Published in final edited form as:
J Alzheimers Dis. 2013 January 1; 36(3): 475–486. doi:10.3233/JAD-122212.

Improved design of prodromal Alzheimer's disease trials through cohort enrichment and surrogate endpoints

Eric A. Macklin[a,b], **Deborah Blacker**[b,c,d], **Bradley T. Hyman**[b,e], and **Rebecca A. Betensky**[a,f]

[a]Massachusetts General Hospital Biostatistics Center, Boston MA 02114 USA

[b]Harvard Medical School, Boston MA 02115 USA

[c]Department of Psychiatry, Massachusetts General Hospital, Boston MA 02114 USA

[d]Department of Epidemiology, Harvard School of Public Health, Boston MA 02115 USA

[e]Department of Neurology, Massachusetts General Hospital, Boston MA 02114 USA

[f]Department of Biostatistics, Harvard School of Public Health, Boston MA 02115 USA

Summary

Alzheimer's disease (AD) trials initiated during or before the prodrome are costly and lengthy because patients are enrolled long before clinical symptoms are apparent, when disease progression is slow. We hypothesized that design of such trials could be improved by: (1) selecting individuals at moderate near-term risk of progression to AD dementia (the current clinical standard) and (2) by using short-term surrogate endpoints that predict progression to AD dementia. We used a longitudinal cohort of older, initially non-demented, community-dwelling participants (n=358) to derive selection criteria and surrogate endpoints and tested them in an independent national data set (n=6,243). To identify a "mid-risk" subgroup, we applied conditional tree-based survival models to Clinical Dementia Rating (CDR) scale scores and common neuropsychological tests. In the validation cohort, a time-to-AD dementia trial applying these mid-risk selection criteria to a pool of all non-demented individuals could achieve equivalent power with 47% fewer participants than enrolling at random from that pool. We evaluated surrogate endpoints measureable over two years of follow-up based on cross-validated concordance between predictions from Cox models and observed time to AD dementia. The best performing surrogate, rate of change in CDR sum-of-boxes, did not reduce the trial duration required for equivalent power using estimates from the validation cohort, but alternative surrogates with better ability to predict time to AD dementia should be able to do so. The approach tested here might improve efficiency of prodromal AD trials using other potential measures and could be generalized to other diseases with long prodromal phases.

Keywords

Clinical Trials as Topic; Surrogate Endpoint; Alzheimer's Disease; Survival Analysis; National Alzheimer's Coordinating Center Uniform Data Set

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder of late life with an insidious onset [1]. Autopsy and imaging studies suggest that underlying disease pathology

Corresponding author: Eric A. Macklin, PhD, MGH Biostatistics Center, Massachusetts General Hospital, 50 Staniford St, Suite 560, Boston, MA 02114 USA, +1 617-724-9828 phone, +1 617-724-9878 fax, emacklin@partners.org.

begins to develop years, even decades, before the development of clinical dementia [2]. Currently available therapies achieve only modest reductions in rates of cognitive decline [3], and none appears to modify the underlying disease process [4]. The greatest therapeutic opportunity may occur before patients develop dementia or even mild cognitive impairment [5, 6]. In response, increasing attention is being given to prevention trials [7, 8], but the long prodromal phase of AD makes brief clinical trials impractical [9]. Recent advances in neuroimaging offer a potential means to identify individuals with underlying plaque deposition [10], but it is not yet known whether all such individuals will develop clinical dementia or in what time frame. Thus, progression to overt AD dementia—while clearly an arbitrary threshold along an underlying continuous process—remains the standard, clinically definitive endpoint, as well as the one most relevant to patients and clinicians. Draft guidance from the FDA identifies time to AD dementia as an appealing efficacy measure but also suggests that use of cognitive and functional endpoints may allow for more efficient trial designs [11].

Following Schneider's [12] recommendation to use existing data sets to develop improved approaches to prodromal AD trials, we analyzed data from an existing cohort of individuals followed annually for progression to AD. Our goal was to optimize the design of secondary prevention trials in AD in two complementary ways: (1) *cohort enrichment*, defining inclusion criteria for future trials in terms of baseline assessments of common measures of cognition and functional status that would enrich trial cohorts with non-demented individuals who are at moderate risk of progressing to AD dementia within the duration of a typical clinical trial (e.g., 2–3 years), and (2) short-term *surrogate endpoints*, identifying measures of progression in these same cognitive and functional assessments that predict future progression to AD dementia.

Methods

Data sources

Separate training and validation data sets were used to develop and then test proposed cohort enrichment criteria and surrogate endpoints. Training data were obtained from a longitudinal cohort of initially non-demented individuals in work funded by a program project grant (PPG) and used in a number of previous studies (e.g., [13–17]). Data for validation were obtained from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS; see [18, 19]). The original studies were approved by the institutional review boards of participating sites and abided by conventions of the Helsinki Declaration. All participants provided written informed consent.

Training Set – PPG Cohort—The PPG cohort (n = 358) was recruited at Massachusetts General Hospital between 1992 and 2005 from community-dwelling participants age 65 or older who were non-demented, free of significant underlying medical, neurological, or psychiatric illness, and had a Clinical Dementia Rating (CDR; [20, 21]) based on an augmented interview sensitive to subtle cognitive symptoms [13] of 0 (n = 117, 33%) or 0.5 (n = 241, 67%) at baseline. Those with CDR 0.5 were over-sampled to provide more data on the prodromal phase of AD. Data from a medical and psychiatric interview, physical examination, EKG, standard laboratory tests, MRI, genetic analysis, and a battery of functional and neuropsychological tests were collected at baseline. Our analyses of the PPG cohort used the CDR global rating, the sum-of-boxes (CDR-SB), and ratings on the 6 individual components (memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care); a memory Z score calculated from total learning scores on the California Verbal Learning Test (CVLT, [22]) and free recall scores on the Free and Cued Selective Reminding Test (SRT, [23]), normalized to baseline scores

for these measures among participants rated CDR 0; and all 6 neuropsychological assessments shared with the NACC-UDS validation sample: the Mini-Mental State Exam (MMSE, [24]), time to complete parts A and B of the Trail Making Test [25], Digit Span Forward and Backward [26], and Controlled Word Association Test for animals [27]. Scores at the baseline visit on these measures were used to identify optimal criteria for cohort enrichment.

PPG participants were followed through Feb 2007 at roughly annual visits with assessments of medical history, functional and cognitive status, bedside neuropsychological tests, and semi-structured interviews. Progression in CDR and bedside tests over the first 2 years (± 0.5 years) after enrollment were evaluated as possible surrogate endpoints. We were restricted to surrogates that could be evaluated in only 2 years by the short mean follow-up in the validation sample. For each measure, we considered both the last observed value within the 2-year window and the slope or average linear change over time from baseline to two years. Progression to probable AD dementia, our gold-standard endpoint, was evaluated clinically at each annual visit based on the DSM-IV definition of dementia and standard research criteria for probable AD [28].

