UC Davis

UC Davis Previously Published Works

Title

Standardized statistical framework for comparison of biomarkers: Techniques from ADNI.

Permalink

https://escholarship.org/uc/item/9321t5c7

Journal

Alzheimers & Dementia: The Journal of the Alzheimers Association, 20(10)

Authors

Harvey, Danielle

Tosun, Duygu

Jack, Clifford

et al.

Publication Date

2024-10-01

DOI

10.1002/alz.14160

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Received: 1 May 2024 | Revised: 18 June 2024 | Accepted: 20 June 2024

DOI: 10.1002/alz.14160

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

RESEARCH ARTICLE

Standardized statistical framework for comparison of biomarkers: Techniques from ADNI

Danielle J. Harvey[1] | **Duygu Tosun**[2,3] | **Clifford R. Jack Jr**[4] | **Michael Weiner**[2,3,5,6,7] | **Laurel A. Beckett**[1] | **for the Alzheimer's Disease Neuroimaging Initiative**

[1]Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA

[2]Department of Veterans Affairs Medical Center, Center for Imaging of Neurodegenerative Diseases, San Francisco, California, USA

[3]Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA

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

[5]Department of Medicine, University of California, San Francisco, California, USA

[6]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, California, USA

[7]Department of Neurology, University of California, San Francisco, California, USA

Correspondence
Danielle J. Harvey, Division of Biostatistics, Department of Public Health Sciences, University of California, One Shields Avenue, MS1C, Davis, CA 95616, USA.
Email: djharvey@ucdavis.edu

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Funding information
National Institute on Aging, Grant/Award Number: U19 AG024904

Abstract

INTRODUCTION: Well-chosen biomarkers have the potential to increase the efficiency of clinical trials and drug discovery and should show good precision as well as clinical validity.

METHODS: We suggest measures that operationalize these criteria and describe a general approach that can be used for inference-based comparisons of biomarker performance. The methods are applied to measures obtained from structural magnetic resonance imaging (MRI) from individuals with mild dementia ($n = 70$) or mild cognitive impairment (MCI; $n = 303$) enrolled in the Alzheimer's Disease Neuroimaging Initiative.

RESULTS: Ventricular volume and hippocampal volume showed the best precision in detecting change over time in both individuals with MCI and with dementia. Differences in clinical validity varied by group.

DISCUSSION: The methodology presented provides a standardized framework for comparison of biomarkers across modalities and across different methods used to generate similar measures and will help in the search for the most promising biomarkers.

KEYWORDS
ADNI, biomarker comparison, clinical validity, precision

Highlights

- A framework for comparison of biomarkers on pre-defined criteria is presented.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2024 The Author(s). *Alzheimer's & Dementia* published by Wiley Periodicals LLC on behalf of Alzheimer's Association.

- Criteria for comparison include precision in capturing change and clinical validity.
- Ventricular volume has high precision in change for both dementia and mild cognitive impairment (MCI) trials.
- Imaging measures' performance in clinical validity varies more for dementia than for MCI.

1 | BACKGROUND

Dementia is widespread among older adults and poses a huge economic burden on the health care system and families.[1] In addition, the cost of drug development is high, in part due to the length and size of clinical trials and the need for surrogate biomarkers.[2] With the positive results for lecanemab[3] and donanemab,[4] the field can start to evaluate potential markers as surrogate biomarkers, which must not only correlate strongly with clinical change, but must also show differential response to treatment that captures the treatment's effect on the clinical endpoint; surrogate markers will likely be endpoint- and intervention specific.[5]

Well-chosen biomarkers may increase the efficiency of clinical trials through strategies ranging from better defined inclusion and exclusion criteria to the use of surrogate markers as alternative endpoints, and a single marker will likely not serve all purposes. Ideal comparison of markers requires a study acquiring all potential markers and standard clinical outcomes on every participant.[6] The Alzheimer's Disease Neuroimaging Initiative (ADNI) was designed to provide high-quality, uniformly ascertained multi-site data on potential imaging and fluid biomarkers, and on cognitive function and clinical change outcomes, and then to compare the performance of large numbers of potential biomarkers for their precision in capturing change (small variance relative to the estimated change) and association with cognitive change and clinical progression.

Research using data from ADNI or other studies has investigated some of these characteristics related to imaging and fluid biomarkers. Some recent studies have evaluated the precision in annual change,[7] associations with cognitive decline,[8–11] associations with incident dementia,[12–14] and correlations between rates of change in markers and cognitive function.[15] However, comparisons of performance across markers have generally been qualitative rather than inference based, although some used bootstrapping techniques for inference.

In the statistical literature, most development has been done in the context of evaluating markers of binary or time-to-event outcomes. Summary measures derived from a graphical tool have been proposed for the predictiveness of markers of a binary outcome.[16] A semi-parametric estimated-likelihood approach was used to compare individual and combinations of biomarkers as principal surrogate endpoints for a binary clinical endpoint,[17] which was then extended into the setting of time-to-event endpoints.[18] Janes, et al.[19] proposed a comprehensive framework consisting of descriptive and inferential methods, using bootstrapping, for evaluating and comparing candidate markers for patient treatment selection. These approaches are more practical when a small number of markers are being considered. To our knowledge, there has been no proposal for a unified framework for evaluating markers on a set of predefined criteria that can accommodate many markers simultaneously.

In this article, we lay out precision and validity criteria for the performance of biomarkers in dementia and mild cognitive impairment (MCI) and suggest commonly used measures that operationalize these criteria. We then describe a general family of statistical techniques that can be used for inference-based comparisons of biomarker performance and apply these methods to data from the ADNI study. Not all criteria need to be used to assess the performance of a biomarker, but rather researchers can use the specific criteria related to the specific purpose of interest.

