Published in final edited form as:

Neurology. 2003 January 28; 60(2): 253–260.

MRI as a Biomarker of Disease Progression in a Therapeutic Trial of Milameline for AD

C. R. Jack Jr., M.D., M. Slomkowski, Pharm.D., S. Gracon, D.V.M., T. M. Hoover, Ph.D., J. P. Felmlee, Ph.D., K. Stewart, B.S., Y. Xu, M.D., Ph.D., M. Shiung, B.A., P. C. O'Brien, Ph.D., R. Cha, M.S., D. Knopman, M.D., and R. C. Petersen, Ph.D. M.D.

Departments of Diagnostic Radiology (CRJ, YCX, JPF, YCX, KS, MS), Neurology (DK, RCP), and Biostatistics (PCO, RC) Mayo Clinic and Foundation, 200 First Street SW, Rochester, MN 55905. Pharmaceutical Research (MS, SG, TMH) Parke-Davis (Pfizer), 2800 Plymouth Road, Ann Arbor, MI 48105

Abstract

Objective—To assess the feasibility of using magnetic resonance imaging (MRI) measurements as a surrogate end point for disease progression in a therapeutic trial for Alzheimer's disease (AD).

Methods—Three-hundred-sixty-two patients with probable AD from 38 different centers participated in the MRI portion of a 52 week randomized placebo controlled trial of milameline, a muscarinic receptor agonist. The therapeutic trial itself was not completed due to projected lack of efficacy on interim analysis; however, the MRI arm of the study was continued. Of the 362 subjects who underwent a baseline MRI study, 192 subjects underwent a second MRI one-year later. Hippocampal volume and temporal horn volume were measured from the MRI scans.

Results—The annualized percent change in hippocampal volume (−4.9%) and temporal horn volume (16.1%) in the study patients were consistent with data from prior single site studies. Correlations between the rate of MRI volumetric change and change in behavioral/cognitive measures were greater for the temporal horn than for the hippocampus. Decline over time was more consistently seen with imaging measures, 99% of the time for the hippocampus, than behavioral/cognitive measures ($p < 0.001$). Greater consistency in MRI than behavioral/clinical measures resulted in markedly lower estimated sample size requirements for clinical trials. The estimated number of subjects per arm required to detect a 50% reduction in the rate of decline over one year are: ADAS-Cog 320; MMSE 241; hippocampal volume 21; temporal horn volume fifty-four.

Conclusion—The consistency of MRI measurements obtained across sites, and the consistency between the multi-site milameline data and that obtained in prior single site studies, demonstrate the technical feasibility of using structural MRI measures as a surrogate end point of disease progression in therapeutic trials. However, validation of imaging as a biomarker of therapeutic efficacy in AD still awaits a positive trial.

The primary outcome measurements for therapeutic trials in Alzheimer's disease (AD) patients are behavioral or cognitive. Due to the inherent test-retest variability in such measurements however, alternatives have been sought. Magnetic resonance imaging (MRI) measurements of rates of whole brain or hippocampal atrophy have been, and are currently being used as outcome measures in several therapeutic trials for AD. Although imaging has been used in clinical trials on AD and vascular disease for diagnostic purposes, to our knowledge, no publication has appeared describing the MRI results of a therapeutic trial in which structural MRI was used as

an outcome measure. MRI measures were added to this trial to gain a claim for effects on disease progression as opposed to just symptomatic treatment in those instances where treatment effect was shown on the behavioral/cognitive measures¹⁻⁷.

We report the MRI results of a therapeutic trial of milameline, a centrally active muscarinic agonist. The therapeutic objective was augmentation of the diminished cholinergic function characteristic in AD^{8, 9}. The therapeutic trial was not completed due to a projected lack of efficacy on interim analysis. However, the MRI arm of the study was continued in order to collect data for reference purposes. The purpose of this manuscript therefore is not to report on the clinical outcome of the therapeutic trial itself, but rather the MRI portion of the trial in order to illuminate methodologic considerations, document feasibility, and serve as a guideline for future studies using MRI as an outcome measure. Demonstrating that the MR image data obtained in this trial from multiple sites were both internally consistent and consistent with reports in prior single site studies, we believe documents the technical feasibility of using MRI as an outcome measure of disease progression in therapeutic trials for neurodegenerative diseases.