Validation Set – NACC-UDS Cohort—The NACC-UDS data set includes ongoing data collection beginning in 2005 from 27 Alzheimer's Disease Centers located throughout the US [18, 19]. Participants in the NACC-UDS are enrolled by referral from clinicians and family members, by self-referral, by active recruitment through community organizations, and, in some centers, from pre-existing memory and aging study cohorts. All centers enroll subjects with memory complaints and as well as cognitively healthy subjects, but centers differ in the distribution of types and severities of impairment of enrolled subjects. The validation data extracted from the September 2011 freeze of the full NACC-UDS data set consisted of 6,243 individuals age 65 years or older with CDR ratings of 0 (n = 3,980, 64%) or 0.5 (n = 2,263, 36%) who were free of dementia at their baseline visit and had completed at least one follow-up visit. Data contributed by the MGH center were excluded as many PPG participants were subsequently followed in the NACC-UDS sample, and data from one additional NACC site were excluded due to a reported problem with the integrity of neuropsychological data collection during the time interval sampled for this study. Baseline visits for the subsample of the NACC-UDS cohort used in these analyses occurred between Aug 2005 and Sep 2010 with roughly annual follow-up visits recorded through September 2011. The CVLT and SRT were not available in the NACC data; instead a memory Z-score was calculated from the delayed logical memory test score (Logical Memory IIA – Delayed; [29]), standardized against baseline scores of participants rated CDR 0. Comparisons of the PPG training sample and the NACC-UDS validation sample are given in Table 1.

Cohort enrichment

We identified criteria for cohort enrichment in the PPG cohort using tree-based models applied to the 15 baseline variables shared with the NACC-UDS validation sample. Tree-based models assign membership in risk strata using recursive partitioning down a hierarchy of "branches." Tree-based methods are particularly useful for defining cohort selection criteria because their construction identifies stratifying criteria explicitly. Individual trees were fit using a conditional tree-based survival model [30]. Trees for predicting risk of progression to probable AD dementia were grown by selecting the predictor and split that yielded the minimum p-value from a log-rank test. Additional predictors and splits were included until a Bonferroni correction indicated that no additional model complexity yielded significant improvement in fit. Use of this stopping criterion avoids over-fitting of trees without a need for pruning [31].

Criteria for cohort enrichment were obtained by excluding terminal nodes predictive of the longest and shortest median times to probable AD dementia. Individuals who were diagnosed during follow-up with dementias other than probable AD were considered censored for risk of probable AD dementia at the first time they were diagnosed as demented. Nodes with the longest median time to probable AD dementia were excluded to minimize enrollment of individuals with limited risk of progressing to a diagnosis of probable AD dementia during a short-term trial. Nodes with the shortest median times to probable AD dementia were excluded for several reasons: (1) individuals with high short-term risk might be beyond the reach of early intervention due to substantial underlying pathology; (2) they might have too little time to respond to a therapy prior to crossing the threshold for dementia, and (3) they might be so close to the threshold that their status as non-demented at baseline might be unreliably assessed. The remaining "*mid-risk*" individuals would be expected to have the greatest potential for well-defined progression and potentially modifiable risk of developing AD dementia in a short-term trial.

Surrogate endpoints

We used Cox proportional hazards regression to identify surrogate endpoints evaluated after two years of follow-up that best predicted subsequent progression to AD dementia among members of the mid-risk stratum of the PPG cohort. Time at risk of progression to AD dementia was modeled as beginning after the 2-year assessment. To ensure that surrogates were evaluated only with regard to their potential response to treatment, we included the best fitting baseline predictors in all Cox models as covariates.

Prediction accuracy of models was measured using Harrell's c-index of rank concordance [32]. Concordance was estimated by 10-fold cross-validation to best estimate prediction accuracy and limit over-fitting to the data. We took the median concordance from 20 replicate cross-validated concordance estimates obtained by random permutations of the data set to better stabilize the estimates. Our final measure for a given surrogate aimed to show how it could improve prediction of AD dementia over baseline variables by measuring the difference in median cross-validated concordance of the model containing the surrogate minus the median cross-validated concordance of the best baseline model. All models included age, gender, and years of education. Symmetric 90% confidence bounds were estimated as simple percentiles of the distribution of these differences from 100 bootstrap replicates. Additional details describing our evaluation of potential surrogates are given in the supplementary data.

Validation

The generalizability of the baseline risk stratification for cohort enrichment based on the survival tree analysis of the PPG cohort was assessed by comparing the estimated survival curves for progression to AD dementia between the PPG and NACC-UDS baseline risk cohorts. The generalizability of surrogate endpoints identified from analysis of the PPG cohort was assessed by comparing their cross-validated concordance estimates and by comparing the Cox model coefficients with those from the NACC cohort. Note that comparisons of concordance between PPG and NACC-UDS cohorts provide only an approximate comparison of prediction accuracy due to differences in the censoring distributions between the two cohorts [33].

Power evaluation

The potential benefit of using the proposed cohort enrichment criteria and surrogate endpoint in a future clinical trial was evaluated with respect to the reduction in required sample size or trial duration relative to a time-to-AD dementia trial designed for equivalent power under two choices of source population: (1) all participants who were non-demented

at baseline, and (2) only those who were non-demented and CDR 0.5 at baseline. In each case, our proposed mid-risk criteria were used to define a subset of the larger non-demented or non-demented and CDR 0.5 populations. Parameters used in calculating sample size requirements were drawn from the NACC-UDS cohort. Three-year trials (including one year to recruit subjects) were considered in order to focus on designs that would be feasible within a normal funding and drug development timeframe.

Effect of cohort enrichment on required sample size—The potential benefit of our proposed cohort enrichment criteria was estimated under the following assumptions: (1) the proportion of potential participants in a future trial who would be eligible under our mid-risk criteria would be equal to that found in NACC-UDS cohort, (2) time to progression to AD dementia would be Weibull distributed with parameters equal to those observed in the NACC-UDS cohort, (3) a proportion of participants equal to the 2-year Kaplan-Meier product-limit estimate of the probability of progression to AD dementia in each risk stratum of the NACC-UDS cohort would not respond to the treatment (i.e., those with non-modifiable or non-detectable risk of progression), and (4) treatment would reduce the hazard of progression to AD dementia for the remaining participants by a fixed hazard ratio. Sample size requirements were calculated for a log-rank test comparing specified survival distributions given parameters estimated from the NACC-UDS cohort assuming 80% power, two-tailed testing at alpha = 0.05, one-year of constant accrual, 1:1 randomization, and a pre-specified treatment-dependent hazard ratios: 0.50 or 0.67 (i.e., a 1/2 or 1/3 reduction in risk; cf. [34]). Additional details describing power calculations are given in the supplementary data.