We recognize that quantitative magnetic resonance imaging (MRI) measures have fallen out of favor as a surrogate biomarker of efficacy in trials of amyloid-removal therapies because the rates of brain volume loss have been shown to be consistently higher with successful amyloid removal compared to placebo. This is the opposite of what was anticipated based on natural history studies, where greater rates of volume loss map onto faster rates of cognitive decline. A variety of explanations have been offered for this "pseudo-atrophy" phenomenon. It is entirely possible, however, that rates of volume loss might be slowed therapeutically by interventions that do not remove amyloid but target other mechanisms, although this has not yet been proven. Nonetheless, given that various morphometric measures have been performed in ADNI, this data set provides a convenient test bed for evaluating the proposed standardized framework for comparing biomarkers that is the topic of this article.

2 | METHODS

2.1 | The Alzheimer's Disease Neuroimaging Initiative

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early Alzheimer's disease (AD).

The determination of sensitive and specific markers of very early progression in the context of dementia is intended to aid researchers

and clinicians to develop new treatments and monitor their effectiveness, as well as to lessen the time and cost of clinical trials. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and participants have been recruited from more than 50 sites across the United States and Canada. For up-to-date information, see www.adni-info.org

2.2 | Participants

Data presented here are from a subset of participants with dementia and MCI from the second phase of ADNI (ADNI-GO/2), who have at least 1 year of clinical and imaging follow-up. All participants were seen at baseline, month 6, and month 12, with an additional one (dementia) or two (MCI) annual visits. Clinical and cognitive assessments were performed, and images were acquired at each visit. All participants underwent 3T MRI.

Full inclusion/exclusion criteria may be found on the ADNI public information website (https://adni.loni.usc.edu/). Briefly, participants were between 55 and 90 years of age (inclusive) with a study partner able to provide an independent evaluation of functional abilities and spoke either English or Spanish. AD participants had Mini-Mental State Examination (MMSE) scores between 20 and 26 (inclusive), a Clinical Dementia Rating (CDR) of 0.5 or 1, and met the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA)[20] criteria for probable AD. These individuals are referred to as having a diagnosis of dementia throughout the article because of the nonspecificity of a clinical diagnosis of probable AD for actual biologically determined AD. MCI participants had MMSE scores between 24 and 30 (inclusive), a CDR of 0.5, and a memory complaint. In addition, they had objective memory loss as measured by education adjusted scores on the Wechsler Memory Scale Logical Memory II and an absence of dementia. Clinical diagnosis was reassessed at each visit. All participants gave written, informed consent prior to participation through the local institutional review boards at the participating institutions.

2.3 | Cognitive measures

Although numerous cognitive tests were given at each assessment, we selected the Alzheimer's Disease Assessment Scale—Cognitive subscale (ADAS-Cog),[21] the most widely used cognitive outcome in clinical trials; the MMSE[22] and the Rey Auditory Verbal Learning Test (RAVLT),[23] and sum of the five learning trials for use in the analyses. The ADAS-Cog contains 13 items that address fundamental cognitive abilities such as comprehension, language, memory, and constructional and ideational praxis. The total score ranges from 0 to 85, with higher values indicating worse performance. The MMSE is a global measure of mental status covering the domains of orientation, registration, attention and calculations, recall, and language; scores range from 0 to 30, with lower scores indicating worse ability. The RAVLT sum of five trials ranges from 0 to 75, with lower values indicating worse memory.

RESEARCH IN CONTEXT

1. **Systematic review**: The literature was reviewed through PubMed with a specific focus on methods for the comparison of biomarkers and evaluation of markers in the context of dementia.
2. **Interpretation**: A standardized framework for the statistical comparison of potential markers on criteria including precision in capturing change over time and clinical validity is presented and provides a strategy for the field to identify the most promising markers of disease progression. The method is illustrated using magnetic resonance imaging data from the Alzheimer's Disease Neuroimaging Initiative.
3. **Future directions**: The framework may be used in the future to compare across markers quantified from multiple imaging modalities and fluid samples or to compare across processing pipelines or assays measuring the same quantity. Extensions of the work should relax the assumption of a balanced design and identify person-level metrics that account for censoring in the context of survival analysis or longitudinal models accounting for covariates and potential confounders.

2.4 | MRI measures

Details of the ADNI MRI data acquisition protocol are publicly available on the Laboratory of Neuroimaging ADNI website (https://adni.loni.usc.edu/data-samples/data-types/) and have been published.[24] Cortical reconstruction and volumetric segmentation were performed with the FreeSurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). To extract reliable volume and thickness estimates, images were automatically processed with the longitudinal stream[25] in FreeSurfer. Specifically, an unbiased within-subject template space and image is created using robust, inverse consistent registration.[26] Several processing steps, such as skull stripping, Talairach transforms, atlas registration, as well as spherical surface maps and parcellations are then initialized with common information from the within-subject template, significantly increasing reliability and statistical power.[25] Of the many regions and measurements provided by the FreeSurfer suite, volumetric measures (mm^3), averaged across hemispheres, of the hippocampus, entorhinal cortex, ventricles, and whole brain (the sum of cortical gray matter volume, cortical white matter volume, subcortical gray volume, cerebellum cortical volume, and cerebellum white matter volume) and a meta region of interest (meta-ROI; mm), defined as the surface-area weighted average of the mean cortical thickness in the entorhinal, inferior temporal, middle temporal, and fusiform.[27] The ventricular volume in each hemisphere was the sum of the lateral ventricle and the inferior lateral ventricle. Volumetric measures were normalized to intracranial

TABLE 1 Summary of statistical framework for comparisons across criteria.