METHODS

Study Design

The study was a 52-week randomized double blind placebo controlled parallel group multicenter trial of milameline, a muscarinic receptor agonist. All patients were titrated up to the maximum tolerated daily dose with a ceiling of 4mg/day. Clinical followup was at months one, three, six, and twelve. The primary outcome measure was the Alzheimer's Disease Assessment Scale-Cognitive sub-scale (ADAS-Cog). Several secondary outcome measures were employed including the Mini-Mental State Exam (MMSE), Global Deterioration Scale (GDS), and MRI measurements of the rates of atrophy of the hippocampi, and the rates of enlargement of the temporal horns¹⁰.

Patients

All patients had probable AD of mild to moderate severity and were greater than 50 years old. MMSE scores were 10-27, inclusive. Patients were excluded for evidence of substantial cerebral vascular disease, non-AD dementia, major psychiatric disorders, major or unstable medical conditions, seizure disorder, Parkinson's disease, substance abuse, major head trauma, or tumor.

Planned enrollment was a total of 450 subjects; all from US sites with an average of ten subjects (range 10-15 maximum) from each of 45 sites. The planned study duration was from June 1996 to December 1998.

MRI

In order to avoid difficulties associated with MRI manufacturer's proprietary data acquisition and storage protocols, only those sites with a General Electric MR Imager were asked to enroll patients in the MRI portion of the trial. The precise hardware and software configuration of scanners at different sites varied. All scanners were 1.5T with the exception of a single site at 0.5T.

The individual imaging sequences in the examination protocol were:

1. A sagittal T1-weighted scan with contiguous 5-mm slices.

2. A 3D volumetric spoiled gradient recalled echo (SPGR) scan obtained in the coronal plane with minimum full echo-time, minimum repetition time, 124 partitions, and 1.6 mm partition thickness.
3. A T2-weighted scan was acquired for pathology screening purposes.

Each site went through an initial qualifying phase prior to scanning patients for the trial. The qualifying phase had two components. First, each site was required to perform the pulse sequences specified for the imaging protocol, film the examination, and send the films to the central analysis site at the Mayo Clinic. The pulse sequence parameters were checked at the central analysis site to insure that each participating site could execute the specified imaging sequences correctly. Second, the scanner on which study patients would be imaged at each site underwent a quality control evaluation. In order to begin studying patients, signal to noise ratio (SNR) and geometric distortion analyses must have been completed and evaluated at the central site, and must have fallen within predetermined specifications.

SNR and geometric distortion measurements were made on the specified scanner at each site both during the qualification phase, and throughout the trial on a monthly basis. Each site performed and submitted to the central image analysis site an axial spin echo and sagittal, coronal, and axial gradient echo acquisitions which were specifically designed to evaluate SNR and geometric distortion. The quality control (QC) imaging protocol was approximately 15 minutes in duration, and was done with a standardized QC phantom supplied by the manufacturer. The phantom contained spatially uniform regions where SNR measurements were obtained and also contained fiducial markers which were used to assess geometric distortion¹¹. The data from the QC MR protocol was transmitted to Mayo, where an ongoing record of SNR, image artifacts, RF power, and geometric distortion along all three axes was kept for each scanner throughout the duration of the trial.

All image data (both patients' imaging data and phantom QC imaging data) was sent by either tape or disc to Mayo and archived electronically. For each patient the total intracranial volume was measured from the sagittal T1 sequence. The volumes of the right and left hippocampi and temporal horns were measured according to previously described criteria^{6, 12, 13} on the 3D SPGR sequence. All the measurements were done at Mayo by a single individual over two years. Intra-rater co-efficient of variation for serial hippocampal measurements has been documented at 0.28%⁶. The individual performing the MR image analyses was blinded to all clinical information including the center and the scan date – ie whether each scan was the 1st or 2nd in the pair. The order of first and second scans was randomized.