Effect of surrogate endpoints on required follow-up time—Assuming one year to complete accrual and surrogates measured over 2 years of follow-up, a proposed surrogate-endpoint trial could be completed in 3 years. For comparison, assuming designs with 80% power, we determined (1) the trial duration required for time-to-AD dementia designs with the same sample size and treatment effect, and (2) the minimum detectable effect of treatment on the surrogate given the sample size required for a time-to-event trial with specified treatment effect and duration. For trials analyzing time to progression to AD dementia, we calculated power using estimates from the mid-risk stratum of the total non-demented population or of the CDR 0.5 non-demented population in the NACC-UDS cohort. For trials analyzing surrogate endpoints, we calculated power for a simple t-test of the surrogate endpoint. For simplicity, we assumed that the percent reduction in the hazard due to treatment could be compared to the percent reduction in the surrogate mean, recognizing that the two endpoints are measured on different scales.

Results

Cohort characteristics

While the NACC-UDS cohort was selected to match the PPG eligibility criteria for age and lack of dementia at baseline, the NACC-UDS participants tended to be older, more racially diverse, slightly less well educated, and somewhat less impaired (based on the fraction with a CDR rating of 0.5, reflecting the oversampling of such individuals in the PPG cohort)s (Table 1). MMSE scores, digit spans, Trail making B completion times, and memory Z-scores were slightly worse among NACC-UDS participants. PPG participants were followed for over 6 years on average, twice as long as the NACC-UDS cohort as of September 2011. In the total non-demented population, the hazard of progression to AD dementia was comparable (hazard ratio (HR) = 0.93, p = 0.59). The mean rate of increase in CDR-SB was only half or a sixth as large in the NACC-UDS cohort relative to the PPG cohort for the total non-demented and CDR 0.5 only non-demented populations, respectively. Conversely,

MMSE scores declined more rapidly on average in the NACC-UDS cohort. Lower neuropsychological scores and more rapid decline despite much slower changes in CDR-SB scores among the NACC-UDS cohort may reflect scoring differences on the CDR between the two samples, at least at the low end of the scale, or educational differences beyond those measured by years of education, or both.

Cohort enrichment

Stratification of the PPG participants identified four risk strata based on participants' baseline CDR-SB and composite memory Z-scores (Figure 1). Forty-six percent of all non-demented NACC-UDS participants were classified as mid-risk based on criteria of CDR-SB ≤ 1 and memory Z-score ≤ −0.428 or CDR-SB > 1 and memory Z-score > −2.00 (Table 2). The NACC-UDS and PPG cohorts had nearly equal risk of progression in the mid-risk stratum (HR = 1.13 for NACC-UDS relative to PPG hazard, 95% CI 0.81 to 1.59, p = 0.47), reflecting good generality of the PPG-defined mid-risk criteria. In the total non-demented NACC-UDS population, participants in the low-risk stratum had 3-yr cumulative rates of progression to probable AD dementia of 1.4%. Inclusion of such individuals in any short-term trial would add almost nothing to estimates of treatment effect given the very low event rates even in the absence of treatment. Conversely, NACC-UDS participants in the high-risk stratum with CDR-SB > 1 and memory Z-score ≤ −2.00 had substantial risk of rapid progression to probable AD dementia (median 2.30 years, 95% CI 2.07 to 2.87 years). While power increases with more observed outcomes, all else being equal, the large percentage of participants progressing to probable AD dementia within a few years of baseline (40% in 2 years) suggests that many of these individuals might be too close to the threshold for dementia to reliably assess the transition or might have insufficient time to respond to therapy. The benefit to a trial of enrolling such individuals depends on the degree to which their progression can be cleanly detected and their disease is still amenable to the specific treatment under study.

For the total non-demented and the CDR 0.5 only non-demented populations, sample sizes of n = 5,214 and 2,308, respectively, would be required for 80% power to detect a treatment hazard ratio of 0.67 given the assumptions described in the Methods. If enrollment for a secondary prevention trial were restricted to the mid-risk stratum defined by our proposed cohort enrichment criteria, the required sample sizes would be n = 2,784 for the total non-demented population and 1,780 for the CDR 0.5 only non-demented population, reductions of 46 and 23%, respectively (Figure 2). If we remove the assumption that subjects within two years of the developing dementia do not benefit from treatment, then the required sample sizes for an unselected population vs. the mid-risk subset would be n = 3,598 and 2,402 (33% fewer), respectively, for the total non-demented population, and n = 1,408 and 1,370 (3% fewer) for the CDR 0.5 only non-demented population.

Surrogate endpoints

Among variables available in the PPG cohort, rate of change in CDR-SB and 2-year CDR-SB were found to be the best predictors of future progression to AD dementia in the total non-demented PPG mid-risk stratum based on baseline-adjusted increments in cross-validated predication accuracy (Table 3). In the NACC-UDS validation sample, concordance estimates of models based on rate of change in CDR-SB were significantly greater than those based on baseline data alone in the total non-demented mid-risk stratum (Table 3), but the increment was modest at 4.0% (90% CI 1.6 to 6.4%). In the mid-risk stratum of the CDR 0.5 only non-demented population, rate of change in CDR-SB also improved prediction of time to progression to AD dementia beyond what was identified at baseline, but the increment was still more modest (2.8%, 90% CI 0.0 to 6.6%). Of note, the estimated hazard ratio for a one-unit increase in change in CDR-SB per year in the NACC-

UDS cohort was much smaller than that in the PPG cohort in both the total non-demented population (NACC-UDS HR = exp(0.90) = 2.46 versus PPG HR = exp(3.31) = 27.4) and the CDR 0.5 only non-demented population (NACC-UDS HR = 1.99 vs. PPG HR = 9.58; Table 3). This may reflect site-to-site variation in CDR scoring among sites contributing to the NACC-UDS data set and may suggest that rate of change of CDR-SB is not an optimal surrogate endpoint, particularly when combining data from a wide number of sites.

A trial designed to detect a 33% reduction in the mean rate of change in CDR-SB over two years in the total non-demented NACC-UDS population (mean = 0.098 / yr) would require a sample size of n = 5,897 based on the observed standard deviation in rate of change in CDR-SB (SD = 0.418 / yr, Table 3) and an estimate that 6.6% of participants would not respond to treatment (based on the observed 2-year event rate in the total non-demented NACC-UDS population). Assuming one year for accrual, the full surrogate trial would require 3 years to complete, although follow-up of individual participants for the purpose of estimating the surrogate was assumed to end after their two year assessment. To achieve 80% power in a time-to-event trial with n = 5,897 given the assumptions above would require only 2.1 years of total follow-up (29% shorter, Figure 3A). Even if all participants could be enrolled at the start of the trial and a surrogate trial would be completed in 2 years, a time-to-event trial of the same sample size powered for a 67% hazard ratio would require only 1.7 years, 15% shorter than the surrogate design. Within the CDR 0.5 only non-demented population, time-to-event trials would be even shorter than trials based on change in CDR-SB when designed with the same sample size and power and equating percent reduction in hazard and surrogate mean as measures of treatment effect (Figure 3B). These estimates reflect the limited statistical power of slope measures, particularly for noisy data, and the limited latitude for detecting any benefit of a given treatment under our assumption that the rate of change in CDR-SB can at best be reduced to zero. The utility of a surrogate trial based on CDR-SB would be greater if it were plausible that treatment could lead to negative slopes, i.e., improvements in cognitive function, not just reductions in the rate of cognitive decline.