Criteria	Operationalized metric	Person-specific contribution	W_{ij}
Precision: Annual change	Sample size calculations	$(R_{ij} - \bar{R}_{.j})^2$	$\frac{(R_{ij}-\bar{R}_{.j})^2}{(\bar{R}_{.j})^2}$
Validity: Clinical differentiation	Effect size	$(-1)^{k-1}(X_{ijk} - \bar{X}_{.j.})\frac{1}{n_k}$	$\frac{(-1)^{k-1}(X_{ijk}-\bar{X}_{.j.})\frac{1}{n_k}}{s_{p_j}}$
Validity: Change (dementia)	Correlation	$r_{ij}^2/\hat{\rho}_j$	$r_{ij}^2/\hat{\rho}_j$
Validity: Change (MCI)	Effect size	$(-1)^{k-1}(R_{ijk} - \bar{R}_{.j.})\frac{1}{n_k}$	$\frac{(-1)^{k-1}(R_{ijk}-\bar{R}_{.j.})\frac{1}{n_k}}{s_{p_j}}$

Note: W_{ij} is a dimensionless or common dimension quantity, derived from the person-specific contribution, which enables comparison of markers on different scales. R_{ij} is the annualized rate of change for biomarker j, person i. $\bar{R}_{.j}$ is the average annualized rate of change for biomarker j. X_{ijk} is the value of biomarker j, person I, in group k. $\bar{X}_{.j.}$ is the global mean of biomarker j across all participants in both groups. n_k is the number of participants in group k. s_{p_j} is the pooled standard deviation for biomarker j. r_{ij} is the residual for person i, biomarker j, and $\hat{\rho}_j$ is the estimated correlation for biomarker j.
Abbreviation: MCI, mild cognitive impairment.

volume (ICV) by first fitting a linear regression between each measure and ICV in amyloid negative cognitively normal participants, applying that model to the values in the impaired groups and obtaining the residual and then adding the mean volume in the cognitively normal group.[28]

2.5 | Framework for statistical comparisons of biomarkers

The ultimate goal of the analyses is to be able to compare across potentially very different measures based on a set of established criteria. As discussed earlier, common criteria used to assess the performance of markers may be classified into two major categories: precision and validity. The proposed statistical framework for comparing biomarkers on a particular criterion consists of the following five steps.

1. Operationalize the criterion with a statistical measure.
2. Identify subject-level contributions for individual i and measure j. This step makes use of the fact that most statistical measures are linear combinations of person-level information. This person-level information is the contribution of interest.
3. Transform the contributions to a dimensionless or common dimension quantity, W_{ij}, to be comparable across markers on different scales.
4. Construct an m x n matrix of dimensionless (or common dimension) quantities for m participants and n markers for statistical analysis.
5. Use standard statistical procedures for an overall test of differences across markers and multiple comparison procedures for identifying specific differences.

Specific criteria of interest include: (1) precision in measurement of change; (2) validity at baseline (clinical differentiation); and (3) validity in change. The first three steps for metrics to be used in ADNI for each of these criteria are described separately in detail below and summarized in Table 1. Details of Steps 4 and 5, which will use the same approach regardless of the criteria or metric used to assess the criteria, are then described. Markers performing well on these criteria may be candidates for future study as true surrogate markers that also capture the effect of a treatment on a clinical outcome or as markers to be used for inclusion/exclusion criteria or other purposes.

2.5.1 | Precision: Annual change

Participants with dementia are expected to show cognitive decline within 1 year, although there may still be variability in the amount of change experienced by the individuals. MCI participants, on the other hand, show much more heterogeneity in their cognitive performance, with some declining and others staying stable. Potential markers in either population, however, should exhibit precise estimates of mean change, meaning small variability relative to the estimated change, which leads to smaller sample sizes and efficiency of the design of a trial. Therefore, this criterion is operationalized (Step 1) as sample size calculations required for a two-arm (equal-sized) 1-year clinical trial to detect a 25% reduction in annual rate of change, assuming 80% power, a two-sided test, and $\alpha = 0.05$. In this context, most sample size calculations are a direct function of the dimensionless quantity σ/δ, where we estimate δ by the sample mean rate of change and σ by the sample standard deviation (SD). This quantity is a measure of precision and is related to the coefficient of variation. Because most sample size calculations, for any study design aimed at measuring change, are a function of this quantity, σ/δ is useful for comparing the performance of markers in any longitudinal study design. The components that depend on data are the variance in rate of change (which itself is proportional to the sum of the squared deviations between person-specific rates and the average rate) and the average change. The person-level contribution to this quantity (Step 2) is $(R_{ij} - \bar{R}_{.j})^2$, where $\bar{R}_{.j} = \sum_{i=1}^{n} \frac{R_{ij}}{n}$, $R_{ij} = D_{ij}/t_{ij}$, D_{ij} is the difference in marker j between the month 12 and baseline measures for participant i, t_{ij} is the time between the two assessments for marker j, participant i, and n is the number of participants. This quantity is invariant with respect to the desired α, power, and targeted percentage reduction. The dimensionless quantity (Step 3) is, therefore, $W_{ij} = \frac{(R_{ij}-\bar{R}_{.j})^2}{(\bar{R}_{.j})^2}$. Multiplying this quantity by *1/(n-1)* and adding across

participants yields an estimate of σ/δ. Measures that are more precise will have smaller values of these scaled squared deviations.

