Published in final edited form as:
Neurobiol Aging. 2014 April ; 35(4): 808–818. doi:10.1016/j.neurobiolaging.2013.09.039.

Operationalizing hippocampal volume as an enrichment biomarker for amnestic MCI trials: effect of algorithm, test-retest variability and cut-point on trial cost, duration and sample size

P. Yu[1], **J. Sun**[1,2], **R. Wolz**[3,4], **D. Stephenson**[5], **J. Brewer**[6], **N.C. Fox**[7], **P.E. Cole**[1], **C.R. Jack Jr**[8], **D.L.G. Hill**[3,7], **A.J. Schwarz**[1,*], and **for the Coalition Against Major Diseases and the Alzheimer's Disease Neuroimaging Initiative†**

[1]Eli Lilly and Company, Indianapolis, IN, USA

[2]University of Texas, Houston, USA

[3]IXICO Ltd., London, UK

[4]Imperial College, London, UK

[5]Critical Path Institute, Tucson, AZ, USA

[6]University of San Diego, San Diego, CA, USA

[7]University College, London, UK

[8]Mayo Clinic, Rochester, MN, USA

Abstract

Objective—To evaluate the effect of computational algorithm, measurement variability and cut-point on hippocampal volume (HCV)-based patient selection for clinical trials in mild cognitive impairment (MCI).

Methods—We used normal control and amnestic MCI subjects from ADNI-1 as normative reference and screening cohorts. We evaluated the enrichment performance of four widely-used hippocampal segmentation algorithms (FreeSurfer, HMAPS, LEAP and NeuroQuant) in terms of two-year changes in MMSE, ADAS-Cog and CDR-SB. We modeled the effect of algorithm, test-retest variability and cut-point on sample size, screen fail rates and trial cost and duration.

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

*Corresponding author: Adam J. Schwarz, Eli Lilly and Company, Lilly Corporate Center DC 1940, Indianapolis IN 46285, USA. Tel. (+1) 317 405 7494, Fax. (+1) 317 277 7601, a.schwarz@lilly.com.

Disclosure statement

AJS, PY and PEC are employees and shareholders of Eli Lilly and Company.
JS is a paid contractor for Eli Lilly and Company.
RW is an employee of Philips Healthcare and consultant for IXICO Ltd.
DS is an employee of the Critical Path Institute and was previously employed by Pfizer.
DLGH is an employee and shareholder of IXICO Ltd.

Results—HCV-based patient selection yielded not only reduced sample sizes (by ~40–60%) but also lower trial costs (by ~30–40%) across a wide range of cut-points. Overall, the dependence on the cut-point value was similar for the three clinical instruments considered.

Conclusion—These results provide a guide to the choice of HCV cut-point for aMCI clinical trials, allowing an informed trade-off between statistical and practical considerations.

Keywords

Hippocampus; biomarker; enrichment; clinical trials; inclusion criterion; MRI; hippocampal volume; structural MRI

1. Introduction

There is increasing interest in studying disease modifying Alzheimer's disease (AD) therapies in pre-demented (e.g., mild cognitive impairment (MCI)) populations, but this can be challenging because the clinical trajectories can vary considerably despite well-defined clinical inclusion criteria – some subjects may remain stable for many years whereas others will deteriorate more rapidly (Petersen, 2004, Mitchell and Shiri-Feshki, 2009). This heterogeneity in clinical course is due to heterogeneity of the pathophysiology that underlies the clinical syndrome of MCI. In roughly 60–70% of cases, the clinical syndrome of amnestic MCI (aMCI) is attributable to AD pathology, most commonly mixed with other age related pathophysiology such as cerebrovascular disease or Lewy body disease (Jicha et al., 2006, Petersen et al., 2006). But in the remaining 30–40% of MCI cases, something other than AD dominates and this may include non-progressive etiologies such as depression. Etiological heterogeneity amongst MCI subjects has been one factor that has been proposed as contributing to the failures in clinical trials to date in this patient population (Peterson, 2011). This variability negatively impacts the statistical power and hence feasibility of a trial to detect a slowing of clinical decline.

Histopathological studies have shown early involvement of the hippocampus (Braak and Braak, 1991) and a large number of imaging studies have found early and disproportionate hippocampal atrophy to be a characteristic feature of AD. In amnestic populations, smaller hippocampi as measured from structural magnetic resonance imaging (MRI) scans have been widely associated with poorer short-term clinical prognosis both prior to and subsequent to the onset of dementia (Jack et al., 1999, Killiany et al., 2002, Jack et al., 2005, Devanand et al., 2007, Desikan et al., 2009, Henneman et al., 2009), in keeping with evidence of a temporal sequence of biomarker dynamics associated with AD pathology and progression (Jack et al., 2010, Jedynak et al., 2012, Jack et al., 2013) in which structural atrophy of the medial temporal lobes has the greatest rate of change at the aMCI and mild AD stages of the disease. This suggests utility of hippocampal volume as a "proximity marker" to AD dementia and hence its use as a staging tool to better identify subjects who are more likely to decline clinically. Indeed, based on this strong body of evidence, the measurement of low hippocampal volume from structural MRI has recently (Dec. 2011) been qualified by the European Medicines Agency (EMA) as an enrichment biomarker to select aMCI patients of imminent risk of rapid clinical deterioration for clinical trials (Hill et al., 2013)[1].

However, in order that low hippocampal volume can be applied prospectively as an enrichment biomarker in clinical trials, a number of practical questions relating to its operational implementation must be addressed. Firstly, a procedure to define a specific cut-point to be used as an inclusion criterion is required (Bartlett et al., 2012). One approach to this is to use a defined normative population, along with a specified mathematical model to adjust for covariates, from which the cut-point is defined (Jack et al., 1999). Secondly, an understanding of the expected practical implications (e.g., screen failure rate, effect sizes of clinical scales that may be used as outcome measures, trial duration and cost) is important in order to demonstrate the utility of this approach. Thirdly, although it is standard practice to utilize a single hippocampal volume measurement algorithm and centralized analysis within any individual study, a number of different algorithms are in common use for the quantification of HCV and the algorithm employed will likely differ across core laboratories and trials. At the present time, these algorithms differ both in their definition of the hippocampus itself as well as in the computational details of how its volume is estimated. It is therefore also important to understand how the enrichment performance depends on the algorithm employed. Finally, an understanding of how the intrinsic measurement variability of the hippocampal volume measurement affects the enrichment performance will determine the confidence with which any obtained performance may generalize to other equivalent cohorts.

Our aim in this study was to evaluate a cut-point based enrichment strategy applicable to clinical trials in an aMCI population using HCV data generated from four different widely-used algorithms. The overall hypothesis was that subjects with smaller hippocampi would progress more rapidly, yielding reduced sample sizes and more efficient clinical trials.

2. Methods

2.1 Study Population

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and

[1]EMA/CHMP/SAWP/809208/2011.

Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

We analyzed ADNI 1 data available as of November 2012 (http://www.loni.ucla.edu/ADNI). Subjects were followed for 2–3 years and assessed every 6–12 months. Normal subjects had MMSE scores between 24–30 (inclusive), a CDR of 0, were non-depressed, non-MCI, and nondemented. MCI subjects had Mini-Mental State Examination (MMSE) scores between 24–30 (inclusive), a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a Clinical Dementia Rating (CDR) of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. At each visit, subjects were evaluated using cognitive tests including the mini mental scale evaluation (MMSE; range 0–30 points) where lower scores indicate more cognitive impairment, the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog 13 item; range 0–85) where higher scores indicate worse cognitive function, and Clinical Dementia Rating Sum of Boxes (CDR-SB; range 0–18) where higher scores indicate more severe cognitive and functional impairment.