These estimates also depend on the unknowable assumption of equal effect of treatment on the surrogate and on time to AD dementia. If the treatment effect on the surrogate were equal to an effect size of at least $\delta = 0.22$, then a surrogate trial could be completed in half the duration (3 years vs. 6 years) of an equivalent time-to-event trial powered to detect a 33% reduction in hazard of progression to AD dementia (Figure 3C). Given the higher event rates in the mid-risk stratum of the CDR 0.5 only non-demented NACC-UDS population and thus the greater relative efficiency of time-to-event trials in this population, the minimum required effect size for a surrogate trial requiring half the duration of an equivalently powered time-to-event trial would be $\delta = 0.28$ for treatment effects that reduced hazard by 33% (Figure 3D).

Discussion

Significant advances are being made in elucidating the pathophysiology of AD [35], but effective prevention and treatment of AD are slow to emerge. Greater efficiency in the conduct of AD trials will enhance our ability to evaluate a wide spectrum of possible therapies. Given the current understanding that the greatest benefit from some treatment strategies may be achieved by initiating therapies early in the disease process, well before progression to dementia, much of the focus of AD clinical trials is on early intervention. Our analysis suggests that reductions in sample size requirements can be achieved for such trials through use of well-guided criteria for enriching the trial cohort with participants who are at risk of progression but still responsive to treatment. These individuals are exactly those who are far enough from the “tipping point” to respond clearly to therapy. In principle, we believe that trials could be shortened by evaluating efficacy on the basis of surrogate

endpoints that predict eventual progression to AD dementia over shorter time-spans, but the surrogates identified in this study from the short list available for validation would not substantively shorten a trial drawing from the NACC-UDS cohort.

Not surprisingly, we found that memory scores and CDR measures were useful metrics in defining cohort enrichment criteria. Baseline risk stratification based on CDR-SB and a memory Z-score derived either from the SRT and CVLT or from delayed recall on a logical memory test consistently identified a mid-risk stratum of patients in two data sets. Under the proposed mid-risk criteria, 45% to 65% of potential non-demented participants or CDR 0.5 only non-demented participants would be eligible for enrollment, but the sample size required for 80% power would be reduced by 47% and 23%, respectively. We have previously suggested that recognition of graded severity across the spectrum of mild cognitive impairment might improve the design of early treatment trials [17]. Our mid-risk criteria broaden the target population for prodromal AD trials beyond those meeting criteria for MCI or with a CDR rating of 0.5 to include a subset of subjects with normal cognition but relatively poor performance on memory tests and excludes more severely impaired subjects from the CDR 0.5 group.

We did not identify useful surrogates among the measures available for analysis in the PPG cohort and validation in the NACC-UDS data set. Based on event rates in the total non-demented NACC-UDS population, a surrogate trial 2-year rate of change in CDR-SB would require n = 5,897 participants for 80% power to detect a 33% reduction in mean rate of change and would require more time to complete than an equivalent time-to-AD trial, even assuming full accrual at study initiation. Coley et al. [36] recommend change in CDR-SB as an outcome measure in clinical trials based on good responsiveness in a population of patients with AD dementia, even in those rated CDR 0.5. The FDA endorsed CDR-SB in their draft guidance on early stage AD trials as a well-validated and reliable measure for the longitudinal assessment of patients with cognitive and functional deficits that do not rise to the level of a diagnosis of overt dementia [11]. We also observed that CDR-SB is the best surrogate among those we considered, but our analysis of trial duration suggests that CDR-SB is not a useful surrogate in a population of non-demented subjects due to substantial floor effect and coarse granularity of the scale, particularly in its lower range. Rate of change in neuropsychological test scores, particularly in memory, might have performed better had they been available for analysis. Use of item response theory to define an internally consistent scoring for such instruments would also likely have improved performance of such measures. Measures of change in beta-amyloid plaques, as proposed for the Anti-Amyloid Treatment in Asymptomatic AD (A4) Trial, or other physiological biomarkers might also have been superior surrogates had those data been available. Based on estimates from the NACC-UDS cohort, even modest effects of treatment on a responsive surrogate could reduce the required duration of a trial. Fleming and DeMets discuss a number of situations where treatments may differentially affect the proposed surrogate and the clinical endpoint [37]. Surrogate effect sizes of only $\delta = 0.1$ to 0.2 would be required for good power and substantial reduction in trial duration relative to a time-to-event trial designed to detect a hazard ratio of 0.67.

Limitations

We are encouraged by the consistency in results between our training and validation samples in the identification of baseline risks of progression to AD dementia. Nevertheless, our study has limitations. Neither cohort is an unbiased sample of patients at risk for AD dementia. While the NACC-UDS cohort does not reflect the distribution of mild symptoms in the community, it has the advantage of actually being used to recruit for AD trials and offered consistent, structured assessments for multiple measures shared with the PPG cohort. The thresholds for differentiating low-, mid-, and high-risk strata were selected by the survival-

tree models on the basis of maximal log-rank statistics, while optimal thresholds will depend on the relationship between risk of rapid progression and potential for therapeutic efficacy. Overall, our exploration of cohort enrichment criteria and possible surrogate endpoints was limited by short follow-up in the NACC-UDS validation cohort, the short list of cognitive and functional variables measured in both cohorts, the lack of neuroimaging or cerebrospinal fluid biomarkers in either cohort, and the lack of serial neuropsychological testing in the PPG cohort. In particular, the PPG cohort lacked serial memory assessments, which would be expected to perform better as surrogates than the CDR-SB and its components. We also chose to restrict ourselves to easily obtained measures from the primary care setting and did not include variables such as APOE genotype that might not be widely available. Note as well that our model of times to AD dementia assumes continuous, Weibull-distributed event times, but time to AD dementia may not follow a Weibull distribution and diagnosis of progression to AD dementia requires a clinical evaluation and thus happens discretely in time with events clustered around scheduled follow-up visits. With annual follow-up, we will slightly over-estimate power using the assumptions described. Also important in limiting the estimated benefit of a surrogate was our assumption that treatment can slow but not reverse cognitive decline. Adaptive clinic trials (cf. [38–40]) might provide a mechanism for collecting the necessary data on treatment mediation by proposed surrogates during an initial phase of a larger efficacy trial.

In spite of these caveats, we feel that the proposed cohort enrichment criteria could help accelerate development of AD therapies. Our analysis presents a general approach to improving clinical trials through better targeted enrollment and use of short-term endpoints that effectively reflect treatment effects on an accepted clinical endpoint. While we applied the technique to a relatively small sample of functional and neurocognitive measures, the approach can be directly generalized to other measures, e.g., Rasch scores of memory assessments, other symptom scales, serum biomarkers, gene expression, and imaging measures. Moreover, this same approach could be extended to other diseases with long prodromal stages where primary prevention trials with time-to-event endpoints are prohibitively large due to low incidence rates in the general population, e.g., Parkinson's disease, amyotrophic lateral sclerosis, and schizophrenia.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants P01 AG004953, P50 AG005134-24S1, and U01 AG016976 from the National Institute on Aging, R01 CA075971 from the National Cancer Institute, and funds from the Harvard NeuroDiscovery Center. We thank the editor and reviewers for constructive suggestions for improvements to the manuscript.