2.5.2 | Validity at baseline: Clinical differentiation

Good markers should be able to differentiate individuals that are clinically different from one another. For example, one would expect that a marker of disease would be "worse" at baseline in participants with dementia compared to MCI participants, because those with dementia, clinically, are further along in the disease process. Interest lies in comparing marker levels between two groups. Groups may be defined according to the clinical diagnosis at the time the marker was acquired. This criterion, therefore, is operationalized (Step 1) as the effect size $\frac{|\bar{X}_{j1}-\bar{X}_{j2}|}{s_{p_j}}$, where $\bar{X}_{jk}$ is the mean level of marker j (i.e., a regional brain volume from MRI adjusted for head size) in group k (dementia or MCI) for measure j, and s_{p_j} is the common SD in the two groups. By adding and subtracting the global mean level across all participants, this effect size is then proportional to the difference in the deviation from the global mean in one group and that same deviation in the other group. If n_k is the number of participants in group k, the person-level contribution to this quantity (Step 2) is $(-1)^{k-1}(X_{ijk}-\bar{X}_{j.})\frac{1}{n_k}$, where X_{ijk} is the level of marker j (adjusted for head size, by residualizing it) for person i, in group k, and $\bar{X}_{j.}$ is the global mean of marker j across participants in both groups. Dividing by the pooled SD generates a dimensionless quantity (Step 3), $W_{ijk}=\frac{(-1)^{k-1}(X_{ijk}-\bar{X}_{j.})\frac{1}{n_k}}{s_{p_j}}$, which when added across participants and scaled yields the usual two-sample t-statistic. For these analyses, it is important to force all markers of interest to move in the same direction. For example, larger hippocampi are good, whereas larger ventricles are bad. To put them in the same direction, multiply the ventricular volumes by −1, so that larger (less-negative) values correspond to better outcomes. Markers that show large effects between the groups will have larger values for the W_{ijk}.

2.5.3 | Validity in change

Operationalizing this criterion depends on the diagnostic group of interest. For example, because the cognitive function of participants with dementia is expected to decline, change in a marker should also correlate with change in cognitive function; if not, the marker may be changing for reasons unrelated to the disease process. In MCI participants, where cognitive function is more variable, we might expect that those who did progress to dementia showed more change on the marker than those who remained stable. The approaches for each diagnostic group are described separately.

One natural metric of performance in participants with dementia is the correlation of change in the biomarker with clinical change, for example, change in a cognitive test score. Specifically, first calculate individual-level estimates of annual cognitive change by using linear regression on all available cognitive data for each individual, with the score as the outcome and time in years as the predictor. The slope of this model represents an estimate of annual change in cognitive function for that person. Using a similar approach, the estimated annual change in marker j may be calculated and then used as a predictor in a regression model with the estimated annual change of cognitive function as the outcome. The Pearson correlation coefficient from this model is a natural way to operationalize this criterion (Step 1). The t-statistic for the coefficient of marker j in the regression model may be shown to be a function of the squared residuals and the estimated correlation. Therefore, the person-level contribution is $r_{ij}^2/\hat{\rho}_j$ (Step 2), where r_{ij} is the residual for person i, marker j, and $\hat{\rho}_j$ is the estimated correlation for marker j. Although these values are not dimensionless, they are units-free with regard to the markers because the deviations are all measured in the units of the common outcome measure, not the predictor. Thus, for example, a cortical thickness measure could be compared with hippocampal volume for its accuracy in predicting ADAS-Cog because the outcome for each person would be a function of the squared error of prediction in ADAS-Cog units for each biomarker. Therefore, this value is also used as the person-level component (Step 3). These scaled squared residuals will be small for points close to the regression line and large for points further away from the regression line, and we expect that measurements that are more highly correlated (in magnitude) with cognitive change will have smaller scaled squared residuals.

Similarly to the validity described for clinical differentiation, in MCI participants, validity in change can be assessed through a two-sample comparison where the groups are defined according to progression to dementia or remaining stable over a fixed time period. Steps 1–3 are identical to those defined in Section 2.5.2, where the rates of change are used in place of the baseline levels of the markers.

2.5.4 | Statistical methods for comparison (Steps 4 and 5)

For simplicity, it is assumed that all measures have been collected on each person, so that we have a completely balanced design. This assumption means that once we have calculated the W_{ij} for a particular criterion, we can construct an array with one row per person and one column per marker. Friedman's rank test or randomized block analysis of variance (for data meeting the assumption of normality), with participants as blocks and measures as "treatments," is used to assess an overall difference between the measures. Post hoc pairwise comparisons, adjusted for multiple comparisons using a distribution-free approach based on the Friedman's rank sums[29] or Tukey's honestly significant difference (HSD) approach, are used to identify specific differences across the markers. For this study, Friedman's rank test and the distribution-free approach for multiple comparisons were used to compare methods. To illustrate differences between markers, columns are added to the tables with shaded cells. Markers that have shaded cells within a column do not differ significantly from one another.

TABLE 2 Sample characteristics.

Variable	MCI ($n = 303$)	Dementia ($n = 70$)	p-value
Age	71.7 (7.4)	73.7 (7.5)	0.04
Education (years)	16.1 (2.7)	15.4 (2.4)	0.04
Male (%)	53.1	58.6	0.49
Caucasian (%)	94.1	91.4	0.42
MMSE	28.0 (1.7)	23.2 (2.0)	<0.001
ADAS-Cog (total 13)	14.9 (6.6)	29.3 (7.5)	<0.001
RAVLT	36.6 (10.5)	22.6 (6.5)	<0.001
CDR sum of boxes	1.48 (0.91)	4.24 (1.65)	<0.001
APOE ε4+ (%)	51.1	80.0	<0.001

Abbreviations: ADAS-Cog, Alzheimer's Disease Assessment Scale—Cognitive subscale; CDR, Clinical Dementia Rating; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; RAVLT, Rey Auditory Verbal Learning Test.

