| | 7 | 3.23 | -0.0796 | -0.0369 | 0.374 | 0.868 | -9.93 | -0.558 | 1.99 | -1.81 | 0.879 |
| Stroop Discrepancy | 50 | -23 | 0.164 | -- | 1.7 | -0.0398 | 9.23 | 5.94 | 2.19 | -2.03 | 0.79 |
| | 7 | -75.7 | -0.0247 | -- | 3.63 | -0.821 | 21.4 | 8.61 | 4.94 | -1.51 | 0.678 |
| Digit Span Forward | 50 | 1.75 | -0.0259 | -- | 0.0759 | -0.0047 | -0.719 | -0.423 | -0.201 | 0.105 | 0.835 |
| | 7 | 1.64 | -0.0192 | -- | -0.133 | 0.0985 | -0.483 | -0.387 | -0.226 | 0.0616 | 0.653 |
| Digit Span Backward | 50 | 1.55 | -0.000684 | -0.00196 | -0.142 | -0.027 | -0.338 | -0.286 | -0.111 | 0.118 | 0.792 |
| | 7 | 0.664 | -0.0144 | -0.00164 | -0.177 | -0.0408 | -0.636 | 0.0494 | -0.348 | 0.179 | 0.612 |
| Digit Span Discrepancy | 50 | -2.02 | 0.0103 | -0.000927 | -0.169 | -0.028 | 0.0608 | 0.0706 | 0.0381 | 0.0251 | 0.353 |
| | 7 | -4.54 | 0.0165 | -0.000314 | -0.23 | -0.0814 | 0.731 | 0.423 | 0.135 | -0.0608 | 0.439 |

Table 4.

Summary of ordinal regression results predicting most recent research diagnosis from Visit 3 ACP and AUP. Reported p-values have been adjusted using the Benjamini-Hochberg procedure.

| Outcome | ACP odds ratio (CI) | ACP p-value | AUP odds ratio (CI) | AUP p-value |
|---|---|---|---|---|
| AVLT Total | 1.21 (0.36-4.07) | 0.818 | 5.24 (1.43-19.26) | 0.056 |
| AVLT Delayed | 2.02 (0.69-5.94) | 0.818 | 5.8 (1.85-18.14) | 0.009 |
| AVLT Discrepancy | 1.47 (0.41-5.35) | 0.818 | 1.93 (0.53-7.05) | 0.218 |
| TMT Part A | 0.67 (0.13-3.4) | 0.818 | 1.2 (0.23-6.23) | 0.979 |
| TMT Part B | 1.22 (0.38-3.91) | 0.818 | 5.24 (1.56-17.62) | 0.014 |
| TMT Discrepancy | 1.97 (0.51-7.64) | 0.818 | 0.35 (0.08-1.6) | 0.402 |
| Stroop Color | 1.18 (0.36-3.84) | 0.818 | 1.18 (0.36-3.84) | 0.740 |
| Stroop Color-Word | 1.54 (0.53-4.52) | 0.818 | 1.87 (0.58-6.08) | 0.288 |
| Stroop Discrepancy | 1.84 (0.65-5.15) | 0.818 | 1 (0.29-3.41) | 0.693 |
| Digit Span Forward | 0.87 (0.27-2.82) | 0.818 | 1.82 (0.76-4.35) | 0.388 |
| Digit Span Backward | 0.55 (0.14-2.11) | 0.818 | 1.65 (0.61-4.47) | 0.693 |
| Digit Span Discrepancy | 0.62 (0.1-3.82) | 0.818 | 0.62 (0.1-3.82) | 0.400 |

Table 5.

Summary of GEE model predicting ACP from informant reports of cognitive functioning (IQCODE). Higher scores on this measure indicate worse functioning; accordingly, odds ratios represent change in risk associated with one-point increase in reports of worsened function. Reported p-values have been adjusted using the Benjamini-Hochberg procedure.