References

1. Jack CR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011; 7:257–262. [PubMed: 21514247]
2. Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 2010; 9:119–128. [PubMed: 20083042]
3. American Psychiatric Association. Practice guideline for the treatment of patients with Alzheimer's disease and other dementias. American Psychiatric Association Arlington; VA: 2007.

4. Holmes C, Boche D, Wilkinson D, Yadegarfar G, Hopkins V, Bayer A, Jones RW, Bullock R, Love S, Neal JW, Zotova E, Nicoll JA. Long-term effects of Abeta42 immunisation in Alzheimer's disease: follow-up of a randomised, placebo-controlled phase I trial. Lancet. 2008; 372:216–223. [PubMed: 18640458]
5. Blennow K. Biomarkers in Alzheimer's disease drug development. Nat Med. 2010; 16:1218–1222. [PubMed: 21052077]
6. Emery VO. Alzheimer disease: are we intervening too late? Pro. J Neural Transm. 2011; 118:1361–1378. [PubMed: 21647682]
7. Feldman HH, Jacova C. Primary prevention and delay of onset of AD/dementia. Can J Neurol Sci. 2007; 34(Suppl 1):S84–89. [PubMed: 17469689]
8. Sano M, Grossman H, Van Dyk K. Preventing Alzheimer's disease : separating fact from fiction. CNS Drugs. 2008; 22:887–902. [PubMed: 18840031]
9. Plassman BL, Langa KM, McCammon RJ, Fisher GG, Potter GG, Burke JR, Steffens DC, Foster NL, Giordani B, Unverzagt FW, Welsh-Bohmer KA, Heeringa SG, Weir DR, Wallace RB. Incidence of dementia and cognitive impairment, not dementia in the United States. Ann Neurol. 2011; 70:418–426. [PubMed: 21425187]
10. Vandenberghe R, Van Laere K, Ivanoiu A, Salmon E, Bastin C, Triau E, Hasselbalch S, Law I, Andersen A, Korner A, Minthon L, Garraux G, Nelissen N, Bormans G, Buckley C, Owenius R, Thurfjell L, Farrar G, Brooks DJ. 18F-flutemetamol amyloid imaging in Alzheimer disease and mild cognitive impairment: a phase 2 trial. Ann Neurol. 2010; 68:319–329. [PubMed: 20687209]
11. Food and Drug Administration. [Accessed 23 February 2013] Guidance for industry: Alzheimer's disease: Developing drugs for the treatment of early stage disease. Feburary. 2013 Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM338287.pdf
12. Schneider LS. The potential and limits for clinical trials for early Alzheimer's disease and some recommendations. J Nutr Health Aging. 2010; 14:295–298. [PubMed: 20305999]
13. Daly E, Zaitchik D, Copeland M, Schmahmann J, Gunther J, Albert M. Predicting conversion to Alzheimer disease using standardized clinical information. Arch Neurol. 2000; 57:675–680. [PubMed: 10815133]
14. Albert MS, Moss MB, Tanzi R, Jones K. Preclinical prediction of AD using neuropsychological tests. J Int Neuropsychol Soc. 2001; 7:631–639. [PubMed: 11459114]
15. Albert M, Blacker D, Moss MB, Tanzi R, McArdle JJ. Longitudinal change in cognitive performance among individuals with mild cognitive impairment. Neuropsychology. 2007; 21:158–169. [PubMed: 17402816]
16. Blacker D, Lee H, Muzikansky A, Martin EC, Tanzi R, McArdle JJ, Moss M, Albert M. Neuropsychological measures in normal individuals that predict subsequent cognitive decline. Arch Neurol. 2007; 64:862–871. [PubMed: 17562935]
17. Dickerson BC, Sperling RA, Hyman BT, Albert MS, Blacker D. Clinical prediction of Alzheimer disease dementia across the spectrum of mild cognitive impairment. Arch Gen Psychiatry. 2007; 64:1443–1450. [PubMed: 18056553]
18. Morris JC, Weintraub S, Chui HC, Cummings J, Decarli C, Ferris S, Foster NL, Galasko D, Graff-Radford N, Peskind ER, Beekly D, Ramos EM, Kukull WA. The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. Alzheimer Dis Assoc Disord. 2006; 20:210–216. [PubMed: 17132964]
19. Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, Hubbard JL, Koepsell TD, Morris JC, Kukull WA. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Dis Assoc Disord. 2007; 21:249–258. [PubMed: 17804958]
20. Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry. 1982; 140:566–572. [PubMed: 7104545]
21. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]
22. Delis, D.; Kramer, J.; Kaplan, E.; Ober, B. The California Verbal Learning Test. The Psychological Corporation; New York, NY: 1987.

23. Grober E, Buschke H. Genuine memory deficits in dementia. Developmental Neuropsychology. 1987; 3:13–36.
24. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975; 12:189–198. [PubMed: 1202204]
25. Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. Perceptual and Motor Skills. 1958; 8:271–276.
26. Wechsler, D. The Wechsler Adult Intelligence Scale–Revised. The Psychological Corporation; New York, NY: 1988.
27. Benton, A.; Hamsher, K. Multilingual Aphasia Examination. University of Iowa Press; Iowa City, IA: 1976.
28. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer’s disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease. Neurology. 1984; 34:939–944. [PubMed: 6610841]
29. Wechsler, D. Wechsler Memory Scale. 3. The Psychological Corporation; San Antonio, TX: 1997.
30. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics. 2006; 15:651–674.
31. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008; 9:307. [PubMed: 18620558]
32. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982; 247:2543–2546. [PubMed: 7069920]
33. Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 30:1105–1117. [PubMed: 21484848]
34. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics. 1988; 44:229–241. [PubMed: 3358991]
35. Blennow K, de Leon MJ, Zetterberg H. Alzheimer’s disease. Lancet. 2006; 368:387–403. [PubMed: 16876668]
36. Coley N, Andrieu S, Jaros M, Weiner M, Cedarbaum J, Vellas B. Suitability of the Clinical Dementia Rating-Sum of Boxes as a single primary endpoint for Alzheimer’s disease trials. Alzheimers Dement. 2011; 7:602–610. e602. [PubMed: 21745761]
37. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996; 125:605–613. [PubMed: 8815760]
38. Freidlin B, Korn EL. Biomarker-adaptive clinical trial designs. Pharmacogenomics. 2010; 11:1679–1682. [PubMed: 21142910]
39. Cook T, DeMets DL. Review of draft FDA adaptive design guidance. J Biopharm Stat. 2010; 20:1132–1142. [PubMed: 21058109]
40. Renfro LA, Carlin BP, Sargent DJ. Bayesian adaptive trial design for a newly validated surrogate endpoint. Biometrics. 2011