MRI Data Analysis

Adjustments of Raw Hippocampal Volume for Total Intracranial Volume—The raw hippocampal volumes in each subject were adjusted first by total intracranial volume (TIV) and then second by referencing the TIV adjusted raw hippocampal volume in that subject to age and gender specific percentile values in a normal elderly population¹³. Hippocampal volumes in cognitively normal subjects vary with head size, (individuals with larger TIV have larger hippocampi) and age (volume declines with advancing age).

In order to identify the relationship between hippocampal volume and TIV in normals, free of the effect of aging, the sagittal T1 and coronal 3D SPGR pulse sequences were performed in 79 young subjects between the age of 17 and 45, who were documented to be free of neurologic disease. The relationship between hippocampus and TIV in these normal young individuals was deduced using a regression analysis of hippocampus on TIV. The regression equations describing expected hippocampal volume as a function of measured TIV was derived separately for men and women and are:

$$\text{Men: } H' = H / (3.8 \times 10^{-3}(\text{TIV}))$$

$$\text{Women: } H' = H / (2443 + 2.0 \times 10^{-3}(\text{TIV}))$$

where H = raw unadjusted hippocampal volume, and H' = hippocampal volume adjusted by TIV. From this data in young people we are able to project what the volume of the hippocampi should be for a given TIV in any individual *unaffected* by aging.

Age and Gender Specific Elderly Norms for TIV Adjusted Hippocampal Volume

—We next turned to hippocampal volumes in normal elderly people. The sagittal and 3D SPGR pulse sequence had been performed in 181 normal elderly individuals (61 men and 120 women, mean age 80.4 years, range 62-100 years) as part of the ongoing clinical research protocols in the Mayo Clinic Alzheimer's Disease Research Center (ADRC) and Alzheimer's Disease Patient Registry (ADPR). For each of these normal elderly subjects the TIV adjusted hippocampal volume was computed based on the relationship between TIV and hippocampal volume that had been derived for the normal young cohort as described above. Age and gender specific normative percentiles of this TIV adjusted hippocampal volume value were then computed for the normal elderly cohort. The method for computing age and gender specific normative percentiles for hippocampal volume was identical to the method we have previously described¹³, except that here normative percentiles were calculated in the elderly subjects for the TIV adjusted hippocampal volume.

Age and TIV Adjusted Hippocampal Volume in Milameline Study Patients—The baseline hippocampal volumes in the AD cases in the milameline study were expressed as W scores, as we have done previously¹³. The W score corresponds to the percentile value in a standard normal distribution. A W score of zero corresponds to the 50th percentile of normal elderly individuals. A W score of 1.645 corresponds to the 95th percentile of normal elderly individuals, and a W score of -1.645 corresponds to the 5th percentile.

Serial MRI Measures in Milameline Study Patients—In each study patient volume measurements of the hippocampi were obtained at two different time points. From these, the raw change in volume (in mm³), raw annualized change in mm³/year, and annualized percent change were calculated. No relationship was found between TIV and the rates of change in hippocampal volume, and therefore no TIV adjustment was needed⁶.

We did not have measurements of temporal horn volume available in young normals, or old normal subjects. For this reason, adjustment of temporal horn volume for TIV, and the W score method could not be developed for the temporal horn measurements. We analyzed the annualized raw volumetric change, and the annualized percent change in temporal horn volume in a manner identical to that described for the hippocampus.