| Outcome | ACP odds ratio (CI) | ACP LR $\chi^2(1)$ | ACP p-value |
|---|---|---|---|
| AVLT Total | 1.02 (0.96-1.08) | 0.29 | 0.683 |
| AVLT Delayed | 1.19 (1.07-1.32) | 10.15 | 0.009 |
| AVLT Discrepancy | 1.06 (1-1.14) | 3.61 | 0.172 |
| TMT Part A | 1.03 (0.96-1.09) | 0.61 | 0.683 |
| TMT Part B | 1.12 (1.03-1.22) | 7.37 | 0.027 |
| TMT Discrepancy | 0.99 (0.95-1.03) | 0.35 | 0.683 |
| Stroop Color | 1 (0.96-1.04) | 0.01 | 0.904 |
| Stroop Color-Word | 1.15 (1.06-1.24) | 12.44 | 0.005 |
| Stroop Discrepancy | 1.03 (0.98-1.07) | 1.24 | 0.638 |
| Digit Span Forward | 1.01 (0.97-1.05) | 0.24 | 0.683 |
| Digit Span Backward | 1.02 (0.96-1.07) | 0.30 | 0.683 |
| Digit Span Discrepancy | 1.02 (0.97-1.08) | 0.54 | 0.683 |

Table 6.

Frequency of ACP and AUP flags at Visit 3, by subtest.

| Outcome | Normal | AUP Only | ACP Only | ACP and AUP |
|---|---|---|---|---|
| AVLT Total | 977 (89.8%) | 44 (4.0%) | 39 (3.6%) | 28 (2.6%) |
| AVLT Delayed | 972 (89.3%) | 48 (4.4%) | 32 (2.9%) | 36 (3.3%) |
| AVLT Discrepancy | 998 (91.7%) | 19 (1.7%) | 20 (1.8%) | 51 (4.7%) |
| TMT Part A | 992 (91.6%) | 24 (2.2%) | 33 (3.0%) | 34 (3.1%) |
| TMT Part B | 976 (90.2%) | 34 (3.1%) | 42 (3.9%) | 30 (2.8%) |
| TMT Discrepancy | 976 (90.5%) | 32 (3.0%) | 22 (2.0%) | 48 (4.5%) |
| Stroop Color | 962 (89.3%) | 45 (4.2%) | 35 (3.2%) | 35 (3.2%) |
| Stroop Color-Word | 963 (89.8%) | 38 (3.5%) | 40 (3.7%) | 31 (2.9%) |
| Stroop Discrepancy | 970 (90.6%) | 29 (2.7%) | 40 (3.7%) | 32 (3.0%) |
| Digit Span Forward | 950 (87.2%) | 66 (6.1%) | 31 (2.8%) | 42 (3.9%) |
| Digit Span Backward | 951 (87.3%) | 68 (6.2%) | 27 (2.5%) | 43 (3.9%) |
| Digit Span Discrepancy | 998 (91.6%) | 23 (2.1%) | 17 (1.6%) | 51 (4.7%) |

Table S1.

Summary of two GEE models predicting ACP and AUP from subjective memory ratings. Odds ratios represent difference in risk associated with one-unit increment in subjective memory rating on a seven-point Likert scale. Reported p-values have been adjusted using the Benjamini-Hochberg procedure.

| Outcome | ACP odds ratio (CI) | ACP LR $\chi^2(1)$ | ACP p-value | AUP odds ratio (CI) | AUP LR $\chi^2(1)$ | AUP p-value |
|---|---|---|---|---|---|---|
| AVLT Delayed | 0.78 (0.67-0.9) | 11.09 | 0.010 | 0.76 (0.66-0.88) | 13.78 | 0.002 |
| AVLT Discrepancy | 1.08 (0.95-1.24) | 1.35 | 0.420 | 0.94 (0.82-1.09) | 0.66 | 0.514 |
| AVLT Total | 0.79 (0.68-0.92) | 9.10 | 0.015 | 0.85 (0.73-0.99) | 4.27 | 0.078 |
| Digit Span Backward | 0.98 (0.85-1.14) | 0.05 | 0.818 | 0.87 (0.77-0.98) | 5.35 | 0.062 |
| Digit Span Discrepancy | 0.95 (0.83-1.08) | 0.60 | 0.524 | 1 (0.87-1.14) | 0.00 | 0.979 |
| Digit Span Forward | 0.93 (0.8-1.08) | 0.95 | 0.496 | 0.95 (0.83-1.09) | 0.52 | 0.514 |
| Stroop Color | 0.9 (0.77-1.04) | 2.02 | 0.311 | 1.06 (0.91-1.24) | 0.54 | 0.514 |
| Stroop Color-Word | 0.82 (0.7-0.96) | 5.88 | 0.061 | 0.74 (0.62-0.89) | 10.75 | 0.004 |
| Stroop Discrepancy | 1.04 (0.9-1.21) | 0.28 | 0.654 | 0.92 (0.77-1.08) | 1.07 | 0.514 |
| TMT Discrepancy | 1.14 (0.99-1.31) | 3.30 | 0.166 | 1.17 (1.02-1.34) | 4.96 | 0.062 |
| TMT Part A | 0.93 (0.8-1.09) | 0.78 | 0.501 | 0.93 (0.77-1.11) | 0.66 | 0.514 |
| TMT Part B | 0.85 (0.73-0.98) | 4.68 | 0.092 | 0.74 (0.63-0.87) | 13.04 | 0.002 |

Table S2.