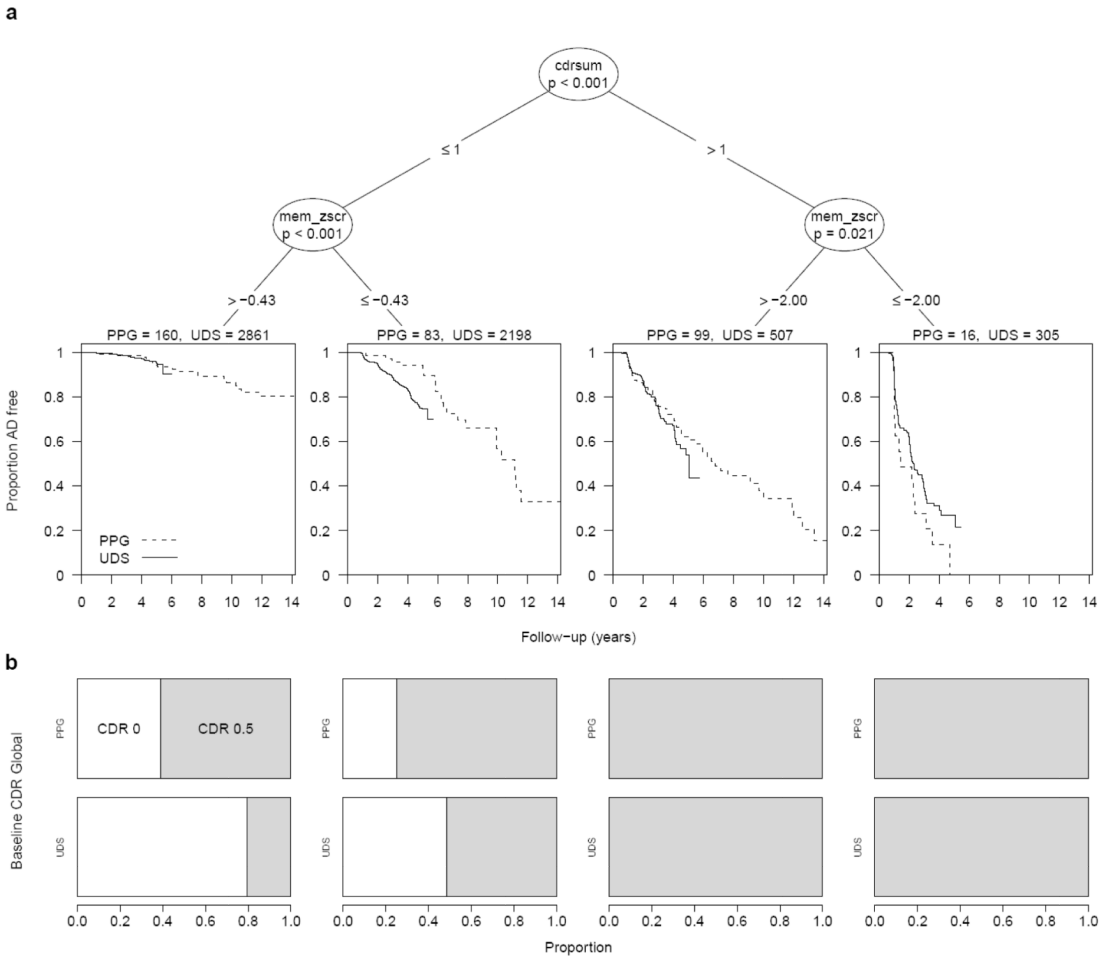

Figure 1.

(A) Baseline risk stratification of the PPG cohort using shared baseline variables. Terminal nodes overlay Kaplan-Meier survival curves for freedom from probable AD for the PPG cohort (dashed line) and for the NACC-UDS validation cohort (solid line). (B) The distribution of CDR Global scores at baseline (0 = white, 0.5 = shaded) in the PPG and NACC-UDS cohorts for each of the four risk strata defined by the shared-variable survival tree model.

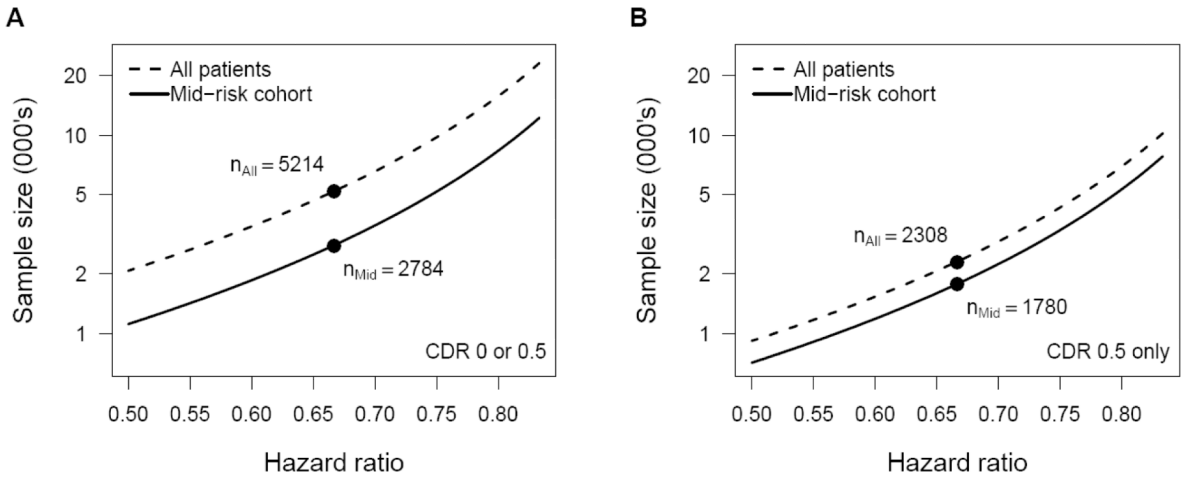

Figure 2.

Sample sizes required for clinical trials with 80% power to detect a treatment effect on risk of progression to probable AD dementia by Cox regression based on survival curves estimated from the NACC-UDS cohort for all non-demented subjects (A) and CDR 0.5 non-demented subjects only (B) comparing all subjects (dashed line) with those meeting the PPG-defined mid-risk criteria (solid line). Estimates assume 1:1 randomization, one year accrual, two years minimal follow-up, and two-tailed testing at alpha = 0.05.