Statistical Analysis

The variables of interest in the analyses were age, gender, education, ADAS-Cog score, MMSE, GDS, and MRI volume variables. The MRI volume variables include raw data in mm³ for the right, left, and total hippocampus at baseline, W score of total hippocampal volume at baseline, change in total hippocampal volume, raw baseline temporal horn volumes, and change in total temporal horn volume. Rank sum tests were used for skewed data, and two sample t tests were used for normally distributed data. Spearman correlation was used to test for associations between MRI volume measurements and behavioral/cognitive measures, education, gender, and age. Univariate and multivariate modeling were used to test for associations between baseline MRI volumes and both demographic and behavioral/cognitive measures. Univariate and multivariate modeling were used to test for associations between change in MRI volumes and baseline MRI volume, demographic, and behavioral/cognitive

measures. Rank transformations were used for all data in the multivariate modeling. One way ANOVA was used to test for differences in baseline hippocampal W score between the different centers. The degree to which serial measures, imaging and behavioral/cognitive, declined consistently over time was compared using the sign test.

Sample size calculations for clinical trials

Because of the potential interest in using structural MRI as a biomarker of disease progression in clinical AD trials, we performed sample size calculations to detect treatment effect based on the annual change data in MMSE, ADAS-Cog, hippocampus, and temporal horn (the GDS was felt not to be a competitive metric). Sample size calculations were based on the assumption of a 50% treatment effect over 12 months—i.e. a 50% reduction in the change over one year in a treated vs. a placebo group. Tests were one sided with power set at 90%.

RESULTS

Baseline Demographic, Behavioral/Cognitive, and MRI Data

Out of 453 subjects enrolled in the drug trial, 362 participated in the MRI portion. Analysis of baseline data is restricted to these subjects (Table 1). Among these 362 subjects, 188 had been randomized to treatment and 174 to placebo. No differences in gender distribution (roughly 60% female), education, behavioral/cognitive measures, or baseline MRI volume measures were present between the treated and placebo groups.

Women had less education, ($p = 0.001$) and worse baseline cognitive performance on the ADAS-Cog ($p = 0.02$) and MMSE ($p = 0.01$). The raw right, left, and total hippocampal ($p < 0.001$) and temporal horn ($p < 0.001$) volumes were larger in men than women. There was a trend toward greater hippocampal atrophy (lower W score and normal percentile ranking) in women than men, but this did not reach significance.

The right hippocampus was approximately 100-200 mm³ (roughly 5%) larger than the left in both treated and placebo groups. However, no right-left differences were present in any of the correlations between hippocampal or temporal horn volume and baseline demographic or behavioral/cognitive measures, nor were right-left differences present in correlations between change in MRI volumes and change in behavioral/cognitive measures. Therefore, all subsequent analyses of MRI volume measures are reported for the total (sum of right plus left) volume rather than right or left volumes separately.

Correlation among Baseline Hippocampal Volume, Demographic, and Behavioral/Cognitive Measures

The placebo and treated groups were combined for baseline correlation analyses because the baseline MRI and behavioral/cognitive measures were completed prior to treatment initiation. Univariate correlations between baseline hippocampal W score, age, gender, education, ADAS-Cog, MMSE and GDS were performed. No significant correlations were present with the exception of that between hippocampal W score and age ($r = 0.15$, $p = 0.004$). A series of multivariate models were constructed with performance on the behavioral/cognitive measures listed above as the dependent variable, and baseline hippocampal W score, age, gender, and education as independent variables. None of the partial correlations between baseline hippocampal W score and baseline behavioral/cognitive performance were significant.

Normative temporal horn measurements were not available and thus a system for adjusting raw baseline temporal horn volumes of milameline study patients for age and gender in normals was not possible. We therefore did not assess correlations among baseline temporal horn volume, demographic, and behavioral/cognitive measures.

Changes from Baseline

All analyses of change (change in MRI volumes and change in behavioral/cognitive scores) are based on the 192 subjects who had both a baseline and a second MRI scan (Table 2). No difference in annual change was present between the treated and placebo groups in any of the behavioral/cognitive measures, nor for either of the MRI volume variables. For each of these variables, however, the median annualized change was different from 0 ($p < 0.001$, for all, signed rank test). In addition, all variables in Tables 2 changed in the expected direction over time – ie. behavioral/cognitive performance worsened, hippocampal volume decreased, and temporal volume increased.