2.2 HCV analysis algorithms

We characterized the enrichment performance achieved in the same cohort of subjects using four widely used HCV algorithms. In each case, tabulated HCV data were provided to CAMD by the following image analysis laboratories. The algorithms used were: (1) FreeSurfer v4.3 (http://surfer.nmr.mgh.harvard.edu/) (Fischl et al., 2002, Fischl, 2012). FreeSurfer data were computed at the University of California, San Francisco and uploaded to the ADNI LONI website (http://adni.loni.ucla.edu/). (2) Hippocampus Multi-Atlas Propagation and Segmentation (HMAPS) (Leung et al., 2010). HMAPS data were provided by N. Fox (University College, London). (3) Learning Embeddings Atlas Propagation (LEAP) (Wolz et al., 2010). LEAP data were provided by R. Wolz (Ixico Ltd. and Imperial College, London). (4) NeuroQuant (http://www.cortechs.net/products/neuroquant.php) (Brewer, 2009). NeuroQuant data were provided by J. Brewer (University of California, San Diego).

These four algorithms are all highly, and in most cases fully, automated. The FreeSurfer processing protocol applied to the present data included thorough human quality control (QC) supervision, by which the parcellation results were designated as pass or fail for the hippocampus (http://adni.loni.ucla.edu/wp-content/uploads/2010/12/ADNI_UCSF_Freesurfer-Overview-and-QC.pdf). HMAPS is a fully automated hippocampal segmentation technique; as input it uses a volumetric T1-weighted brain scan with the brain regions delineated. No human intervention (e.g. fine-tuning of parameters or editing or discarding of the regions) was used (or required) for the HMAPS regions in this

manuscript. LEAP is fully automated with a visual QC check applied to the final segmentation to accept or reject the result; for the present data set, all hippocampal segmentations were acceptable. NeuroQuant is fully automated, deterministic and 510(c) approved for commercial use; no human intervention is permitted and the software includes automated QC checks of image input parameters and to reject scans where alignment to the probabilistic atlas is poor, such as when a lesion causing mass effect is present.

2.3 Adjusted hippocampal volume

Raw hippocampal volumes were adjusted for age and head size. As a measure of head size, we used the intracranial volume (ICV) calculated by NeuroQuant and provided by J. Brewer (University of California, San Diego). This measure was based on a combination of the white matter, grey matter and cerebral spinal fluid distribution maps determined using segmentation of each subject's image. For each HCV measurement algorithm, we built a linear regression model using the raw hippocampal volumes in the normal subjects as HCV = $\beta_0 + \beta_1 \times \text{ICV} + \beta_2 \times \text{age} + \varepsilon$, where β_i are regression coefficients to be computed from the data (HCV, ICV and age) and ε represents residual error. This model was then applied to the NL and MCI subjects to compute the adjusted HCV (aHCV) as aHCV = HCV - $\beta_1 \times$ ICV - $\beta_2 \times$ age.

2.4 Assessment of enrichment performance

The enrichment performance was profiled over a range of cutoff values *c*, determined as the 1^{st} to the 50^{th} percentile (in single percentile increments) of aHCV values in the healthy control cohort, which served as a normative reference population. For each cutoff value, and for each of the four HCV algorithms, we calculated the clinical characteristics of the selected subpopulation of aMCI subjects and derived implications for clinical trials where only this subpopulation would be enrolled.

For application to clinical trials the primary advantage of HCV-based enrichment is the likelihood of enrolling subjects with more rapidly changing trajectories of clinical decline and thereby also leading to a more homogeneous study population. We thus calculated the change in three widely-used clinical instruments (MMSE, ADAS-Cog13, CDR-SB) over 24 months following the MRI scan and quantified the enrichment performance in terms of several measures. First, we calculated the effect size (mean/standard deviation) of the 24-month change in each of the three clinical outcome measures. Denoting the effect size in the enriched sample by *ES*' and that in the unenriched sample by *ES*, we then calculated the sample size $N' = (ES/ES')^2 N$ required to achieve a statistical power equivalent to an unenriched sample of size *N*. We considered the scenario where the same relative change in the clinical scale is sought. Intuitively, a more homogeneous (less variable) or steeper average rate of decline in an enriched population should require a smaller sample size to detect such a change.

We also characterized practical implications for conducting a clinical trial using a HCV-based inclusion criterion, including the screen failure fraction (SFF), being the fraction of aMCI subjects not meeting the HCV cutoff criterion (i.e., with aHCV > cut-point), which ranges from 0 when *c*=100% (all subjects included) and increases toward unity as the cutoff

percentile *c* is decreased. We thus calculated the number of subjects needed to screen with MRI (NNS_{HCV}), $NNS_{HCV}=N'/(1-SFF)$ in order to achieve the required sample size *N*'. We did not explicitly model subject dropout.

2.5 Modeling trial cost and execution time

Next, we modeled indicative trial costs and execution times in the enriched sample as a function of cutpoint and in comparison with the unenriched scenario.

In the absence of the HCV enrichment strategy, the cost of a trial can be simply modeled as the sum of the screening cost plus the maintenance cost:

$$C_T = N_s.C_s + N.D.C_m \quad (1)$$

Where N_s is the number of subjects needed to enter standard trial screening, C_s is the cost per patient of the screening phase, *N* is the number of subjects to be randomized, *D* is the trial duration in years and C_m is the annual cost of maintaining each patient in the study.

For our current enrichment scenario, we modeled the additional cost of the HCV assessment via an additional term. We assumed that the HCV assessment will only be performed on subjects that have already successfully passed the other inclusion and screening criteria. We further assumed that an MRI scan is already included as part of the screening procedures, and that this is the last screening procedure performed. Under these assumptions, the trial cost equation is modified to:

$$C'_T = N'_s.C_s + NNS_{HCV}.C_{HCV} + N'.D.C_m \quad (2)$$

Here C_{HCV} is the additional cost associated with obtaining a HCV measurement on each subject, and NNS_{HCV} represents the number of subjects needed to undergo a screening HCV measurement in order to obtain the required sample size *N*'. In turn, N'_s is the number of subjects needed to enter screening to obtain NNS_{HCV}.

In an enrichment scenario, more screen fails will occur due to HCV-based exclusion, providing upward pressure on N'_s (and hence the overall screening cost) for a given enrollment target *N*'. However, *N*' would be expected to decrease relative to the unenriched case (*N*), due to the higher effect sizes in the clinical endpoints, providing downward pressure on the trial maintenance cost. Eq. (2) provides a simple model to capture the impact of these competing influences of the enrichment strategy on trial cost.

The time required to prosecute a clinical trial can similarly be considered as the sum of the screening time and the trial observation period following randomization of the last subject:

$$T'_T = \frac{N'_s}{R_s} + D \quad (3)$$

Here R_s is the rate of subject screening (y^{-1}) and in the unenriched scenario $N'_s = N_s$. This model allows straightforward calculation of trial cost and execution time implications for a given level of enrichment and the concomitant estimate of N'_s.

In order to provide indicative estimates of the effect of enrichment on trial execution time and cost, we used parameters estimated based on recent experience at Lilly (Table 1). The cost calculations were based on the amount paid to investigators per patient for screening and for maintenance during the trial. These are the parameters typically used for costing calculations as additional trial costs depend on the degree to which trial operation functions that are outsourced, and can be appropriately scaled by the sample sizes under consideration. The simulated trial scenario assumed that an MRI scan was already being performed at screening for radiological purposes. The additional cost for HCV based inclusion was estimated as the additional cost required for the core lab overheads relating to the implementation of the structural MRI acquisition suitable for quantification (as part of the same screening MRI scanning session), site and core lab training and process implementation (amortized across subjects) and the cost for the HCV computation itself performed by the core lab. We further assumed that the HCV measurement would occur in parallel to the radiological review of the MR images, hence incurring no additional screening delay.

Sample sizes were determined assuming a two-year treatment trial, powered for a 25% slowing of the rate of decline in the clinical endpoint (MMSE, ADAS-Cog13 or CDR-SB), a power of 0.8 and alpha of 0.05.