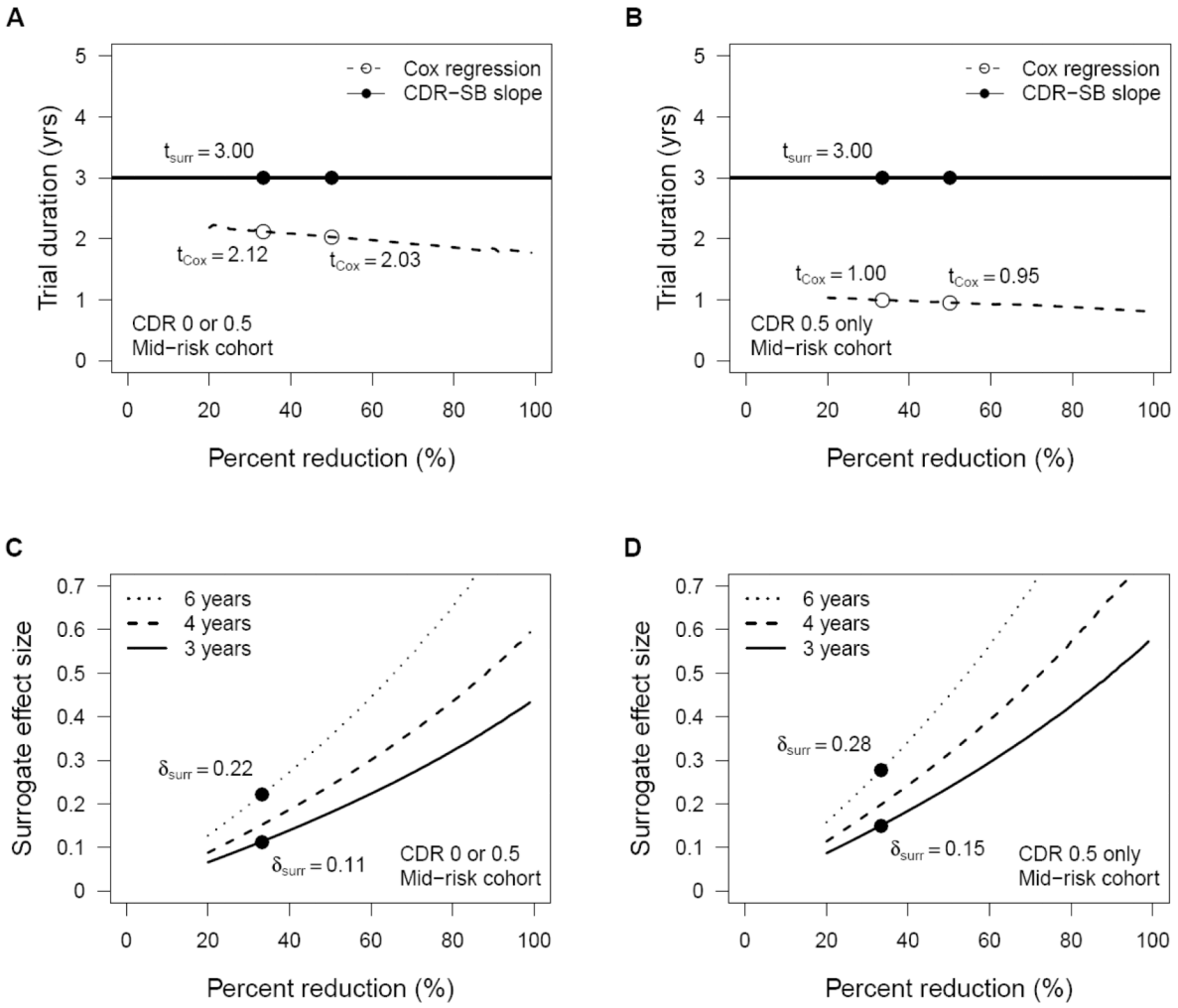

Figure 3.

Trial duration (A and B) and minimum surrogate treatment effects (C and D) required for 80% power in all non-demented subjects (A and C) and CDR 0.5 only non-demented subjects (B and D) populations. Estimates assume 1:1 randomization, one year accrual, and two-tailed testing at alpha = 0.05.

J Alzheimers Dis. Author manuscript; available in PMC 2013 September 23.

Table 1

Cohort characteristics.

Variable	All non-demented				Only CDR 0.5 non-demented			
	PPG (mean±SD (range) or % (N))	NACC-UDS (mean ±SD (range) or % (N))	Cohort comparison (estimate, 95% CI)	P-value	PPG (mean±SD (range) or % (N))	NACC-UDS (mean ±SD (range) or % (N))	Cohort comparison (estimate, 95% CI)	P-value
N	358	6243			241	2263		
Demographics								
Age (yrs)	73.5±5.2 (65,88)	76.9±7.3 (65,110)	3.39 (2.83,3.96)[1]	<.001	74.1±5.3 (65,88)	77.4±7.2 (65,110)	3.33 (2.60,4.07)[1]	<.001
Male	44.1% (158)	41.1% (2568)		0.26	44.8% (108)	49.9% (1130)		0.13
Race				<.001				0.006
Asian	2.5% (9)	2.4% (150)			2.9% (7)	3.3% (74)		
Black	5.6% (20)	13.0% (809)			5.4% (13)	12.4% (281)		
White	91.3% (327)	82.5% (5151)			90.9% (219)	81.0% (1833)		
Other/Unknown	0.6% (2)	2.1% (131)			0.8% (2)	3.3% (75)		
Education (yrs)	15.5±2.9 (5,24)	15.2±3.3 (0,29)	−0.34 (−0.66,−0.03)[1]	0.033	15.5±3.0 (5,24)	14.8±3.7 (0,29)	−0.71 (−1.12,−0.31)[1]	<.001
Clinical Dementia Rating								
CDR rating Distribution:				<.001				N/A
0	32.7% (117)	63.8% (3980)			0% (0)	0% (0)		
0.5	67.3% (241)	36.2% (2263)			100% (241)	100% (2263)		
CDR-SB	0.90±0.89 (0.0,3.5)	0.47±0.81 (0.0,5.5)	−0.43 (−0.52,−0.33)[1]	<.001	1.33±0.77 (0.5,3.5)	1.26±0.89 (0.5,5.5)	−0.07 (−0.17,0.04)[1]	0.21
Memory rating Distribution:				<.001				0.88
0	33.0% (118)	64.6% (4034)			0.4% (1)	2.4% (54)		
0.5	59.8% (214)	31.1% (1939)			88.8% (214)	85.7% (1939)		
1	7.3% (26)	4.3% (267)			10.8% (26)	11.8% (267)		
2	0.0% (0)	0.0% (3)			0.0% (0)	0.1% (3)		
Neuropsych scores								
MMSE	29.2±1.1 (24,30)	28.3±2.0 (7,30)	−0.93 (−1.06,−0.81)[1]	<.001	29.1±1.2 (24,30)	27.3±2.4 (7,30)	−1.74 (−1.92,−1.56)[1]	<.001
Digit-span forwards	6.9±1.3 (4,9)	6.6±1.1 (0,8)	−0.33 (−0.47,−0.19)[1]	<.001	6.8±1.3 (4,9)	6.4±1.2 (0,8)	−0.44 (−0.61,−0.26)[1]	<.001
Digit-span backwards	5.3±1.4 (0,8)	4.7±1.2 (0,7)	−0.62 (−0.77,−0.47)[1]	<.001	5.2±1.4 (0,8)	4.4±1.2 (0,7)	−0.77 (−0.96,−0.58)[1]	<.001
Trail making B (sec)	103±53.1 (23,300)	114±65.9 (17,300)	11.2 (5.46,17.0)[1]	<.001	108±59.5 (32,300)	140±76.0 (22,300)	32.3 (24.1,40.5)[1]	<.001
CVLT total	48.3±11.5 (15,75)	N/A	N/A	.	46.6±11.7 (15,73)	N/A	N/A	.
SRT free	41.4±8.3 (10,59)	N/A	N/A	.	40.6±8.8 (10,58)	N/A	N/A	.