3 | RESULTS

Our sample consisted of a subset of 70 dementia and 303 MCI participants from the ADNI parent study who had at least 1 year of clinical follow-up, baseline, and month 12 MRI scans available and volumetric measures from FreeSurfer available on those scans. Table 2 provides a description of the participants included in this study. Those with dementia were slightly older than the MCI participants and had slightly fewer years of education. The dementia group had a higher prevalence of having at least one apolipoprotein E (*APOE*) ε4 allele. As expected, the cognitive scores were worse, on average, in the dementia group compared to the MCI group.

3.1 | Precision: Annual change

Tables 3 and 4 present the sample size required for each arm of a two-arm 1-year dementia or MCI trial to have 80% power to detect a 25% reduction in annual rate of change. Three commonly used measures of cognitive function in clinical trials (ADAS-Cog, MMSE, and RAVLT) are included in the comparison as a reference to the precision of currently used outcomes. Measures corresponding to a shaded cell within a column do not differ significantly after multiple comparison adjustment. For both dementia and MCI trials there is variable performance across measures, with an overall significant difference across measures ($p < 0.001$). In dementia trials, the commonly used cognitive measures (MMSE, ADAS-Cog, and RAVLT) require the most participants, and their precision to detect change did not differ significantly with sample sizes ranging from 415 to 1349 individuals per arm of a trial. Volume of the hippocampus and ventricles had significantly greater precision than the cognitive measures, resulting in a substantial reduction in the required sample size (118–200 per arm). The cortical thickness meta-ROI and whole brain tissue volume had greater precision than the MMSE and RAVLT, whereas the entorhinal cortex volume had greater precision than RAVLT. In MCI trials the cognitive tests require the most participants per arm. The imaging measures all performed better, resulting in at least a 71.6% reduction in required sample size. Ventricular volume required the smallest number of participants per arm at just over 300 participants per arm.

3.2 | Validity: Clinical differentiation of dementia from MCI participants

Means and SDs of the imaging variables at baseline are presented for MCI and dementia participants in Table 5. There is a statistically significant difference in all of the markers under consideration between MCI and dementia, with effect sizes ranging from 0.51 to 1.5. Although there was a global difference across methods ($p < 0.001$), effect sizes did not differ significantly for the majority of markers. The cortical thickness meta-ROI and hippocampal volume had a significantly larger effect size than whole brain tissue volume, when comparing dementia to MCI participants.

3.3 | Validity: Change in dementia and MCI participants

In participants with dementia, estimated change in the cortical thickness meta-ROI correlated significantly with change in ADAS-Cog (Table 6). When compared across the imaging metrics, cortical thickness meta-ROI and ventricular volume had significantly higher correlations (in magnitude) than the other metrics. Volume of the entorhinal cortex had the smallest correlation (in magnitude).

Table 7 presents the mean annual change in the imaging measures for the MCI participants that progressed to dementia and those who remained stable over 36 months, based on the subset of 239 individuals who had either progressed to dementia by the 36-month visit or who had been followed for the entire 3-year period. All imaging metrics considered had significantly different average change in those that progressed versus those that did not, with those progressing showing the most change. Hippocampal volume performed the best across the imaging metrics for this criterion, with an effect size of 1.30.

4 | DISCUSSION

We presented a framework for comparing the performance of potential markers on the basis of pre-defined precision and validity criteria, which allow for the comparison of markers measured on different scales. We provide an example of a metric to operationalize each criterion, how to identify the person-level contribution to that metric, and how to achieve unitless (or same units) contributions, which can then be used in analyses. We illustrated the methodology using a subset of MCI and dementia participants from ADNI and four MRI volumes and one cortical thickness measure associated with dementia. In dementia participants, differences existed in precision in change and clinical

TABLE 3 Sample size required per arm of a two-arm 1-year clinical trial in dementia participants to detect a 25% reduction in annual rate of change ($n = 70$), assuming $\alpha = 0.05$, a two-sided test and 80% power.[a]

Variable	Mean annual change (SD)	n/arm[b]
Ventricular volume (mm^3)	2136.6 (1459.4)	118
Hippocampal volume (mm^3)	−111.3 (99.1)	200
Cortical thickness meta-ROI (mm)	−0.08 (0.07)	201
Whole brain tissue volume (mm^3)	−18345.8 (17386.9)	226
Entorhinal cortex volume (mm^3)	−70.0 (79.8)	327
ADAS-Cog (total 13)	4.4 (5.7)	415
MMSE	−1.6 (2.9)	799
RAVLT	−1.9 (4.3)	1349

Abbreviations: ADAS-Cog, Alzheimer's Disease Assessment Scale—Cognitive subscale; meta-ROI, meta region of interest; MMSE, Mini-Mental State Examination; RAVLT, Rey Auditory Verbal Learning Test.

[a]To illustrate comparisons between measures, including imaging markers and cognitive test scores, we have included columns with shaded cells. Measures corresponding to a shaded cell within a column are not statistically different.

[b]Sample sizes are computed with the non-rounded means and standard deviations.