2.6 Estimating the effect of measurement error on enrichment

Finally, we used the results of a recent test-retest analysis (Wolz et al., 2012) with one of the algorithms (LEAP) to estimate the effect of measurement variability on the enrichment performance. In that work, test-retest differences of $-0.34\pm1.93\%$ were reported using 1.5T data on 74 ADNI aMCI subjects. Based on this finding, we simulated 1000 datasets by randomly sampling hippocampus volumes for each of the 287 subjects in the present study, based on a Gaussian distribution with zero mean and standard deviation of 1.93% around the actual measured LEAP hippocampal volumes at each cut point, and evaluated the distributions of the derived parameters over the 1000 simulated datasets. This enabled the sensitivity of these parameters to measurement variability in the underlying HCV measure to be assessed. This was quantified in terms of the standard deviation (across the 1000 simulated runs) in the afore-mentioned enrichment assessments.

3. Results

3.1 Sample characteristics

In this study, we used subjects from the standardized lists of ADNI-1 baseline 1.5T MRI scans for normal (to define the normative reference range of HCVs) and aMCI (as a putative screening population) subjects as recommended by the ADNI magnetic resonance core (Wyman et al., 2012). From the standardized lists, one control subject and one aMCI subject had no FreeSurfer parcellation available and were thus excluded. Furthermore, 104 aMCI subjects did not have ADAS-Cog13, CDR-SB and MMSE measurements at both baseline and 24 months, and 5 subjects had ambiguous clinical diagnoses. The final HCV data set comprised measurements from N=228 normal subjects and N=287 aMCI subjects with

baseline HCV measurements from all four algorithms and 24 month clinical follow-up. A list of subject IDs used in this study is provided in the Supplemental Material.

The baseline subject characteristics for the normal and aMCI subjects used in this study are provided in Table 2. All four algorithms showed a similar reduction (12.2% to 15.6%) in mean HCV in the aMCI group relative to controls. Subjects excluded due to absence of clinical follow-up data for all three clinical scales are also summarized. There were trends to more severe cognitive scores and smaller hippocampi at baseline in the subjects excluded due to missing follow-up data. Only the difference in ADAS-Cog13 (mean 19.8 vs. 18.2) was significant (p=0.029, t-test, uncorrected for multiple comparisons).

3.2 Impact of different cut-points on HCV-based enrichment

3.2.1 Clinical scale effect sizes and required sample sizes—The effect size (mean/standard deviation of change over 24 months) of all three clinical scales, MMSE, ADAS-Cog(13) and CDR-SB, was increased in all sub-groups of aMCI subjects selected according to the aHCV inclusion criteria, relative to the unenriched scenario in which all aMCI subjects would be enrolled (Figure 1(a–c)). In general, more stringent cut-points (smaller aHCV thresholds) led to greater effect size increases. Cut-points corresponding to the 10^{th}, 25^{th} and 40^{th} percentiles of the normal control distribution yielded effect sizes of between 15% and 61% (MMSE), 12% and 39% (ADAS-Cog) and 14% and 44% (CDR-SB) greater than the unenriched scenario (Table 3).

A consequence of the increased effect size is that a smaller number of subjects need to be enrolled in order to obtain an equivalent statistical power to the unenriched scenario (Figure 2). Reduced sample sizes were obtained across all the cut-points evaluated, for all algorithms tested. For cut-points corresponding to the 10^{th}, 25^{th} and 40^{th} percentiles of the normal control population, sample sizes corresponding to 39–76% (MMSE), 52–80% (ADAS-Cog) or 48–78% (CDR-SB) of the unenriched case (100%) were estimated (Table 3).

3.2.2 Screen fail fraction and number needed to screen—A consequence of more stringent patient selection strategy is that additional screen fails will occur, as subjects with an aHCV value above the cut-point will be excluded from the study. For the range of cutpoints considered in this study, the percentage of the aMCI screening population that would be excluded ranged from 14–20% at a cut-point corresponding to the 50^{th} percentile of the normal controls to 76–85% at a cut-point corresponding to the 1^{st} percentile of the normal controls (across algorithms). For cut-points corresponding to the 10^{th}, 25^{th} and 40^{th} percentiles of the normal control population, the screen fail fraction ranged from 19–57 (Table 3).

However, the number of subjects needed to screen in order to enroll the required number of subjects reflects both the increased screen failures and the decreased sample size due to enrichment. In the present analysis the resulting number of subjects needed to screen, relative to the unenriched scenario, depended upon the clinical scale. For MMSE, fewer subjects would need to be screened for cut-points greater than ~15%; in contrast, using ADAS-Cog or CDR-SB would require screening a similar overall number of subjects for

cut-points greater than ~25%. More stringent cut-points would result in more subjects being screened relative to the unenriched scenario (Figure 1(d–f)). For cut-points corresponding to the 10^{th}, 25^{th} and 40^{th} percentiles of the normal control population, the number needed to screen ranged from 74%–101% (MMSE), 77%–129% (ADAS-Cog) and 88%–115% (CDR-SB) of those needed to screen in the absence of enrichment (Table 3).

3.2.3 Implications for trial duration and cost—Estimates of trial cost and total duration for enriched and unenriched scenarios are shown for all cut-points in Figure 3 and enumerated for selected cut-points in Table 4. HCV-based patient selection resulted in substantially reduced estimated trial costs for all clinical scales and across the full range of cut-points, with the exception of the very lowest (~1%) cut-point for ADAS-Cog (Figure 3(a–c)). In contrast, estimated trial duration was reduced for MMSE and cut-points >15%, but similar to the unenriched case for ADAS-Cog and CDR with cut-points >25%. More stringent cut-points resulted in increased estimated trial duration (Figure 3(d–f)).

3.3 Effects of measurement variability and algorithm

The enrichment characteristics in terms of the metrics considered here followed similar curves against cut-point for all four HCV algorithms tested. Simulations based on the empirically determined measurement error in hippocampal volume estimates were used to determine the impact of this variability on the enrichment parameters (Figures 1, 3). For most cut-points considered, the estimated standard deviation in these parameters due to intrinsic (test-retest) measurement variability was similar to or smaller in magnitude than the variability due to the different algorithms.

4. Discussion

In this study, we systematically characterized the operational performance of hippocampal volume (HCV) as an enrichment biomarker for clinical trials in an amnestic MCI population. Hippocampal volume was recently qualified for this purpose by the EMA – the first imaging biomarker to achieve regulatory qualification – based on a large number of studies showing associations between hippocampal volume measurements and subsequent clinical progression (typically, conversion to dementia) (Hill et al., 2013). This work included a demonstration that the same four algorithms as assessed here perform very similarly in predicting progression to clinical dementia within two years. However, to date little attention has been directed to the operational implementation of HCV-based patient selection as a screening tool. Here, we addressed four important aspects directly impacting how this approach would affect clinical trials in practice: (1) how to select the cut-point, (2) the effect of measurement variability on enrichment performance, (3) the difference in performance across different HCV algorithms, and (4) how different clinical instruments behave as outcome measures in this enrichment scenario.

This work was motivated by the substantial current interest in conducting AD treatment trials in earlier, pre-dementia stages of the disease. This interest is driven by factors including recent failed trials in AD dementia populations, the observation of a trend toward greater treatment effect of solanezumab in patients with milder symptoms, and an increasing body of post-mortem and biomarker evidence showing that the key pathologies of AD are

present prior to the onset of dementia. However, amnestic MCI populations are clinically heterogeneous, with a wide range of symptom trajectories, providing substantial challenges to conducting efficient clinical trials in this population. A measurement of hippocampal volume from a structural MRI scan provides a non-invasive, widely available and well-established biomarker to select a fraction of the screening population that is likely to imminently progress most rapidly. Importantly, this biomarker is applicable to putative treatments of any mechanism of action. However, for routine use, clear and simple operational guidelines are required. Similar to recent work on operationalizing the NINDS criteria (Albert et al., 2011) for clinical use (Jack et al., 2012), our aim here was to answer key practical questions regarding the application of hippocampal volume as a screening tool in aMCI clinical trials.