To assess how consistently imaging and behavioral/cognitive measures declined over time, we computed the proportion of individuals in whom the measures in Table 2 declined. Decline was defined as a decrease in hippocampal volume, an increase in temporal horn volume, a decrease in MMSE score, and an increase in the GDS and ADAS-Cog scores. The volume of the hippocampus decreased in 99% of subjects whereas only 60.4% of subjects ADAS-Cog and 66.2% of subjects' MMSE scores declined. In pair-wise comparisons, the proportion of decliners was greater for the hippocampus than any of the behavioral/cognitive measures or the temporal horn ($p < 0.001$). The proportion of decliners was also greater for the temporal horn (85.4%) than for any of the behavioral/cognitive measures ($p < 0.001$).

Sample size calculations for clinical trials

There were two outliers in the MMSE data. After deleting these values for purposes of computing sample size requirements, the MMSE and temporal horn data were not highly skewed and transformations of these data were not needed. For these variables, we computed the effect size to be 50% of the observed mean annual rate. The hippocampus and ADAS-cog were highly skewed and transformations of these data were needed. For these we computed the effect size after transformation as follows: transformed(median) - transformed(.5xmedian).

For ADAS-cog, we first transformed to $Y' = Y - \min + 1$ ($\min = -59.9$), then used $Y' * 10^{-2}$, where “min” refers to the minimum observed value.

For hippocampus, Y' is defined as above ($\min = -15.2$), and we used $Y' * 10^2$.

Notice that we are using a 50% reduction in mean rate of change where transformations are not required and a 50% reduction in the median rate of change where transformations are required. After deletion of the two outliers from the MMSE data, the mean annual percent change was 10.7% (19.9). The sample size required to detect a 50% reduction in this rate of change in a one year placebo controlled trial with power of 90% (one-sided t test at the .05 level) is 241 per arm. For ADAS-Cognitive, the data were highly skewed with a median of 16.4% (the associated SD of the transformed data was 65.4). In order to detect of 50% reduction, $N=320$ per arm are needed. For the hippocampus, the data were again highly skewed with a median of -4.9% (the associated SD of the transformed data was 2.1). In order to detect a 50% reduction, $N=21$ per arm are needed. The data for the temporal horn were not skewed, with mean 16.1% (SD14.1), and $N=54$ per arm are needed.

Factors that Influence Change in MRI Volume

In order to assess whether demographic variables or baseline MRI volume influenced the annualized change in volume, a multivariate model was constructed with the raw annualized change in hippocampal volume (mm^3) as the dependent variable, and with age, gender, education, and baseline hippocampal W score as independent variables. The slope (Beta), standard error, partial Spearman's correlation, and associated p value for each of the four independent variables in this model appear in Table 3. Baseline hippocampal W scores were

inversely associated with rates of hippocampal atrophy (ie smaller hippocampi at baseline were associated with greater volume loss over time). Age, gender, and education at baseline were not associated with the annualized rate of hippocampal atrophy.

Similar modeling was performed using the raw annualized change in temporal horn volume (mm^3) as the dependent variable, and age, gender, education, and baseline temporal horn volume as predictor variables. Because W scores were not available for temporal horn measurements, the raw temporal horn volume in mm^3 at baseline was used as the independent baseline volume variable. Younger age at baseline was associated with a greater annualized change in temporal horn volume (i.e., a greater rate of atrophy). A larger temporal horn volume at baseline was also associated with a larger annualized change in temporal horn volume.

Similar modeling was performed using the raw annualized change in volume (hippocampal and temporal horn) as the dependent variable, and age, gender, education, and baseline behavioral/cognitive variables as predictor variables. None of the baseline behavioral/cognitive variables were associated with the annualized change in hippocampal volume. ADAS-Cog was the only baseline cognitive variable associated with the annualized change in temporal horn volume ($r = 0.38$, $p < 0.001$).