TABLE 4 Sample size required per arm of a two-arm 1-year clinical trial in MCI participants to detect a 25% reduction in annual rate of change ($n = 303$), assuming $\alpha = 0.05$, a two-sided test and 80% power.[a]

Variable	Mean annual change (SD)	n/arm[b]
Ventricular volume (mm^3)	1084.0 (1191.3)	304
Hippocampal volume (mm^3)	−57.0 (86.3)	576
Whole brain tissue volume (mm^3)	−9851.4 (15639.6)	634
Cortical thickness meta-ROI (mm)	−0.04 (0.07)	714
Entorhinal cortex volume (mm^3)	−45.5 (84.4)	866
MMSE	−0.5 (1.8)	3054
RAVLT	−0.6 (7.0)	37715
ADAS-Cog (total 13)	−0.04 (4.5)	3138886

Abbreviations: ADAS-Cog, Alzheimer's Disease Assessment Scale—Cognitive subscale; meta-ROI, meta region of interest; MMSE, Mini-Mental State Examination; RAVLT, Rey Auditory Verbal Learning Test.

[a]To illustrate comparisons between measures, including imaging markers and cognitive test scores, we have included columns with shaded cells. Measures corresponding to a shaded cell within a column are not statistically different.

[b]Sample sizes are computed with the non-rounded means and standard deviations.

TABLE 5 Baseline imaging measures (mean [SD]) for dementia and MCI participants.[a]

Variable	MCI ($n = 303$)	Dementia ($n = 70$)	p-value	Effect size[b]
Cortical thickness meta-ROI (mm)	2.8 (0.2)	2.5 (0.2)	<0.001	1.5
Hippocampal volume (mm^3)	3400.6 (579.2)	2832.1 (457.4)	<0.001	1.0
Entorhinal cortex volume (mm^3)	1786.4 (366.3)	1469.5 (299.7)	<0.001	0.89
Ventricular volume (mm^3)	18,064.8 (8968.9)	23,991.9 (9543.2)	<0.001	0.65
Whole brain tissue volume (mm^3)	1,211,612 (69,450)	1,175,832 (72,126.4)	<0.001	0.51

Abbreviations: MCI, mild cognitive impairment; meta-ROI, meta region of interest.

[a]To illustrate comparisons between imaging measures, we have included columns with shaded cells. Measures corresponding to a shaded cell within a column are not statistically different.

[b]Effect size computed with the pooled estimate of the standard deviation.

TABLE 6 Correlations between change in MRI and change in ADAS-Cog in dementia participants ($n = 70$).[a]

Variable	Correlation	*p*-value
Cortical thickness meta-ROI (mm)	−0.38	0.001
Ventricular volume (mm^3)	0.23	0.054
Hippocampal volume (mm^3)	0.11	0.36
Whole brain tissue volume (mm^3)	−0.07	0.55
Entorhinal cortex volume (mm^3)	0.01	0.93

Abbreviation: meta-ROI, meta region of interest.
[a]To illustrate comparisons between imaging measures, we have included columns with shaded cells. Measures corresponding to a shaded cell within a column are not statistically different.

TABLE 7 Imaging change measures (mean [SD]) by progression status within 3 years in MCI.[a]

Variable	Progressors ($n = 75$)	Non-progressors ($n = 164$)	*p*-value	Effect size[b]
Hippocampal volume (mm^3)	−109.7 (59.1)	−44.5 (45.2)	<0.001	1.30
Ventricular volume (mm^3)	1970.4 (1526.1)	770.5 (705.2)	<0.001	1.16
Whole brain tissue volume (mm^3)	−18580.2 (13572.9)	−7993.8 (6942.1)	<0.001	1.11
Entorhinal cortex volume (mm^3)	−80.6 (59.4)	−31.1 (41.4)	<0.001	1.03
Cortical thickness meta-ROI (mm)	−0.06 (0.05)	−0.02 (0.03)	<0.001	1.00

Abbreviation: meta-ROI, meta region of interest.
[a]To illustrate comparisons between imaging measures, we have included columns with shaded cells. Measures corresponding to a shaded cell within a column are not statistically different.
[b]Effect size computed using the pooled estimate of the standard deviation.

validity of change with ventricular volume performing well under both criteria. Additional well-performing markers from those selected in this analysis include the cortical thickness meta-ROI. In those with MCI, the main difference between markers was in precision in change, with the volume of the ventricles performing the best. Many of the considered markers were promising in showing differences between MCI and dementia participants at baseline, whereas change in hippocampal volume showed the largest differences between stable-MCI and those who progressed to dementia. Although hippocampal volume and ventricles performed well among the markers included in the comparison, they are likely not useful as surrogate markers, in part due to the increased atrophy in groups treated with anti-amyloid therapies relative to the placebo group as mentioned in the Background, but also because the correlation between change in these measures and change in cognitive function is fairly weak (Table 6). However, our focus was on illustration of the proposed framework and how it can be used to compare biomarkers rather than promoting specific markers for use in clinical trials.

Although our illustration of the framework utilizes individual biomarkers and individual outcomes (single cognitive test scores or progression in the comparison), in clinical trials there may be composite outcomes combining two or more single outcomes, co-primary outcomes (which may have different constructs), secondary domain-specific outcomes, or even interest in composite markers. When considering a comparison of biomarkers, researchers should consider how the biomarker will be used, the specific clinical outcome(s) of interest, as well as the intervention. Different biomarkers may be included in a comparison depending on these decisions. A comparison of performance across biomarkers for multiple outcomes, for example, correlations with cognitive change across multiple cognitive scores, may be warranted. It is unlikely that a single biomarker will work for multiple purposes and multiple interventions.