4.1 Choice of cut-point

For use as an inclusion criterion a predefined cut-point must be specified. Currently, different HCV algorithms generate different absolute hippocampal volumes, precluding the selection of a universal cut-point. Moreover, the inclusion of covariates in an adjusted model leads to a departure from absolute physical volumes. We therefore specified cut-points in terms of percentiles of a distribution of adjusted hippocampal volumes from a healthy aged control cohort. We assessed the enrichment performance over a wide range of plausible cut-point values, corresponding to the first through fiftieth percentiles of the normative reference distribution. This approach allowed the same cut-point selection methodology to be applied to data generated using different HCV algorithms, on the understanding that the same algorithm was applied to both the reference and screening cohorts. However, an effort is underway to establish a consensus standard definition of the hippocampus for MRI-based volumetry (Boccardi et al., 2011), which is expected to yield increased convergence in the performance across different algorithms.

Improvements in both effect size and sample size in the enriched population were found, beyond variability due to LEAP test-retest error and choice of algorithm, over the full range of cut-points evaluated, with effect sizes maximized, and hence sample sizes minimized, at more stringent cut-points less than ~25% (~0. 7 standard deviations below the mean) of the ADNI normative reference population. However, since more stringent cut-points result in a higher screen failure rate (Lorenzi et al., 2010), the dependence of trial time and cost on cut-point does not necessarily track the sample size curves. The reduced sample size and increased screen failure rate act to decrease and increase, respectively, the trial time and cost. Taking into account both screening and maintenance activities, our results indicate that HCV-based enrichment would result in reduced investigator costs across virtually the full range of cut-points and was minimized in the 10%–25% range with 30–40% cost savings predicted. Although we note that this cost calculation pertains only to investigator and imaging costs for the trial itself, and does not take into account other financial factors such as a reduced on-patent marketing window if enrichment causes overall trial duration to be longer due to an extended period of screening. However, the estimated trial duration was actually reduced (by 5–15%) for an MMSE outcome with HCV cut-points greater than ~15%. For ADAS-Cog13 and CDR-SB endpoints, the estimated trial time was unchanged

from the unenriched case for HCV cut-points greater than ~25%. At lower cut-point values, overall trial duration was increased.

4.2 Effect of measurement variability and hippocampal measurement algorithm

To our knowledge, the present study is the first to undertake explicit modeling of the effect of measurement variability on the enrichment performance. Overall, we found that the effect of measurement error was small relative to the dynamic range of the parameters studied, and generally smaller than the differences due to the different algorithms.

Despite the fact that different algorithms perform a different segmentation of the hippocampus, the enrichment performance was for the most part very similar across algorithms and for the different parameters evaluated. Given that each of these algorithms (and others) are used by specialist commercial and academic analysis groups that serve as core laboratories for clinical trials, this is an important finding and suggests a general robustness of hippocampal volume measures in this context. An ongoing effort under the auspices of the European Alzheimer's Disease Consortium (EADC), ADNI and the Alzheimer's Association to define a standard definition of the hippocampal boundaries (Boccardi et al., 2011) is expected to substantially reduce variability between hippocampal volumes computed by different algorithms once they are re-tuned to this standard definition and accredited as recently proposed (Jack et al., 2011). In the context of enrichment, this methodological standardization is anticipated to reduce the inter-algorithm differences to within the range of measurement variability.

4.3 General considerations

Hippocampal atrophy represents a measure of gross neurodegeneration, which is most prominent in the medial temporal lobes from the prodromal into later phases of AD. Thus, when coupled with existing amnestic MCI clinical inclusion criteria, our results show that it has utility as a means of identifying subjects at increased risk of imminent clinical decline. This is consistent with previous findings that HCV-based selection increases the rate of clinical conversion to AD dementia – in other words, that HCV can serve as a "proximity marker" to dementia in amnestic MCI populations. Moreover, HCV measures may be combined with other, pathologically specific (e.g., amyloid, tau), biomarkers to further refine the patient population and ensure presence of target pathology (Vos et al., 2012, Yu et al., 2012).

4.4 Limitations

The present analysis used data from the ADNI study, a population that may perhaps have slightly different characteristics from those typically recruited for participation in industry sponsored trials. The generalizability of the enrichment performance shown here to other clinical cohorts thus remains an important outstanding issue that is necessary to address in order to determine the extent to which the performance indicated in the present study can be expected when used prospectively in clinical trials. Nevertheless, the strong body of convergent evidence regarding the utility of hippocampal volume measures in predicting conversion to clinical dementia (Hill et al., 2013) suggests that the operational advantages indicated here are likely to obtain.

We did not explicitly model subject dropout, which in practice would affect both statistical power (hence, likely increase number needed to enroll) and thus also the number needed to screen. For the present analysis, the number needed to enroll represents the number of completers, and would need to be inflated based on an expected drop-out rate. The baseline subject characteristics indicate that the subjects without 24 month clinical follow-up data may have slightly more severe cognitive function at baseline. However, this was only borderline significant ($p<0.05$, uncorrected for multiple comparisons), suggesting that the dropout rate is not strongly related to hippocampal volume per se. Our simple model thus still provides an indication of the time and cost relative to the unenriched case.

We also note that the results presented here are based on the common assumption that treatment effects will be proportional to mean outcome – an assumption that may not hold, for example if the molecular pathology relevant to the treatment target is heterogeneous in the population. Results from actual treatment trials in the future will be required to clarify the validity of this assumption and refine how sample size calculations should be modeled in this context (cf. (Barnes et al., 2013)).

5. Conclusions

We evaluated a cut-point based enrichment strategy for hippocampal volume-based patient selection of aMCI clinical trial subjects, explicitly evaluating four different widely-used hippocampal segmentation algorithms and test-retest variability of one of them. In addition to standard sample size considerations, we also modeled the practical implications of this method in terms of predicted trial cost and duration. We found that HCV-based selection yielded not only reduced sample sizes but also lower trial costs across a wide range of cut-points, although trial duration was not substantially reduced. Overall, the dependence on the cut-point value was similar for the three clinical instruments considered. These results provide a guide to the choice of HCV cut-point for aMCI clinical trials, allowing an informed trade-off between statistical and practical considerations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was performed in collaboration with the Coalition Against Major Diseases (CAMD), part of the Critical Path Institute. The Critical Path Institute's CAMD is supported by the US FDA [grant number U01FD003865] and Science Foundation Arizona [grant number SRG 0335-08].

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health

(www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, Rev October 16, 2012 San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

JB's research is supported by NINDS K02 NS067427, NIA U01 AG10483, NIA P50 AG005131, NIA R01AG034062, and General Electric Medical Foundation. He was an investigator for, and received research funds from Janssen Alzheimer Immunotherapy. He has served on advisory boards for Elan, Bristol-Myers Squibb, Avanir, and Eli Lilly and holds stock options in CorTechs Labs, Inc.

NCF's research group has received payment for consultancy or for conducting studies from AVID, Bristol-Myers Squibb, Elan Pharmaceuticals, Eisai, Lilly Research Laboratories, GE Healthcare, IXICO, Janssen Alzheimer Immunotherapy, Johnson & Johnson, Janssen-Cilig, Lundbeck, Neurochem Inc, Novartis Pharma AG, Pfizer Inc, Sanofi-Aventis and Wyeth Pharmaceuticals. NCF has an NIHR Senior Investigator award and receives support from the Wolfson Foundation; NIHR Biomedical Research Unit (Dementia) at UCL; the EPSRC; Alzheimer's Research UK and the NIA. NCF receives no personal compensation for the activities mentioned above.

CRJ provides consulting services for Siemens Healthcare. He receives research funding from the National Institutes of Health (R01-AG011378, RO1-AG041851, RO1-AG037551, U01-HL096917, U01-AG032438, U01-AG024904), and the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Foundation.