Summary of GEE model predicting ACP from participants' responses when asked whether they believe that they have a memory problem (0=No, 1=Yes). Odds ratios greater than one indicate increased risk associated with endorsing this item. Reported p-values have been adjusted using the Benjamini-Hochberg procedure.

| Outcome | ACP odds ratio (CI) | ACP LR $\chi^2(1)$ | ACP p-value |
|---|---|---|---|
| AVLT Total | 2.13 (1.42-3.2) | 13.23 | 0.003 |
| AVLT Delayed | 1.89 (1.23-2.91) | 8.46 | 0.022 |
| AVLT Discrepancy | 0.97 (0.61-1.55) | 0.01 | 0.912 |
| TMT Part A | 1.53 (1.02-2.3) | 4.30 | 0.092 |
| TMT Part B | 1.64 (1.07-2.54) | 5.04 | 0.074 |
| TMT Discrepancy | 0.74 (0.45-1.21) | 1.44 | 0.399 |
| Stroop Color | 1.11 (0.71-1.74) | 0.21 | 0.835 |
| Stroop Color-Word | 1.63 (1.08-2.44) | 5.50 | 0.074 |
| Stroop Discrepancy | 1.07 (0.67-1.72) | 0.09 | 0.835 |
| Digit Span Forward | 1.07 (0.69-1.68) | 0.10 | 0.835 |
| Digit Span Backward | 0.73 (0.43-1.23) | 1.42 | 0.399 |
| Digit Span Discrepancy | 0.84 (0.52-1.36) | 0.52 | 0.706 |

Figure captions

Figure 1. Longitudinal performance of three individuals (green) on each of three cognitive tests. Performances are plotted against demographically-adjusted unconditional standard lines for several percentiles (grey). Circles (red) indicate abnormal conditional performance (ACP).

Figure 2. Confidence intervals on odds ratio estimates for ordinal regression predicting research diagnosis from concurrent ACP.

Figure 3. Predicted probability plots for models of ACP (first panel) and AUP (second panel) on AVLT Total as a function of subjective memory performance (x-axis). The linear predictor is shown in red, and the observed proportions at Visit 3 in black; the label indicates the total N observed at Visit 3 for each x-value.

Figure 1.