The methodology presented here assumes that all markers of interest are available for all participants. This assumption has two possible impacts on the analysis. First, the number of participants in the comparison may be reduced to a subset on which all markers are available. Second, certain markers may be removed from consideration due to lack of sufficient overlap with the other markers. The criteria for precision in annual change also assume that a 25% reduction (or whatever reduction of interest) in annual change means the same thing across markers and cognitive outcomes, which may not be the case. A modification to the proposed strategy might instead consider some percentage reduction in the relative change in MCI or dementia compared to cognitively normal. In our framework, we have opted to focus on MCI and dementia as key diagnostic groups of interest. It is important to note that the diagnosis of MCI and dementia is left censored (we do not know exactly when individuals would have first been diagnosed with MCI or dementia), which requires a strong linear assumption of decline within the dementia group and a comparison of MCI who progress to those who do not. Over a relatively short period of time (2–3 years), linear change is often a reasonable approximation to the amount of change, although researchers should consider whether this assumption is appropriate in their particular setting. Despite the limitations, the methodology does provide a framework for comparing

markers across a variety of modalities, including different imaging platforms, fluid biomarkers, and different cognitive function tests. The methodology is also easily adaptable to compare across criteria of interest. We present precision and validity criteria by which to compare the markers, but they are not the only ones possible, and the person-level components are not the only ones that may be chosen for a given set of criteria. Our proposed framework is not the only possible approach for comparison either, although it does provide a means for the simultaneous comparison of a larger number of potential markers. Bootstrapping is another option for assessing pairwise comparisons of markers. However, when the number of potential markers is large, such a strategy requires considerable computational effort. In some cases, regression modeling, such as logistic or linear regression, may be sufficient for identifying promising markers, although when markers under consideration are highly correlated, such a strategy may not be appropriate. To our knowledge, this project is the first to illustrate the comparison of performance across cognitive and imaging biomarkers that have been statistically normed to a common standard. The participants included in this analysis come from ADNI, which itself is a strength, because all participants were imaged and clinically and cognitively evaluated using a standardized protocol.

Although this study uses only a small subset of the available markers restricted to a single processing method of the MRI scans for illustration of the methodology, future analyses will compare markers across the imaging modalities (MRI and PET scans). Comparing fluid biomarkers may also be of interest, as would more targeted comparisons for a specific clinical endpoint or with a specific intervention in mind. Multiple labs have also provided volumetric data on the same regions, such as the hippocampus and the ventricles, using different processing methods. Additional future work may compare across the methods to determine if one approach performs better than others. Similarly, comparison of amyloid or tau burden quantified from lumbar punctures or amyloid or tau PET imaging is also of interest. From a statistical standpoint, this methodology can be extended readily into the unbalanced design setting, in which not all markers are available on all participants. Further extensions to account for censoring in a survival analysis setting and to enable the comparison of predictors of longitudinal change of cognitive function after adjustment for covariates and potential confounders should also be made.

ACKNOWLEDGMENTS

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

CONFLICT OF INTEREST STATEMENT

Drs. Harvey, Tosun, and Beckett receive research funding from the National Institutes of Health (NIH). Dr. Harvey serves as a Statistical Advisor for *PLOS ONE*. Dr. Weiner serves on editorial boards for *Alzheimer's & Dementia*, *Magnetic Resonance Imaging* and *Topics in Magnetic Resonance Imaging*. He has served on advisory boards for Acumen Pharmaceutical, Alzheimer's Disease Neuroimaging Initiative (ADNI), Alzheon, Inc., Biogen, Brain Health Registry, Cerecin, Dolby Family Ventures, Eli Lilly, Merck Sharp & Dohme Corp., National Institute on Aging (NIA), Nestle/Nestec, Patient-Centered Outcomes Research Institute /Patient Partnered Research Network (PCORI/PPRN), Roche, University of Southern California (USC), and NervGen. He has provided consulting to Baird Equity Capital, BioClinica, Cerecin, Inc., Cytox, Dolby Family Ventures, Duke University, Eisai, FUJIFILM-Toyama Chemical (Japan), Garfield Weston, Genentech, Guidepoint Global, Indiana University, Japanese Organization for Medical Device Development, Inc. (JOMDD), Medscape, Nestle/Nestec, NIH, Peerview Internal Medicine, Roche, T3D Therapeutics, USC, and Vida Ventures. He has acted as a speaker/lecturer to The Buck Institute for Research on Aging, China Association for Alzheimer's Disease (CAAD), Japan Society for Dementia Research, and Korean Dementia Society. He holds stock options with Alzheon, Inc., Alzeca, and Anven. The following entities have provided funding for academic travel: USC, NervGen, American Society of Functional Neuroradiology (ASFNR), and Clinical Trials on Alzheimer's Disease (CTAD) Congress. Dr. Jack is employed by Mayo Clinic. He receives no personal compensation from any commercial entity. He receives research support from NIH and the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic. Author disclosures are available in the supporting information.

CONSENT STATEMENT

All participants gave written, informed consent prior to participation through the local institutional review boards at the participating institutions.