Correlation between MRI Volumes and Change in Behavioral/Cognitive Measures

Univariate Analyses—To assess the association between change in behavioral/cognitive performance and baseline hippocampal and temporal horn volume, univariate analyses were performed with the annual percent change in each of the three behavioral/cognitive measures as dependent variables and each baseline MR structure's volume as the independent variable. With one exception (MMSE change score and baseline temporal horn volume) none of these correlations were significant (Table 4 and 5).

To assess the association between change in behavioral/cognitive performance and change in hippocampal volume or temporal horn volume over the same time period, a series of univariate analyses were performed with the annualized percent change in each of the behavioral/cognitive measures as dependent variables, and the annualized raw change in MR volume (mm^3) as the independent variables. None of these correlations were significant for hippocampal volume. Change in all three cognitive variables was associated with the annual raw change of temporal horn volume (Tables 4 and 5).

Multivariate Analyses—Multivariate models were then constructed, one for each of the three behavioral/cognitive variables. In each model, the annualized percent change in behavioral/cognitive performance was the dependent variable, and the independent variables were age, gender, education, MRI volume at baseline, and the annualized raw change in MRI volume (mm^3). For baseline volumes, only baseline hippocampal W score was significant and only with change scores on the GDS.

In multi-variate modeling, greater annualized change in temporal horn volume (i.e. greater rate of atrophy) was associated with a greater change (worse cognitive performance) on all three behavioral/cognitive measures. In contrast, annualized change in hippocampal volume was not associated with change scores of any of the cognitive behavioral measures.

MRI Quality Control Measures

All 38 participating sites eventually met qualifying criteria from the standpoint of MRI quality control. However, over the course of the study, results of QC analysis warranted site contact 19 times. Most of the QC alerts occurred during the one-month qualifying phase, which preceded actual enrollment of patients. The nature of the QC alerts were: gradient coil error

(i.e., geometric distortion), nine alerts at seven sites; low SNR, five alerts at four sites; noise lines, three alerts at three sites; other, two alerts at two sites.

Another measure of quality control is the consistency of measured volumes across sites. The mean hippocampal W score across all sites was -1.78 (SD 0.97). Hippocampal W score did differ among sites (one way ANOVA, $p = 0.04$). We then used the Student-Newman-Keuls multiple range test to do pairwise comparisons, and no pairwise difference between sites was found. We estimated the components of variance: variance arising from difference among sites (V_A) and variance arising from difference among patients within sites (V_W). There were $V_A = 0.04$ and $V_W = 0.90$, and the percent of total variance ($V_T = V_A + V_W$) due to variability among sites was 4.71% ($V_A/V_T \times 100\%$). Stepping down multivariate regression was used with hippocampal W score as the dependent variable; age at MRI scan, female gender, treatment, site, baseline ADAS-Cog score, baseline MMSE score, and baseline GDS score as the independent variables. Only one site was significantly different from the others in this analysis.

DISCUSSION

A major objective of these analyses was to validate the technical feasibility of using MRI as an outcome measure of disease progression in multi-site studies of neurodegenerative disease. One measure of the validity of multicenter data, is the degree to which it is concordant with similar data acquired from a single center or centers. Single center data is less likely to be corrupted by technical non-uniformity than multicenter data. In the milameline study patients, the right hippocampus was roughly 5% larger than the left. This same side to side volume difference has been found repeatedly in analyses of normals and patient groups through the age spectrum by our group and others and is a reflection of normal right-left asymmetry in hippocampal volume^{12, 14-20}. The raw hippocampal and temporal horn volumes in the milameline patients are in close agreement with these values reported previously in AD subjects from our own center^{12, 13}. The average hippocampal W score and corresponding percentile ranking of the milameline study patients, 3rd percentile of normal controls, corresponds very closely to the published W scores of mild to moderately demented AD patients in our own center¹³.