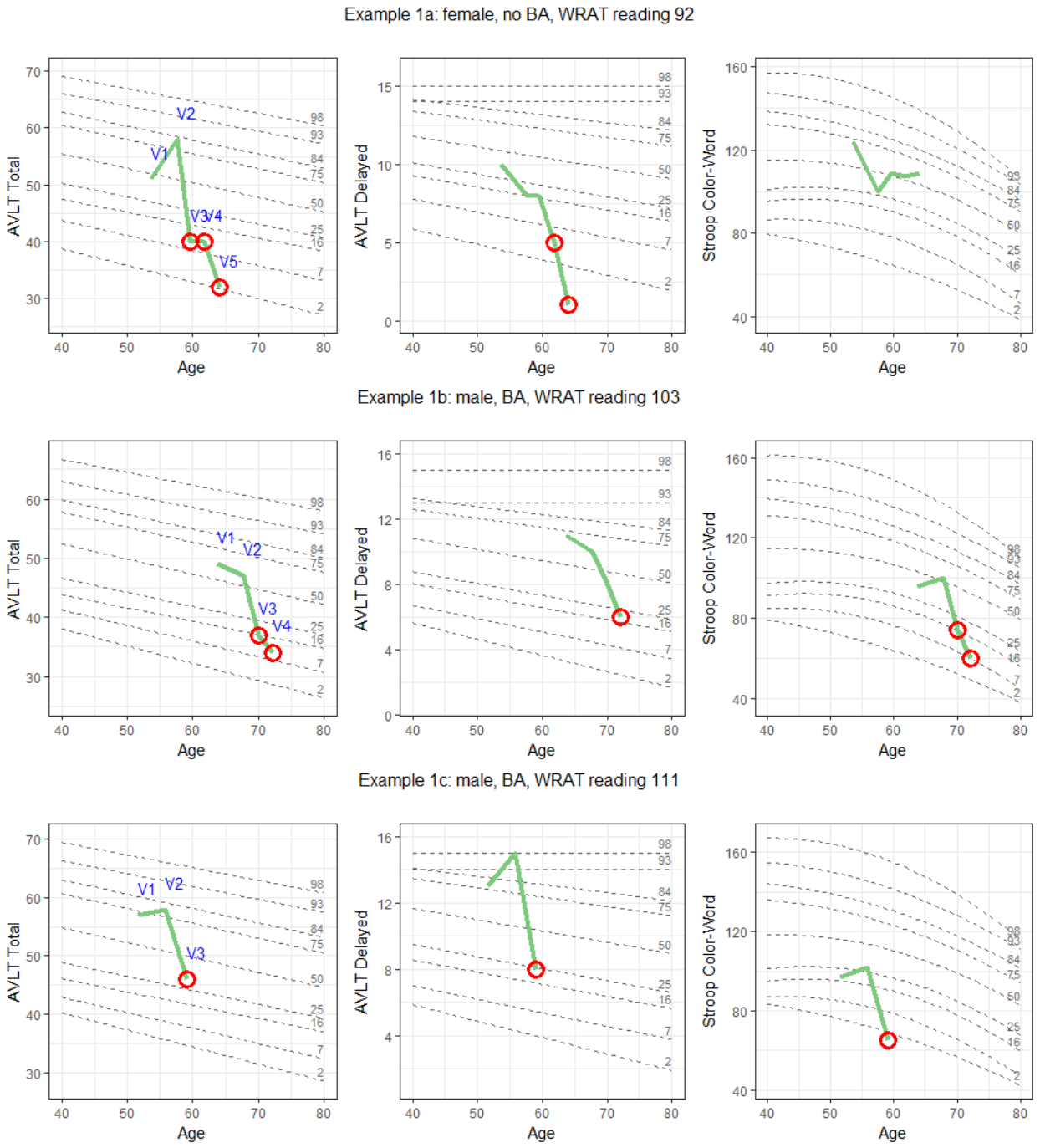

Example 1a: female, no BA, WRAT reading 92
AVLT Total
AVLT Delayed
Stroop Color-Word
Age
V1
V2
V3
V4
V5
Example 1b: male, BA, WRAT reading 103
Example 1c: male, BA, WRAT reading 111

Figure 2.