	All non-demented				Only CDR 0.5 non-demented			
Variable	**PPG (mean±SD (range) or % (N))**	**NACC-UDS (mean ±SD (range) or % (N))**	**Cohort comparison (estimate, 95% CI)**	**P-value**	**PPG (mean±SD (range) or % (N))**	**NACC-UDS (mean ±SD (range) or % (N))**	**Cohort comparison (estimate, 95% CI)**	**P-value**
Delayed logical memory	N/A	10.4±5.0 (0.0,24.0)	N/A	.	N/A	7.6±4.9 (0.0,23.0)	N/A	.
Memory Z-score	−.29±1.02 (−3.8,2.0)	−.38±1.17 (−2.8,2.8)	−0.09 (−0.20,0.02)[1]	0.102	−.43±1.06 (−3.8,1.7)	−1.0±1.15 (−2.8,2.5)	−0.61 (−0.76,−0.46)[1]	<.001
Follow-up characteristics								
2-yr timepoint to AD or censoring (yrs)	5.89±3.60 (0.8,12.1)	1.85±0.76 (0.4,4.0)	−4.03 (−4.50,−3.57)[1]	<.001	5.90±3.74 (0.8,12.1)	1.67±0.71 (0.4,3.6)	−4.23 (−4.83,−3.62)[1]	<.001
CDR-SB slope to 2-yr timepoint (/yr)	0.14±0.29 (−0.5,1.3)	0.06±0.32 (−2.0,3.8)	−0.08 (−0.12,−0.04)[1]	<.001	0.18±0.33 (−0.5,1.3)	0.03±0.47 (−2.0,2.6)	−0.15 (−0.21,−0.08)[1]	<.001
CDR-SB last obs 2-yr timepoint	1.23±1.11 (0.0,5.0)	0.65±1.31 (0.0,18.0)	−0.58 (−0.70,−0.46)[1]	<.001	1.70±1.02 (0.0,5.0)	1.55±1.78 (0.0,18.0)	−0.15 (−0.30,0.01)[1]	0.060
MMSE slope to 2-yr timepoint (/yr)	0.00±0.48 (−1.5,1.8)	−.10±0.86 (−3.9,10.6)	−0.10 (−0.17,−0.03)[1]	0.005	−.01±0.52 (−1.5,1.8)	−.19±1.11 (−3.4,10.6)	−0.18 (−0.30,−0.07)[1]	0.002
AD after 2-yr timepoint	25.5% (59)	5.5% (158)	1.15 (0.72,1.84)[3]	0.55	34.5% (51)	13.6% (103)	2.06 (1.26,3.37)[3]	0.004

[1] Mean difference = mean(NACC-UDS) − mean(PPG)

[2] Odds ratio = odds(NACC-UDS) / odds(PPG)

[3] Hazard ratio = hazard(NACC-UDS) / hazard(PPG)

Table 2

Baseline stratification, 3-yr cumulative rates of progression to probable AD dementia, and Weibull parameters for time to progression to probable AD dementia.

	Definition		PPG Cohort		NACC-UDS Cohort[1]			
Risk Stratum	**CDR sum of boxes**	**Memory Z-score**	**N (%)**	**3-yr cumulative event rate (%, 95% CI)**	**N (%)**	**3-yr cumulative event rate (%, 95% CI)**	**Weibull Shape (a)**	**Weibull Scale (b)**
All non-demented			358 (100%)	11.3 (8.4 to 15.2)	5871 (100%)	9.6 (8.7 to 10.5)	1.62 (1.51 to 1.73)	12.9 (11.6 to 14.3)
Low-risk	≤ 1	> −0.43	160 (45%)	1.3 (0.3 to 5.1)	2861 (49%)	1.4 (1.0 to 2.0)	1.86 (1.74 to 1.98)	27.2 (22.5 to 32.9)
Mid-risk	≤ 1 or > 1	≤ −0.43 > −2.00	182 (51%)	15.0 (10.4 to 21.4)	2705 (46%)	13.4 (11.9 to 15.1)		8.58 (7.90 to 9.32)
High-risk	> 1	≤ −2.00	16 (4%)	72.2 (49.0 to 91.2)	305 (5%)	61.7 (54.8 to 68.6)		3.09 (2.84 to 3.37)
CDR 0.5 only			241 (100%)	16.8 (12.5 to 22.3)	2139 (100%)	24.2 (22.0 to 26.5)	1.68 (1.56 to 1.81)	6.65 (6.16 to 7.19)
Low-risk	≤ 1	> −0.38	77 (32%)	2.7 (0.7 to 10.4)	473 (22%)	5.3 (3.3 to 8.4)	1.83 (1.70 to 1.96)	15.9 (12.4 to 20.3)
Mid-risk	≤ 1 or > 1	≤ −0.38 > −2.00	148 (61%)	18.4 (12.8 to 26.0)	1361 (64%)	22.9 (20.2 to 25.9)		6.32 (5.83 to 6.85)
High-risk	> 1	≤ −2.00	16 (7%)	72.2 (49.0 to 91.2)	305 (14%)	61.7 (54.8 to 68.6)		3.10 (2.84 to 3.39)

[1] Only participants with non-missing memory Z-scores are included. This excludes 248 CDR 0 and 124 CDR 0.5 NACC-UDS participants.

Table 3

Means, standard deviations, Cox regression coefficients, and corcordance estimates for CDR-SB slope (yr^{-1}), the proposed surrogate measure in the mid-risk stratum.

	PPG Cohort					NACC-UDS cohort				
		Cox regression		Concordance			Cox regression		Concordance	
Population	**Mean±SD**	**Beta (95% CI)**	**P value**	**Abs.[1]**	**Change[2] (90% CI[3])**	**Mean±SD**	**Beta (95% CI)**	**P value**	**Abs.[1]**	**Change[2] (90% CI[3])**
All non-demented	0.169±0.299	3.31 (2.08,4.53)	<.001	0.730	0.082 (0.014,0.160)	0.098±0.418	0.90 (0.62,1.17)	<.001	0.801	0.040 (0.016,0.064)
CDR 0.5 only	0.186±0.322	2.26 (1.17,3.35)	<.001	0.691	0.066 (0.005,0.139)	0.061±0.512	0.69 (0.33,1.04)	<.001	0.691	0.028 (−.000,0.066)

[1] Absolute concordance estimated from 100 bootstrap replicates of the median 10-fold cross-validated rank concordance from 20 random permutations of each bootstrap replicate.

[2] Change in median 10-fold cross-validated rank concordance above that of the best baseline model identified for a given bootstrap replicate.

[3] Simple percentiles from the bootstrap distribution of change in median 10-fold cross-validated rank concordance.