References

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011; 7:270–279. [PubMed: 21514249]

Barnes J, Bartlett JW, Fox NC, Schott JM. Targeted recruitment using cerebrospinal fluid biomarkers: implications for Alzheimer's disease therapeutic trials. J Alzheimers Dis. 2013; 34:431–437. [PubMed: 23229078]

Bartlett JW, Frost C, Mattsson N, Skillback T, Blennow K, Zetterberg H, Schott JM. Determining cut-points for Alzheimer's disease biomarkers: statistical issues, methods and challenges. Biomark Med. 2012; 6:391–400. [PubMed: 22917141]

Boccardi M, Ganzola R, Bocchetta M, Pievani M, Redolfi A, Bartzokis G, Camicioli R, Csernansky JG, de Leon MJ, deToledo-Morrell L, Killiany RJ, Lehericy S, Pantel J, Pruessner JC, Soininen H, Watson C, Duchesne S, Jack CR Jr, Frisoni GB. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J Alzheimers Dis. 2011; 26(Suppl 3):61–75. [PubMed: 21971451]

Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991; 82:239–259. [PubMed: 1759558]

Brewer JB. Fully-automated volumetric MRI with normative ranges: translation to clinical practice. Behav Neurol. 2009; 21:21–28. [PubMed: 19847042]

Desikan RS, Cabral HJ, Hess CP, Dillon WP, Glastonbury CM, Weiner MW, Schmansky NJ, Greve DN, Salat DH, Buckner RL, Fischl B. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain. 2009; 132:2048–2057. [PubMed: 19460794]

Devanand DP, Pradhaban G, Liu X, Khandji A, De Santi S, Segal S, Rusinek H, Pelton GH, Honig LS, Mayeux R, Stern Y, Tabert MH, de Leon MJ. Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease. Neurology. 2007; 68:828–836. [PubMed: 17353470]

Fischl B. FreeSurfer. Neuroimage. 2012; 62:774–781. [PubMed: 22248573]

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33:341–355. [PubMed: 11832223]

Henneman WJ, Sluimer JD, Barnes J, van der Flier WM, Sluimer IC, Fox NC, Scheltens P, Vrenken H, Barkhof F. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. Neurology. 2009; 72:999–1007. [PubMed: 19289740]

Hill DLG, Schwarz AJ, Isaac M, Pani L, Vamvakas S, Hemmings R, Carrillo MC, Yu P, Sun J, Beckett L, Boccardi M, Brewer J, Brumfield M, Cole PE, Fox N, Frisoni GB, Jack C, Kelleher T, Luo F, Novak G, Maguire P, Meibach R, Patterson P, Bain L, Sampaio C, Soares H, Suhy J, Wang H, Wolz R, Stephenson D. CAMD/EMA Biomarker Qualification of Hippocampal Volume for Enrichment of Clinical Trials in Predementia Stages of Alzheimer's Disease. 2013 Submitted.

Jack CR Jr, Barkhof F, Bernstein MA, Cantillon M, Cole PE, Decarli C, Dubois B, Duchesne S, Fox NC, Frisoni GB, Hampel H, Hill DL, Johnson K, Mangin JF, Scheltens P, Schwarz AJ, Sperling R, Suhy J, Thompson PM, Weiner M, Foster NL. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheimers Dement. 2011; 7:474–485. e474. [PubMed: 21784356]

Jack CR Jr, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski JQ. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 2013; 12:207–216. [PubMed: 23332364]

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]

Jack CR Jr, Knopman DS, Weigand SD, Wiste HJ, Vemuri P, Lowe V, Kantarci K, Gunter JL, Senjem ML, Ivnik RJ, Roberts RO, Rocca WA, Boeve BF, Petersen RC. An operational approach to National Institute on Aging-Alzheimer's Association criteria for preclinical Alzheimer disease. Ann Neurol. 2012; 71:765–775. [PubMed: 22488240]

Jack CR Jr, Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, Kokmen E. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology. 1999; 52:1397–1403. [PubMed: 10227624]

Jack CR Jr, Shiung MM, Weigand SD, O'Brien PC, Gunter JL, Boeve BF, Knopman DS, Smith GE, Ivnik RJ, Tangalos EG, Petersen RC. Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. Neurology. 2005; 65:1227–1231. [PubMed: 16247049]

Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, Raunig D, Jedynak CP, Caffo B, Prince JL. A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort. Neuroimage. 2012

Jicha GA, Parisi JE, Dickson DW, Johnson K, Cha R, Ivnik RJ, Tangalos EG, Boeve BF, Knopman DS, Braak H, Petersen RC. Neuropathologic outcome of mild cognitive impairment following progression to clinical dementia. Arch Neurol. 2006; 63:674–681. [PubMed: 16682537]

Killiany RJ, Hyman BT, Gomez-Isla T, Moss MB, Kikinis R, Jolesz F, Tanzi R, Jones K, Albert MS. MRI measures of entorhinal cortex vs hippocampus in preclinical AD. Neurology. 2002; 58:1188–1196. [PubMed: 11971085]

Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, Schuff N, Fox NC, Ourselin S. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage. 2010; 51:1345–1359. [PubMed: 20230901]

Lorenzi M, Donohue M, Paternico D, Scarpazza C, Ostrowitzki S, Blin O, Irving E, Frisoni GB. Enrichment through biomarkers in clinical trials of Alzheimer's drugs in patients with mild cognitive impairment. Neurobiol Aging. 2010; 31:1443–1451. 1451 e1441. [PubMed: 20541287]

Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia--meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand. 2009; 119:252–265. [PubMed: 19236314]

Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med. 2004; 256:183–194. [PubMed: 15324362]

Petersen RC, Parisi JE, Dickson DW, Johnson KA, Knopman DS, Boeve BF, Jicha GA, Ivnik RJ, Smith GE, Tangalos EG, Braak H, Kokmen E. Neuropathologic features of amnestic mild cognitive impairment. Arch Neurol. 2006; 63:665–672. [PubMed: 16682536]

Vos S, van Rossum I, Burns L, Knol D, Scheltens P, Soininen H, Wahlund LO, Hampel H, Tsolaki M, Minthon L, Handels R, L'Italien G, van der Flier W, Aalten P, Teunissen C, Barkhof F, Blennow K, Wolz R, Rueckert D, Verhey F, Visser PJ. Test sequence of CSF and MRI biomarkers for

prediction of AD in subjects with MCI. Neurobiol Aging. 2012; 33:2272–2281. [PubMed: 22264648]

Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D. LEAP: learning embeddings for atlas propagation. Neuroimage. 2010; 49:1316–1325. [PubMed: 19815080]

Wolz R, Schwarz AJ, Yu P, Cole PE, Rueckert D, Jack CR, Raunig D, Hill DL. Robustness of Automated Hippocampal Volumetry across MR Field Strengths and Same-Session Repeat Scans. 2012 Submitted.

Wyman BT, Harvey DJ, Crawford K, Bernstein MA, Carmichael O, Cole PE, Crane PK, Decarli C, Fox NC, Gunter JL, Hill D, Killiany RJ, Pachai C, Schwarz AJ, Schuff N, Senjem ML, Suhy J, Thompson PM, Weiner M, Jack CR Jr. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimers Dement. 2012

Yu P, Dean RA, Hall SD, Qi Y, Sethuraman G, Willis BA, Siemers ER, Martenyi F, Tauscher JT, Schwarz AJ. Enriching amnestic mild cognitive impairment populations for clinical trials: optimal combination of biomarkers to predict conversion to dementia. J Alzheimers Dis. 2012; 32:373–385. [PubMed: 22796873]