ORCID

Danielle J. Harvey https://orcid.org/0000-0002-5367-0951

REFERENCES

1. Tay LX, Ong SC, Tay LJ, Ng T, Parumasivam T. Economic burden of Alzheimer's disease: a systematic review. *Value Health Reg Issues*. 2024;40:1-12. doi:10.1016/j.vhri.2023.09.008

2. Scott TJ, O'Connor AC, Link AN, Beaulieu TJ. Economic analysis of opportunities to accelerate Alzheimer's disease research and development. *Ann N Y Acad Sci*. 2014;1313:17-34. doi:10.1111/nyas.12417
3. van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in early Alzheimer's disease. *N Engl J Med*. 2023;388:9-21. doi:10.1056/NEJMoa2212948
4. Sims JR, Zimmer JA, Evans CD, et al. Donanemab in early symptomatic Alzheimer disease. *JAMA*. 2023;330:512. doi:10.1001/jama.2023.13239
5. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989;8:431-440. doi:10.1002/sim.4780080407
6. Becker R, Greig N. Alzheimer's disease drug development: old problems require new priorities. *CNS Neurol Disord Drug Targets*. 2008;7:499-511. doi:10.2174/187152708787122950
7. Cullen NC, Zetterberg H, Insel PS, et al. Comparing progression biomarkers in clinical trials of early Alzheimer's disease. *Ann Clin Transl Neurol*. 2020;7:1661-1673. doi:10.1002/acn3.51158
8. Rajan KB, Aggarwal NT, McAninch EA, et al. Remote blood biomarkers of longitudinal cognitive outcomes in a population study. *Ann Neurol*. 2020;88:1065-1076. doi:10.1002/ana.25874
9. Smith R, Cullen NC, Pichet Binette A, et al. Tau-PET is superior to phospho-tau when predicting cognitive decline in symptomatic AD patients. *Alzheimers Dement*. 2023;19:2497-2507. doi:10.1002/alz.12875
10. Marks JD, Syrjanen JA, Graff-Radford J, et al. Comparison of plasma neurofilament light and total tau as neurodegeneration markers: associations with cognitive and neuroimaging outcomes. *Alzheimers Res Ther*. 2021;13:199. doi:10.1186/s13195-021-00944-y
11. Chen Y-H, Lin R-R, Huang H-F, Xue Y-Y, Tao Q-Q. Microglial activation, tau pathology, and neurodegeneration biomarkers predict longitudinal cognitive decline in Alzheimer's disease continuum. *Front Aging Neurosci*. 2022;14:848180. doi:10.3389/fnagi.2022.848180
12. Planche V, Bouteloup V, Pellegrin I, et al. Validity and performance of blood biomarkers for Alzheimer disease to predict dementia risk in a large clinic-based cohort. *Neurology*. 2023;100:e473-e484. doi:10.1212/WNL.0000000000201479
13. Llano DA, Devanarayan P, Devanarayan V. CSF peptides from VGF and other markers enhance prediction of MCI to AD progression using the ATN framework. *Neurobiol Aging*. 2023;121:15-27. doi:10.1016/j.neurobiolaging.2022.07.015
14. Cullen NC, Leuzy A, Palmqvist S, et al. Individualized prognosis of cognitive decline and dementia in mild cognitive impairment based on plasma biomarker combinations. *Nat Aging*. 2020;1:114-123. doi:10.1038/s43587-020-00003-5
15. Fonseca CS, Baker SL, Dobyns L, Janabi M, Jagust WJ, Harrison TM. Tau accumulation and atrophy predict amyloid independent cognitive decline in aging. *Alzheimers Dement*. 2024;20:2526-2537. doi:10.1002/alz.13654
16. Pepe MS, Feng Z, Huang Y, et al. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol*. 2007;167:362-368. doi:10.1093/aje/kwm305
17. Huang Y, Gilbert PB. Comparing biomarkers as principal surrogate endpoints. *Biometrics*. 2011;67:1442-1451. doi:10.1111/j.1541-0420.2011.01603.x
18. Gabriel EE, Sachs MC, Gilbert PB. Comparing and combining biomarkers as principal surrogates for time-to-event clinical endpoints. *Stat Med*. 2015;34:381-395. doi:10.1002/sim.6349
19. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. *Int J Biostat*. 2014;10:99-121. doi:10.1515/ijb-2012-0052
20. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease. *Neurology*. 1984;34:939-939. doi:10.1212/WNL.34.7.939
21. Mohs RC, Knopman D, Petersen RC, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's disease assessment scale that broaden its scope. The Alzheimer's disease cooperative study. *Alzheimer Dis Assoc Disord*. 1997;11(suppl 2):S13-S21.
22. Folstein MF, Folstein SE, McHugh PR. Mini-mental state. *J Psychiatr Res*. 1975;12:189-198. doi:10.1016/0022-3956(75)90026-6
23. Rey A. L'examen clinique en psychologie. [*The Clinical Examination in Psychology*]. Presses Universitaries De France; 1958.
24. Jack CR, Bernstein MA, Borowski BJ, et al. Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative. *Alzheimers Dement*. 2010;6:212-220. doi:10.1016/j.jalz.2010.03.004
25. Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*. 2012;61:1402-1418. doi:10.1016/j.neuroimage.2012.02.084
26. Reuter M, Rosas HD, Fischl B. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*. 2010;53:1181-1196. doi:10.1016/j.neuroimage.2010.07.020
27. Jack CR, Wiste HJ, Weigand SD, et al. Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement*. 2017;13:205-216. doi:10.1016/j.jalz.2016.08.005
28. Voevodskaya O, Simmons A, Nordenskjöld R, et al. The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Front Aging Neurosci*. 2014;6:264. doi:10.3389/fnagi.2014.00264
29. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. 2nd. John Wiley & Sons; 1999.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Harvey DJ, Tosun D, Jack CR, Weiner M, Beckett LA; for the Alzheimer's Disease Neuroimaging Initiative. Standardized statistical framework for comparison of biomarkers: Techniques from ADNI. *Alzheimer's Dement*. 2024;20:6834–6843. https://doi.org/10.1002/alz.14160