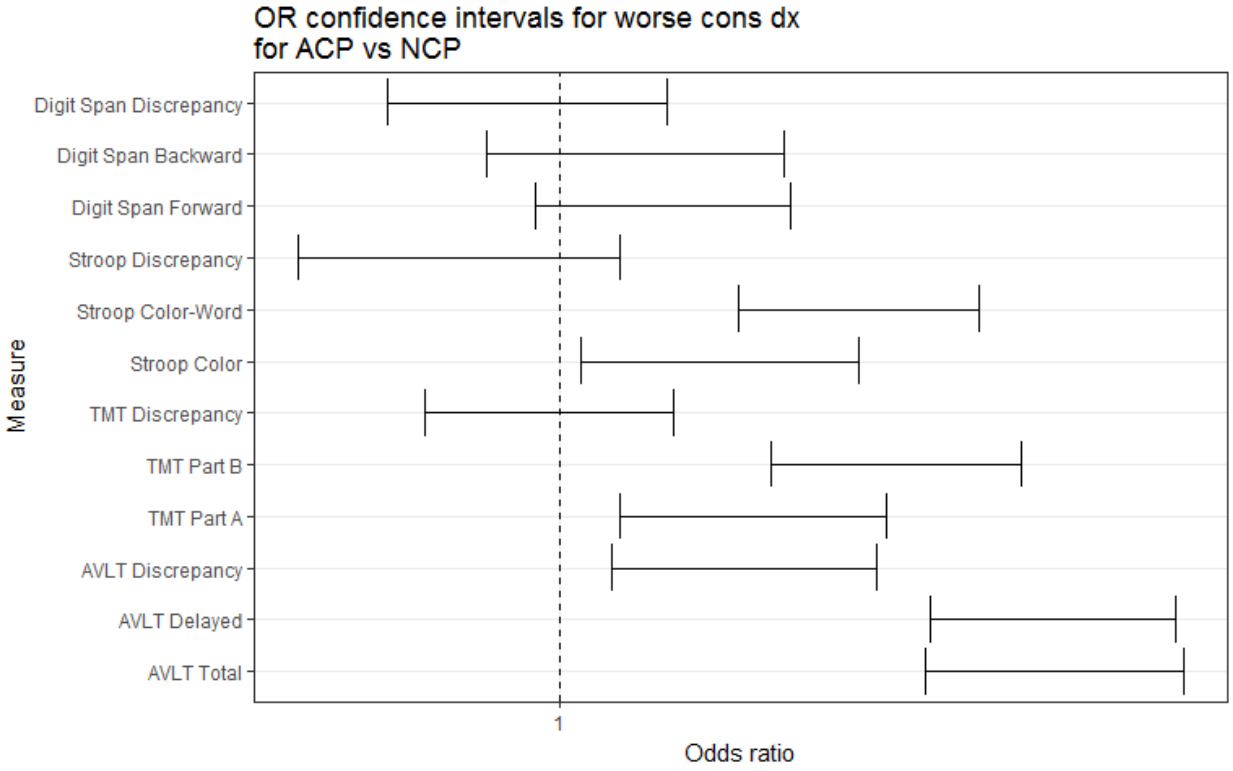
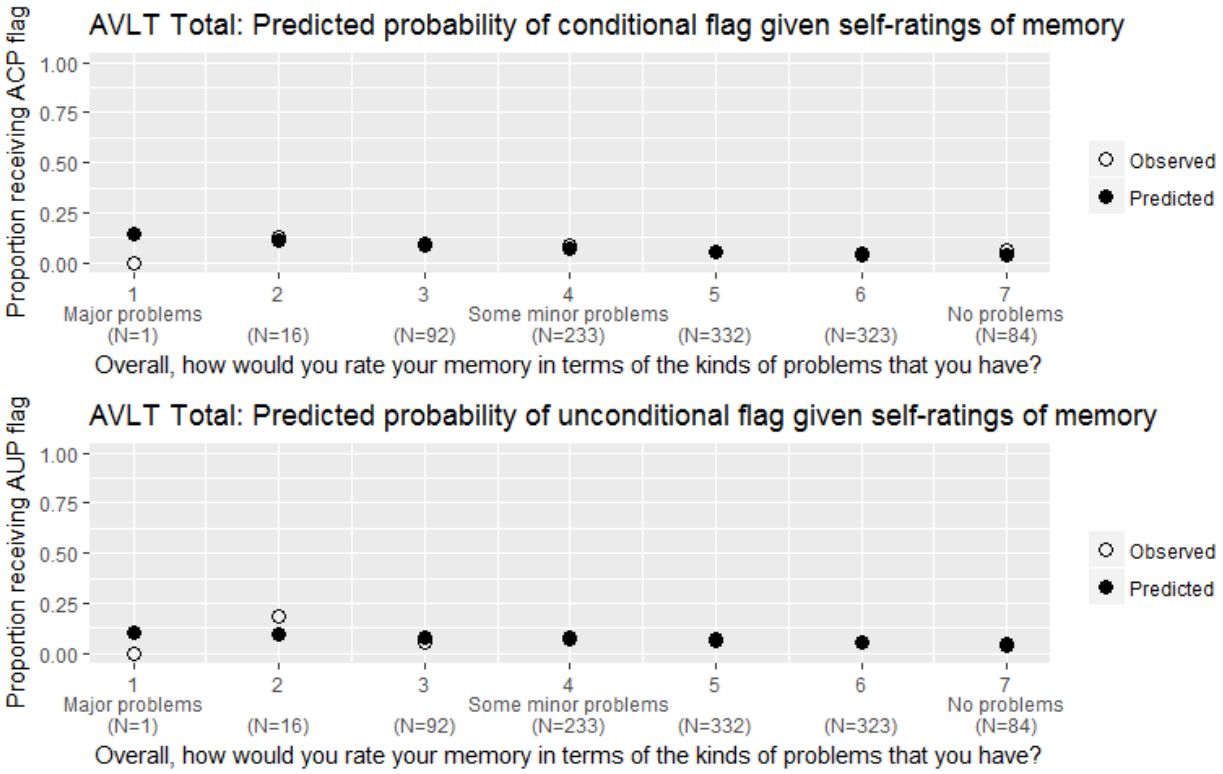

OR confidence intervals for worse cons dx
for ACP vs NCP
Digit Span Discrepancy
Digit Span Backward
Digit Span Forward
Stroop Discrepancy
Stroop Color-Word
Stroop Color
TMT Discrepancy
TMT Part B
TMT Part A
AVLT Discrepancy
AVLT Delayed
AVLT Total
Measure
1
Odds ratio

Figure 3.