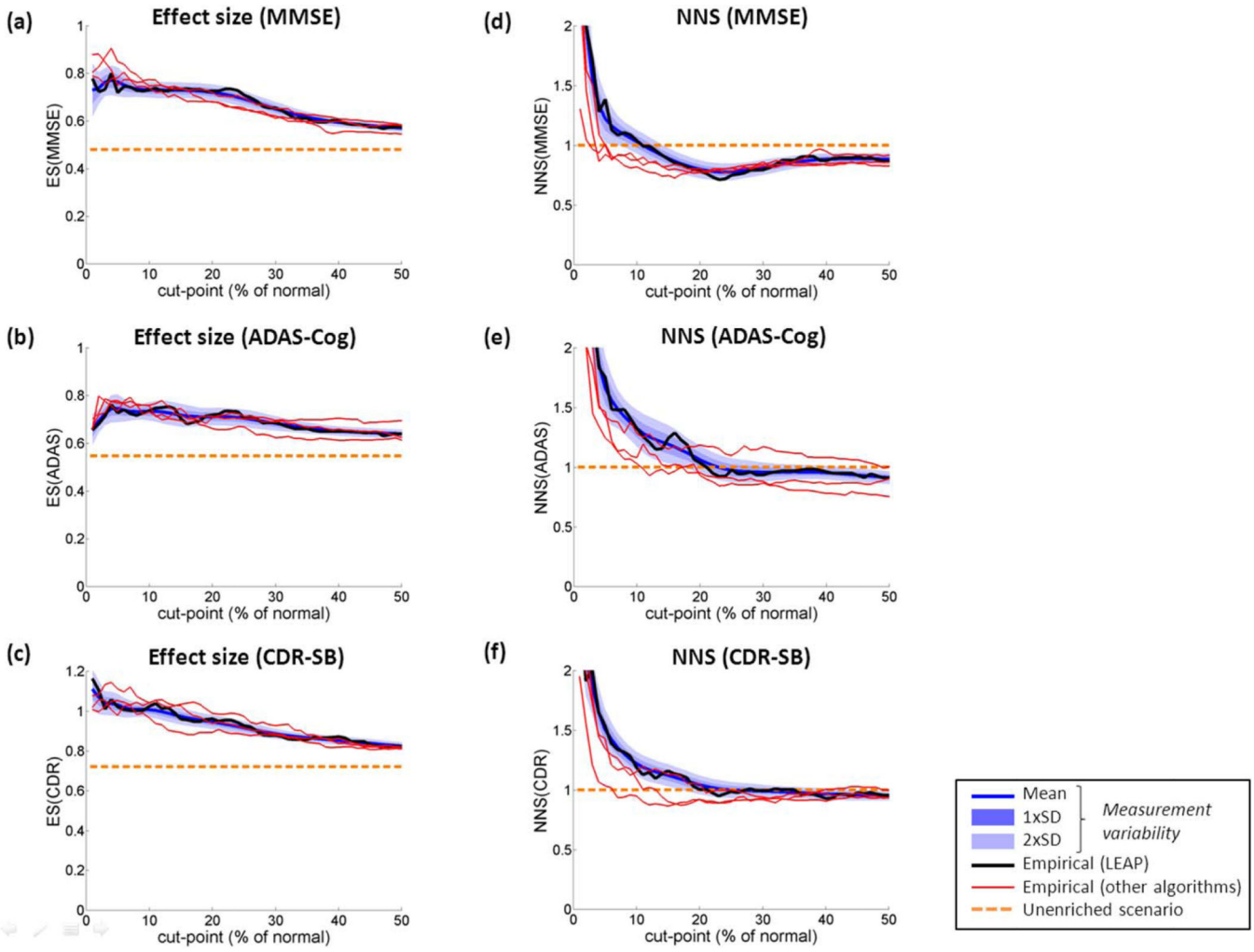

Figure 1.

Effect of aHCV-based enrichment on (a–c) effect sizes and (d–f) number of subjects needed to screen (NNS), as a function of cut-point (specified as percentile, 1%–50%, of the distribution of aHCVs in the normal control population) for (a,d) MMSE, (b,e) ADAS-Cog and (c,f) CDR-SB. Results are shown for four different HCV computational algorithms. Variance due to test-retest measurement variability is shown as the shaded area for one of the four algorithms (LEAP). Effect size was calculated as mean / standard deviation of the two-year change in each clinical scale. Number needed to screen (NNS) is the number of subjects needed to undergo screening MRI scans, in order to enroll the predicted sample sizes and is expressed as a fraction of the unenriched scenario.

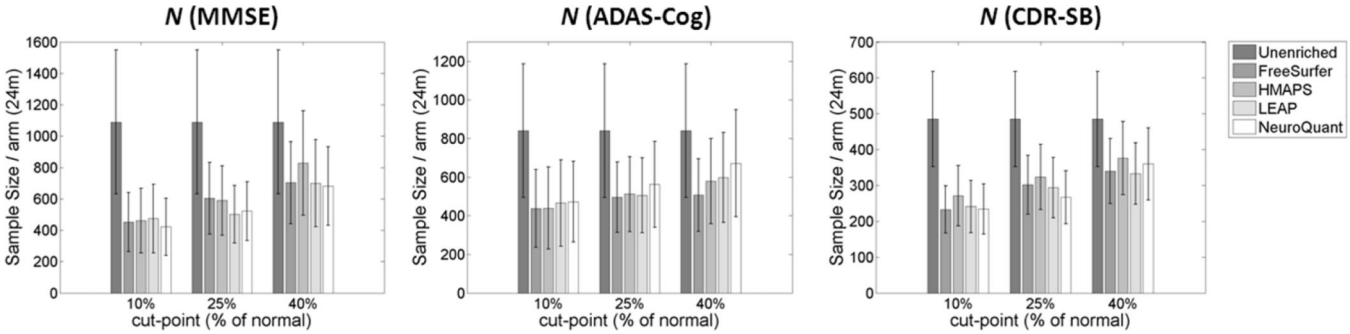

Figure 2.

Sample sizes for a 24 month trial using HCV based patient selection, for three selected cutpoints (10%, 25% and 40%). Sample sizes were calculated as those required to achieve the same absolute change in the clinical trajectories as in the unenriched case. Error bars indicate 95% confidence intervals.