The observed inverse correlation between raw hippocampal volume and age is consistent with our own data and that of others for both AD patients and normals^{1, 5, 12, 13, 21, 22}. We interpret the decline in hippocampal W score with advancing age as an effect of disease duration. Greater age should on average correspond to greater disease duration that in turn will correspond to smaller (more atrophic) hippocampi. The adjustment for age in computing the W score correction only accounts for the effect of normal aging on hippocampal volume. It does not account for the effect of greater disease duration in older AD patients (vs younger AD patients). The fact that no correlations were observed between baseline hippocampal volume and global behavioral/cognitive test performance in the milameline study subjects may at first seem counter intuitive. However, this is consistent with our past experience when radiologic-cognitive correlations are limited to a single clinical group - for example AD patients only or normal subjects only. We believe that the lack of correlation in this circumstance is due to the truncated range of values for both baseline hippocampal volume and cognitive performance. In past studies, when normals were combined with AD subjects, thereby expanding the range of volume and cognitive performance values, highly significant correlations were present²³.

As indicated in Table 2, no treatment effect of milameline was observed in the MRI data. This was an expected finding due to the lack of treatment efficacy observed on interim analysis and patients randomized to treatment had discontinued the drug before completing the full 12-month course when the second MRI was obtained. The annualized rates of hippocampal atrophy in the AD subjects in this trial on average were just under 5%. This is slightly greater

than the annualized rates of hippocampal atrophy observed in our own community/referral AD patients that were just under 4%^{6, 7}. One plausible explanation for this slight difference is that our community/referral based AD patients average around 80 years of age, whereas the average age in the milameline study was 74 years. It may be that a younger group of AD patients have a slightly more aggressive clinical course, and thus a slightly greater rate of atrophy.

It should be noted that substantial overlap in atrophy rates exists between AD patients in the milameline study and cognitively normal elderly subjects from our own center⁶. The range of annualized % change rates in our normal subjects was -4.8% to 0.2% for hippocampus, and -7.7% to 26.3% for temporal horn⁶. Using these values, 90/192 (47%) of the milameline AD patients' hippocampal rates and 142/192 (74%) of the milameline AD patients' temporal horn rates overlapped into the normal range.

The behavioral/cognitive and MRI measures in Table 2 all changed in the expected direction over time. That is, as behavioral/cognitive performance worsened, hippocampal volume declined, and temporal horn volume increased. However, decline was much more consistently seen with imaging, particularly the hippocampus, than with behavioral/cognitive measures. In fact while 99% of subjects showed the expected decline in hippocampal volume over time, only 60% worsened on the ADAS-Cog and 66% on the MMSE and only 39% on the GDS. Consistent results and high test re-test reproducibility are desirable qualities for an outcome measure in therapeutic trials, provided that the measure is sensitive to the biologic features of interest. Because the drug trial itself was not completed, these data do not address whether MRI measures are better or worse than standard cognitive/behavioral measures as an outcome metric in therapeutic trials. However, these data do demonstrate that MRI measures more consistently follow the expected decline due to disease progression, than widely used behavioral/cognitive measures obtained in the same group of subjects over the same period of time.

Assuming a common standard of a 50% reduction in the rate of decline over one year for all measures, the effect size calculations indicate that sample size required for clinical trials should be substantially smaller for the MRI measures vs these generally used clinical/psychometric measures. In practice, attrition would have to be built into sample size estimates, which we did not do in these calculations. In addition, one could argue that the annual rate of decline observed for each individual measure in cognitively normal elderly subjects should be subtracted from the rate in AD patients in order to assess the rate of change due specifically to disease progression and thus "available for therapeutic modification". We did not build this into the effect size calculations either. Finally, the analyses were based on a 50% reduction in the rate of decline over one year, which may be excessively optimistic. However, the purpose of the analysis was to compare the measures head-to-head using a common criteria in the same group of patients, and in these data, the MRI measures outperformed the clinical/psychometric measures.