AVLT Total: Predicted probability of conditional flag given self-ratings of memory
Proportion receiving ACP flag
1.00
0.75
0.50
0.25
0.00
1
Major problems
(N=1)
2
(N=16)
3
(N=92)
4
Some minor problems
(N=233)
5
(N=332)
6
(N=323)
7
No problems
(N=84)
Overall, how would you rate your memory in terms of the kinds of problems that you have?
Observed
Predicted
AVLT Total: Predicted probability of unconditional flag given self-ratings of memory
Proportion receiving AUP flag
1.00
0.75
0.50
0.25
0.00
1
Major problems
(N=1)
2
(N=16)
3
(N=92)
4
Some minor problems
(N=233)
5
(N=332)
6
(N=323)
7
No problems
(N=84)
Overall, how would you rate your memory in terms of the kinds of problems that you have?
Observed
Predicted

## References

Attix, D. K., Story, T. J., Chelune, G. J., Ball, J. D., Stutts, M. L., Hart, R. P., & Barth, J. T. (2009). The prediction of change: Normative neuropsychological trajectories. *The Clinical Neuropsychologist*, *23*(1), 21–38. https://doi.org/10.1080/13854040801945078

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach. *Journal of the Royal Statistical Society Series B-Methodological*, *57*(1), 289–300.

Berkey, C. S., Reed, R. B., & Valadian, I. (1983). Longitudinal growth standards for preschool children. *Annals of Human Biology*, *10*(1), 57–67.

Berman, S. E., Koscik, R. L., Clark, L. R., Mueller, K. D., Bluder, L., Galvin, J. E., & Johnson, S. C. (2017). Use of the Quick Dementia Rating System (QDRS) as an Initial Screening Measure in a Longitudinal Cohort at Risk for Alzheimer's Disease. *JAD Reports*, *1*(1), 9–13. https://doi.org/10.3233/ADR-170004

Bowditch, H. (1891). The growth of children studied by Galton's percentile grades. In *22nd annual report of the State Board of Health of Massachusetts* (pp. 479–525). Boston, MA: Wright and Potter.

Cameron, N. (1980). Conditional standards for growth in height of British children from 5.0 to 15.99 years of age. *Annals of Human Biology*, *7*(4), 331–337.

Cheung, Y. B., Xu, Y., Feng, L., Feng, L., Nyunt, M. S. Z., Chong, M. S., … Ng, T. P. (2015). Unconditional and Conditional Standards Using Cognitive Function Curves for the Modified Mini-Mental State Exam: Cross-Sectional and Longitudinal Analyses in Older Chinese Adults in Singapore. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, *23*(9), 915–924. https://doi.org/10.1016/j.jagp.2014.08.008

Clark, L. R., Koscik, R. L., Nicholas, C. R., Okonkwo, O. C., Engelman, C. D., Bratzke, L. C., … Johnson, S. C. (2016). Mild Cognitive Impairment in Late Middle Age in the Wisconsin Registry for Alzheimer's Prevention Study: Prevalence and Characteristics Using Robust and Standard Neuropsychological Normative Data. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*. https://doi.org/10.1093/arclin/acw024

Cole, T. J. (1995). Conditional reference charts to assess weight gain in British infants. *Archives of Disease in Childhood*, *73*(1), 8–16.