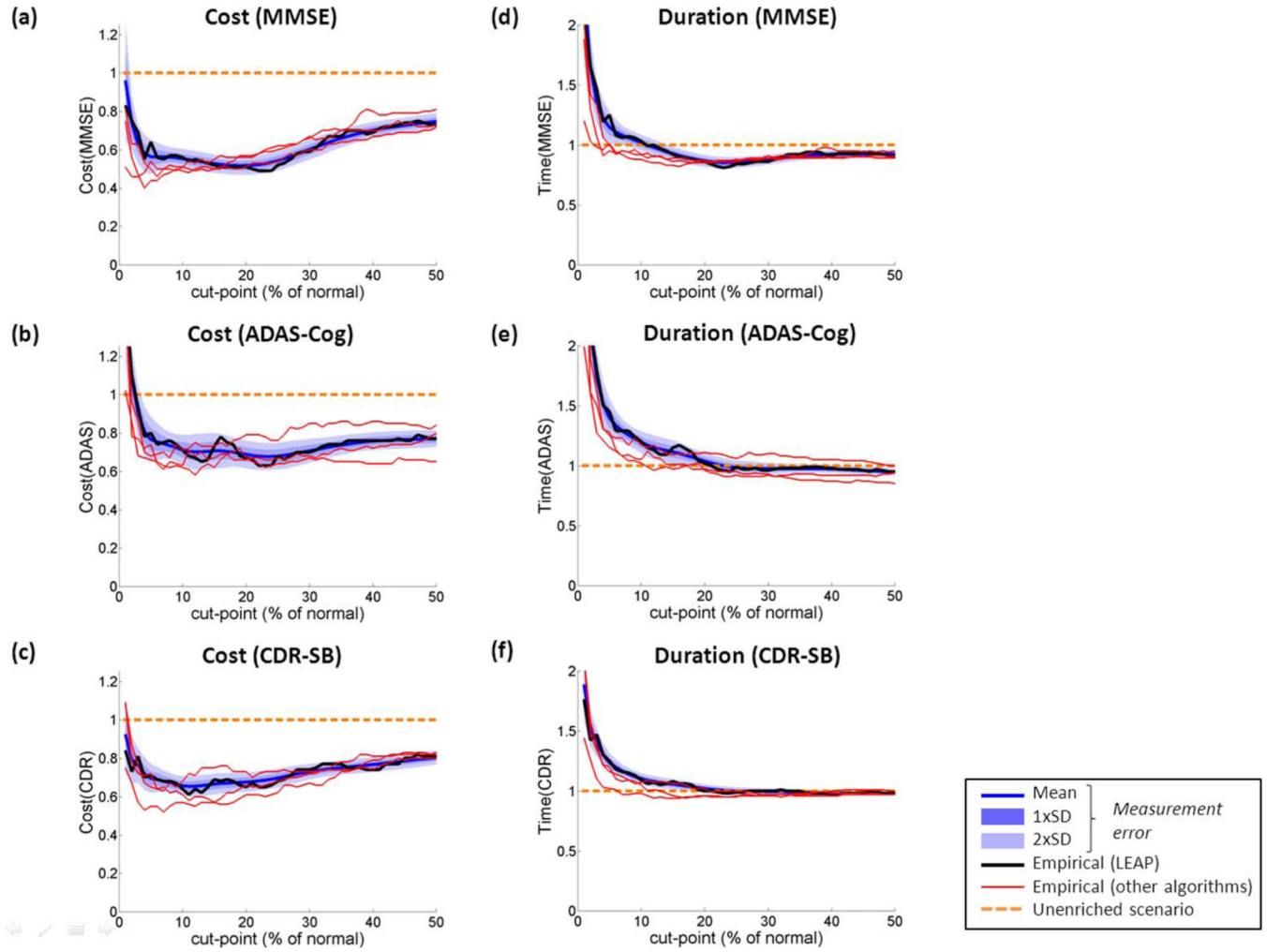

Figure 3.

Implications of aHCV-based enrichment for (a–c) trial cost and (d–f) trial execution time, as a function of cut-point (specified as percentile, 1%–50%, of the distribution of aHCVs in the normal control population) for (a,d) MMSE, (b,e) ADAS-Cog and (c,f) CDR-SB. Results are expressed as fractions of the unenriched scenario, and are shown for four different HCV computational algorithms. Variance due to test-retest measurement variability is shown as the shaded area for one of the four algorithms (LEAP).

Table 1

Parameter values used in trial duration and cost calculations.

Symbol	Parameter	Value
D	Trial treatment duration	24m
C_{HCV}	Additional cost for each HCV measurement	$1,000
R_s	Screening rate	800/y
C_s	Screening cost per patient	$5,800
C_m	Maintenance cost per patient	$18,500/y
NNS_{HCV}/N_s'	Fraction of subjects that enter screening who fulfull screening criteria prior to HCV measurement	0.7

Table 2

Baseline subject characteristics (mean±SD and range provided except for number of subjects and sex).

Variables	NL	aMCI	aMCI (excluded due to incomplete 24m clinical followup)
Number of subjects	228	287	104
Age	75.9±5.0 (60–90)	74.7±7.2 (55–88)	75.0±8.0 (55–89)
Sex (%female)	48%	36%	38%
MMSE	29.1±1.0 (25–30)	27.1±1.7 (23–30)	26.8±1.9 (24–30)
ADAS–Cog13	9.5±4.2 (1–21)	18.2±6.4 (3–40)	19.8±5.9 (8–36)*
CDR–SB	0.0±0.1 (0.0–0.5)	1.6±0.8 (0.5–5)	1.7±0.9 (0–4.5)
Adjusted HCV – FreeSurfer (cm^3)	3.76±0.37 (2.84–5.26)	3.28±0.46 (2.14–4.46)	3.23±0.44 (2.20–4.34)
Adjusted HCV – HMAPS (cm^3)	2.63±0.32 (1.70–3.58)	2.26±0.37 (1.36–3.18)	2.19±0.34 (1.45–3.11)
Adjusted HCV – LEAP (cm^3)	1.80±0.27 (1.12–2.94)	1.52±0.32 (0.76–2.67)	1.48±0.31 (0.80–2.27)
Adjusted HCV – NeuroQuant (cm^3)	3.20±0.36 (2.20–4.21)	2.81±0.46 (1.53–4.09)	2.74±0.47 (1.66–4.11)

* $p<0.05$ (uncorrected).

Table 3

Two-year clinical scale changes and implications for clinical trials (sample size, screen fail fraction and number needed to screen) for three illustrative cut-points, corresponding to the 10th, 25th and 40th percentile of the normative aHCV distribution.

		unenriched	cutpoint at 10th percentile				cutpoint at 25th percentile				cutpoint at 40th percentile			
			FS	HMAPS	LEAP	NQ	FS	HMAPS	LEAP	NQ	FS	HMAPS	LEAP	NQ
MMSE	**Change***	−1.676 ± 3.489	−2.745 ± 3.676	−2.626 ± 3.558	−2.758 ± 3.784	−2.825 ± 3.655	−2.388 ± 3.699	−2.411 ± 3.691	−2.670 ± 3.770	−2.611 ± 3.758	−2.153 ± 3.598	−2.000 ± 3.629	−2.222 ± 3.706	−2.217 ± 3.649
	ES**	−0.480 (*−0.573, −0.387*)	−0.747 (*−0.893, −0.601*)	−0.738 (*−0.889, −0.587*)	−0.729 (*−0.883, −0.575*)	−0.773 (*−0.935, −0.611*)	−0.645 (*−0.76, −0.53*)	−0.653 (*−0.77, −0.536*)	−0.708 (*−0.833, −0.583*)	−0.695 (*−0.816, −0.574*)	−0.598 (*−0.705, −0.491*)	−0.551 (*−0.658, −0.444*)	−0.599 (*−0.712, −0.486*)	−0.608 (*−0.717, −0.499*)
	N/N$_{unenriched}$†	1	0.414 (*0.168, 0.660*)	0.424 (*0.161, 0.686*)	0.434 (*0.159, 0.710*)	0.386 (*0.148, 0.624*)	0.554 (*0.238, 0.870*)	0.541 (*0.237, 0.845*)	0.460 (*0.200, 0.720*)	0.478 (*0.215, 0.741*)	0.644 (*0.283, 1.006*)	0.76 (*0.324, 1.196*)	0.642 (*0.276, 1.009*)	0.625 (*0.28, 0.971*)
	NNS/NNS$_{unenriched}$†	1	0.819 (*0.328, 1.310*)	0.875 (*0.324, 1.426*)	1.039 (*0.365, 1.712*)	0.880 (*0.327, 1.433*)	0.811 (*0.342, 1.281*)	0.808 (*0.347, 1.270*)	0.750 (*0.318, 1.182*)	0.784 (*0.343, 1.225*)	0.833 (*0.36, 1.306*)	0.948 (*0.398, 1.499*)	0.870 (*0.370, 1.370*)	0.847 (*0.375, 1.318*)
ADAS-Cog	**Change***	3.927 ± 7.18	5.097 ± 6.716	5.209 ± 6.886	5.281 ± 7.182	5.572 ± 7.636	5.249 ± 7.366	5.118 ± 7.296	5.303 ± 7.514	5.172 ± 7.734	4.994 ± 7.090	4.704 ± 7.135	4.783 ± 7.369	4.585 ± 7.492
	ES**	0.547 (*0.440, 0.654*)	0.759 (*0.587, 0.931*)	0.756 (*0.574, 0.938*)	0.735 (*0.558, 0.912*)	0.73 (*0.561, 0.899*)	0.713 (*0.581, 0.845*)	0.701 (*0.566, 0.836*)	0.706 (*0.569, 0.843*)	0.669 (*0.537, 0.801*)	0.704 (*0.574, 0.834*)	0.659 (*0.535, 0.783*)	0.649 (*0.525, 0.773*)	0.612 (*0.491, 0.733*)
	N/N$_{unenriched}$†	1	0.519 (*0.195, 0.844*)	0.523 (*0.186, 0.859*)	0.553 (*0.197, 0.910*)	0.562 (*0.216, 0.908*)	0.589 (*0.269, 0.909*)	0.608 (*0.271, 0.945*)	0.601 (*0.270, 0.932*)	0.669 (*0.295, 1.043*)	0.603 (*0.254, 0.951*)	0.688 (*0.288, 1.088*)	0.71 (*0.287, 1.133*)	0.799 (*0.314, 1.283*)
	NNS/NNS$_{unenriched}$†	1	1.028 (*0.368, 1.688*)	1.079 (*0.365, 1.794*)	1.323 (*0.453, 2.193*)	1.280 (*0.462, 2.097*)	0.863 (*0.385, 1.341*)	0.909 (*0.392, 1.426*)	0.979 (*0.428, 1.531*)	1.097 (*0.469, 1.726*)	0.779 (*0.329, 1.230*)	0.859 (*0.358, 1.359*)	0.961 (*0.390, 1.532*)	1.081 (*0.425, 1.738*)
CDR-SB	**Change***	1.409 ± 1.955	2.048 ± 1.967	1.975 ± 2.044	2.117 ± 2.069	2.198 ± 2.114	1.872 ± 2.047	1.820 ± 2.064	1.923 ± 2.076	2.014 ± 2.070	1.736 ± 2.016	1.639 ± 2.002	1.757 ± 2.017	1.731 ± 2.069
	ES**	0.721 (*0.625, 0.817*)	1.041 (*0.892, 1.190*)	0.966 (*0.815, 1.117*)	1.023 (*0.867, 1.179*)	1.040 (*0.882, 1.198*)	0.915 (*0.789, 1.041*)	0.882 (*0.757, 1.007*)	0.926 (*0.792, 1.06*)	0.973 (*0.837, 1.109*)	0.862 (*0.748, 0.976*)	0.819 (*0.709, 0.929*)	0.871 (*0.759, 0.983*)	0.837 (*0.721, 0.953*)
	N/N$_{unenriched}$†	1	0.479 (*0.291, 0.668*)	0.557 (*0.326, 0.787*)	0.497 (*0.288, 0.705*)	0.480 (*0.280, 0.681*)	0.621 (*0.383, 0.860*)	0.668 (*0.406, 0.930*)	0.606 (*0.364, 0.847*)	0.549 (*0.339, 0.759*)	0.700 (*0.440, 0.960*)	0.775 (*0.485, 1.066*)	0.685 (*0.434, 0.936*)	0.742 (*0.457, 1.027*)
	NNS/NNS$_{unenriched}$†	1	0.949 (*0.560, 1.338*)	1.15 (*0.655, 1.644*)	1.188 (*0.664, 1.711*)	1.094 (*0.614, 1.575*)	0.91 (*0.551, 1.269*)	0.999 (*0.595, 1.402*)	0.987 (*0.578, 1.397*)	0.900 (*0.544, 1.257*)	0.905 (*0.560, 1.250*)	0.968 (*0.599, 1.336*)	0.927 (*0.582, 1.273*)	1.005 (*0.611, 1.399*)

		unenriched	cutpoint at 10th percentile				cutpoint at 25th percentile				cutpoint at 40th percentile			
			FS	HMAPS	LEAP	NQ	FS	HMAPS	LEAP	NQ	FS	HMAPS	LEAP	NQ
All	**SFF**†	0	0.495 (*0.438, 0.552*)	0.516 (*0.458, 0.573*)	0.582 (*0.526, 0.638*)	0.561 (*0.503, 0.619*)	0.317 (*0.263, 0.371*)	0.331 (*0.275, 0.387*)	0.387 (*0.330, 0.444*)	0.390 (*0.332, 0.449*)	0.226 (*0.179, 0.274*)	0.199 (*0.153, 0.244*)	0.261 (*0.211, 0.312*)	0.261 (*0.208, 0.315*)

*Specified as mean ± standard deviation.

**Calculated as mean / standard deviation.

†Numbers in parentheses are the 95% confidence intervals.

Table 4

Sample sizes and trial duration and cost calculations for three illustrative cut-points, corresponding to the 10th, 25th and 40th percentile of the

		unenriched	cutpoint at 10th percentile				cutpoint at 25th percentile				cutpoint at 40th percentile			
			FS	HMAPS	LEAP	NQ	FS	HMAPS	LEAP	NQ	FS	HMAPS	LEAP	NQ
MMSE	**Sample size / arm to randomize***	1090 (631,1549)	452 (264,640)	462 (256,668)	474 (255,693)	422 (240,604)	604 (376,832)	590 (369,811)	502 (318,686)	522 (335,709)	703 (441,965)	828 (495,1161)	700 (422,978)	682 (432,932)
	Trial duration (y)	5.9 (4.3,7.5)	5.2 (3.8,6.6)	5.4 (3.8,7)	6.0 (4.1,8)	5.4 (3.9,7)	5.2 (3.9,6.4)	5.1 (3.9,6.4)	4.9 (3.8,6)	5.1 (3.9,6.2)	5.2 (4,6.5)	5.7 (4.2,7.2)	5.4 (4,6.8)	5.3 (4.1,6.5)
	Trialcost (M$)	100.9 (58.4,143.4)	50.1 (29.1,71)	51.9 (28.6,75.2)	56.1 (30,82.3)	49.1 (27.7,70.5)	61.1 (38,84.2)	60.0 (37.5,82.6)	52.4 (33,71.7)	54.5 (35,74.1)	68.9 (43.1,94.7)	80.5 (48.1,112.9)	69.4 (41.7,97.1)	67.6 (42.7,92.5)
ADAS–Cog	**Sample size / arm to randomize***	841 (495,1187)	438 (237,639)	440 (228,652)	466 (242,690)	473 (264,682)	496 (314,678)	512 (317,707)	506 (312,700)	563 (340,786)	508 (320,696)	579 (358,800)	598 (365,831)	672 (395,949)
	Trial duration (y)	5.0 (3.8,6.2)	5.1 (3.6,6.6)	5.2 (3.6,6.9)	6.0 (4,8)	5.8 (4,7.7)	4.6 (3.6,5.6)	4.7 (3.7,5.8)	4.9 (3.8,6.1)	5.3 (3.9,6.7)	4.3 (3.5,5.2)	4.6 (3.6,5.6)	4.9 (3.8,6)	5.2 (3.9,6.6)
	Trial cost (M$)	77.9 (45.9,109.8)	48.5 (26,71)	49.4 (25.4,73.4)	55.2 (28.4,82)	55.0 (30.4,79.6)	50.2 (31.7,68.7)	52.1 (32.1,72.1)	52.8 (32.5,73.1)	58.8 (35.4,82.2)	49.8 (31.4,68.2)	56.3 (34.8,77.7)	59.3 (36.2,82.3)	66.6 (39.2,94.1)
CDR–SB	**Sample size / arm to randomize***	485 (352,618)	233 (167,299)	271 (187,355)	241 (168,314)	234 (164,304)	302 (220,384)	324 (233,415)	294 (210,378)	267 (193,341)	340 (249,431)	376 (274,478)	333 (247,419)	360 (259,461)
	Trial duration (y)	3.7 (3.3,4.2)	3.6 (3.1,4.1)	4.0 (3.3,4.7)	4.1 (3.4,4.7)	3.9 (3.3,4.5)	3.6 (3.1,4)	3.7 (3.2,4.2)	3.7 (3.2,4.2)	3.6 (3.1,4)	3.6 (3.1,4)	3.7 (3.2,4.1)	3.6 (3.2,4)	3.7 (3.2,4.3)
	Trial cost (M$)	44.9 (32.6,57.2)	25.8 (18.5,33.1)	30.4 (21,39.9)	28.5 (19.8,37.3)	27.2 (19,35.5)	30.6 (22.2,38.9)	33.0 (23.6,42.3)	30.7 (21.9,39.5)	27.9 (20.2,35.6)	33.3 (24.3,42.3)	36.5 (26.6,46.5)	33.0 (24.5,41.6)	35.7 (25.6,45.8)

* Sample size based on 25% difference, 80% power, alpha=0.05, trial duration = 2 years. Numbers in brackets are 95% confidence intervals.