Cole, T. J. (2012). The development of growth references and growth charts. *Annals of Human Biology*, *39*(5), 382–394. https://doi.org/10.3109/03014460.2012.694475

Cook, S., & Marsiske, M. (2006). Subjective memory beliefs and cognitive performance in normal and mildly impaired older adults. *Aging & Mental Health*, *10*(4), 413–423. https://doi.org/10.1080/13607860600638487

Corvalan, C. (2014). Unconditional or conditional change: Does it matter? Growth charts for monitoring weight gain during pregnancy. *The American Journal of Clinical Nutrition*, *99*(2), 245–246. https://doi.org/10.3945/ajcn.113.079483

Crawford, J. R., Garthwaite, P. H., Denham, A. K., & Chelune, G. J. (2012). Using regression equations built from summary data in the psychological assessment of the individual case: Extension to multiple regression. *Psychological Assessment*, *24*(4), 801–814. https://doi.org/10.1037/a0027699

Crawford, J. R., Sutherland, D., & Garthwaite, P. H. (2008). On the reliability and standard errors of measurement of contrast measures from the D-KEFS. *Journal of the International Neuropsychological Society: JINS*, *14*(6), 1069–1073. https://doi.org/10.1017/S1355617708081228

De Santi, S., Pirraglia, E., Barr, W., Babb, J., Williams, S., Rogers, K., … de Leon, M. J. (2008). Robust and conventional neuropsychological norms: Diagnosis and prediction of age-related cognitive decline. *Neuropsychology*, *22*(4), 469–484. https://doi.org/10.1037/0894-4105.22.4.469

Duff, K., Schoenberg, M. R., Patton, D., Paulsen, J. S., Bayless, J. D., Mold, J., … Adams, R. L. (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *20*(3), 281–290. https://doi.org/10.1016/j.acn.2004.07.007

Farias, S. T., Mungas, D., & Jagust, W. (2005). Degree of discrepancy between self and other-reported everyday functioning by cognitive status: Dementia, mild cognitive impairment, and healthy elders. *International Journal of Geriatric Psychiatry*, *20*(9), 827–834. https://doi.org/10.1002/gps.1367

Galvin, J. E. (2015). The Quick Dementia Rating System (QDRS): A rapid dementia staging tool. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(2), 249–259. https://doi.org/10.1016/j.dadm.2015.03.003

Gilewski, M. J., Zelinski, E. M., & Schaie, K. W. (1990). The Memory Functioning Questionnaire for assessment of memory complaints in adulthood and old age. *Psychology and Aging*, *5*(4), 482–490.

Hagemann, A. (2017). Cluster-Robust Bootstrap Inference in Quantile Regression Models. *Journal of the American Statistical Association*, *112*(517), 446–456. https://doi.org/10.1080/01621459.2016.1148610

Hayden, K. M., Kuchibhatla, M., Romero, H. R., Plassman, B. L., Burke, J. R., Browndyke, J. N., & Welsh-Bohmer, K. A. (2014). Pre-clinical cognitive phenotypes for Alzheimer disease: A latent profile approach. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, *22*(11), 1364–1374. https://doi.org/10.1016/j.jagp.2013.07.008

Healy, M. J. (1974). Notes on the statistics of growth standards. *Annals of Human Biology*, *1*(1), 41–46. https://doi.org/10.1080/03014467400000041

Hedden, T., Oh, H., Younger, A. P., & Patel, T. A. (2013). Meta-analysis of amyloid-cognition relations in cognitively normal older adults. *Neurology*, *80*(14), 1341–1348. https://doi.org/10.1212/WNL.0b013e31828ab35d

Hermann, B. P., Seidenberg, M., Schoenfeld, J., Peterson, J., Leveroni, C., & Wyler, A. R. (1996). Empirical techniques for determining the reliability, magnitude, and pattern of neuropsychological change after epilepsy surgery. *Epilepsia*, *37*(10), 942–950.

Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, *15/2*, 1–11.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Elliott, C., … Sperling, R. A. (2017). *2018 NIA-AA research framework to investigate the Alzheimers disease continuum* (Draft report).

Jacobson, M. W., Delis, D. C., Bondi, M. W., & Salmon, D. P. (2002). Do neuropsychological tests detect preclinical Alzheimer's disease: Individual-test versus cognitive-discrepancy score analyses. *Neuropsychology*, *16*(2), 132–139.

Johnson, S. C., Koscik, R. L., Jonaitis, E. M., Clark, L. R., Mueller, K. D., Berman, S. E., … Sager, M. A. The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimers Dement (Amst)*, in press.

Jorm, A., & Jacomb, P. (1989). The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): Socio-demographic correlates, reliability, validity and some norms. *Psychological Medicine*, *19*(4), 1015–1022.

Kenny, R. A., Coen, R. F., Frewen, J., Donoghue, O. A., Cronin, H., & Savva, G. M. (2013). Normative values of cognitive and physical function in older adults: Findings from the Irish Longitudinal Study on Ageing. *Journal of the American Geriatrics Society*, *61 Suppl 2*, S279–290. https://doi.org/10.1111/jgs.12195

Koenker, R. (2017). Quantreg: Quantile Regression.

Koscik, R. L., Berman, S. E., Clark, L. R., Mueller, K. D., Okonkwo, O. C., Gleason, C. E., … Johnson, S. C. (2016). Intraindividual Cognitive Variability in Middle Age Predicts Cognitive Impairment. *Journal of the International Neuropsychological Society : JINS*, *22*(10), 1016–1025. https://doi.org/10.1017/S135561771600093X

Koscik, R. L., La Rue, A., Jonaitis, E. M., Okonkwo, O. C., Johnson, S. C., Bendlin, B. B., … Sager, M. A. (2014). Emergence of mild cognitive impairment in late middle-aged adults in the Wisconsin Registry for Alzheimer's Prevention. *Dementia and Geriatric Cognitive Disorders*, *38*(1-2), 16–30. https://doi.org/10.1159/000355682

Maassen, G. H., Bossema, E., & Brand, N. (2009). Reliable change and practice effects: Outcomes of various indices compared. *Journal of Clinical and Experimental Neuropsychology*, *31*(3), 339–352. https://doi.org/10.1080/13803390802169059

Morris, J. C. (1997). Clinical dementia rating: A reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics*, *9 Suppl 1*, 173–176; discussion 177–178.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276.

Roberts, J. L., Clare, L., & Woods, R. T. (2009). Subjective memory complaints and awareness of memory functioning in mild cognitive impairment: A systematic review. *Dementia and Geriatric Cognitive Disorders*, *28*(2), 95–109. https://doi.org/10.1159/000234911

RStudio Team. (2016). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc.

Sager, M. A., Hermann, B., & La Rue, A. (2005). Middle-aged children of persons with Alzheimer's disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimer's Prevention. *Journal of Geriatric Psychiatry and Neurology*, *18*(4), 245–249. https://doi.org/10.1177/0891988705281882

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A handbook.* Los Angeles, CA: Western Psychological Services.

Smith, G. E., Ivnik, R. J., & Lucas, J. (2008). Assessment techniques: Tests, test batteries, norms and methodological approaches. In *Textbook of Clinical Neuropsychology* (pp. 38–58). New York, NY: Taylor & Francis.

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., Ivnik, R. J., & Malec, J. F. (2005). Mayo's Older Americans Normative Studies: Age- and IQ-Adjusted Norms for the Auditory Verbal Learning Test and the Visual Spatial Learning Test. *The Clinical Neuropsychologist*, *19*(3-4), 464–523. https://doi.org/10.1080/13854040590945193

Tanner, J., Whitehouse, R., & Takaishi, M. (1966). Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965 Parts I and II., *41*, 454–471; 613–635.

Touloumis, A. (2015). R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses. *Journal of Statistical Software*, *64*(8), 1–14.

Trenerry, B., Crosson. (1989). *The Stroop Neuropsychological Screening Test*. Odessa, FL: Pychological Assessment Resources, Inc.

Vannini, P., Amariglio, R., Hanseeuw, B., Johnson, K. A., McLaren, D. G., Chhatwal, J., … Sperling, R. A. (2017). Memory self-awareness in the preclinical and prodromal stages of Alzheimer's disease. *Neuropsychologia*, *99*, 343–349. https://doi.org/10.1016/j.neuropsychologia.2017.04.002

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). New York: Springer.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale - III*. San Antonio, TX: The Psychological Corporation.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, *25*(8), 1369–1382. https://doi.org/10.1002/sim.2271

WHO Multicentre Growth Reference Study Group. (2006). WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatrica (Oslo, Norway: 1992). Supplement*, *450*, 76–85.

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, G. S. (1993). *The Wide Range Achievement Test: Manual (Third ed.)*. Wilmington, DE: Jastak Association.

Xie, Y. (2017). *Knitr: A General-Purpose Package for Dynamic Report Generation in R*.

Xu, J., Luntamo, M., Kulmala, T., Ashorn, P., & Cheung, Y. B. (2014). A longitudinal study of weight gain in pregnancy in Malawi: Unconditional and conditional standards. *The American Journal of Clinical Nutrition*, *99*(2), 296–301. https://doi.org/10.3945/ajcn.113.074120