Smaller baseline hippocampal volume was associated with a greater rate of hippocampal atrophy. Older age was associated with a greater rate of temporal horn enlargement. However, the magnitude of both of these relationships was only modest, with slopes of -0.15 and -0.19. The most striking relationship in the analyses in Table 3 occurred between baseline temporal horn volume and the rate of temporal horn enlargement. Larger temporal horn volume at baseline was associated with a greater rate of temporal horn enlargement, and the magnitude of this relationship was substantial with a slope of 0.53. It is not clear why patients with greater atrophy at baseline showed greater rates of atrophy over the course of the study, but it might be that baseline temporal horn volume is an indicator of those patients who have more aggressive pathologic progression.

The absence of consistent correlation between the change in behavioral/cognitive measures and change in hippocampal volume was unexpected (Table 4). We expected to find a strong correlation between worsening cognitive performance and greater rates of hippocampal atrophy. One explanation for a failure to observe such a correlation, aside from noise in the data itself, rests with truncation of the range of values in both rates of hippocampal atrophy and in rates of change in cognitive performance in these uniformly selected AD patients. It might also be true that in patients with established AD, the most substantial pathologic disease progression occurs outside the hippocampus, and therefore hippocampal change measurements and cognitive/behavioral change are simply not strongly correlated. This implies that multiple measures should be evaluated in future clinical trials, as the information provided by different imaging measures may be orthogonal to each another.

At the outset, the major emphasis was on hippocampal volume with temporal horn volume representing a secondary MRI measure. An unexpected result, therefore, was that overall correlations between change in cognitive performance and change in MRI volume were greater with the temporal horn than with the hippocampus (Fig 1). Although modest in magnitude with slopes less than 0.30, correlations were present between rate of temporal horn enlargement and all behavioral/cognitive measures (Fig 2). All of the correlations were in the expected direction, that is greater rates of temporal horn enlargement corresponded to greater rates of behavioral/cognitive decline. Interestingly, significant correlations were only present in the multivariate models, which controlled for age, gender, education, and baseline temporal horn volume. This is logical, as change in temporal horn volume was highly associated with baseline temporal horn volume (Table 3). One possible explanation for this discrepancy between observed correlations with the hippocampus vs the temporal horn, is that of the two measures, the temporal horn should be more sensitive to cerebral atrophy outside the medial temporal lobe²⁴. Temporal horn enlargement occurs not just with atrophy of the hippocampus but also as a reflection of atrophy in other medial temporal lobe limbic areas like the entorhinal cortex as well as the remaining temporal lobe including neocortical association areas. Given that all patients in this study were in the mild to moderate phase of AD, all likely had pathologic involvement that extended well beyond the hippocampus.

In spite of established ongoing quality control programs at every one of the participating sites, 19 alerts were triggered by undetected QC problems¹¹. We conclude from this experience that centrally orchestrated MRI QC surveillance is both feasible and necessary in a multi-site trial using MRI as outcome measure.

We compared baseline hippocampal W scores across centers to assess the consistency of data acquire from multiple sites. The fact that only one outlier was identified when comparing baseline W scores across centers, illustrates that MR data can be acquired at multiple sites and analyzed centrally without obvious systematic errors.

Most published MRI studies have been derived from a single institution. We are not aware of any publications that have demonstrated the feasibility of multi-site acquisition and central analysis of structural MRI data as an outcome measure in an AD therapeutic trial. There are many reasons to suspect a priori that multi-site MRI data might not be internally consistent -- differences in scanner hardware, software, scanner QC maintenance programs, etc. We believe the data provided in this paper validates the technical feasibility of using MRI as an outcome measure of disease progression in multi-center AD therapeutic trials. This study does not, however, prove that imaging measures constitute valid biomarkers of therapeutic efficacy. This will require a positive therapeutic trial that has incorporated serial imaging measures in the study design.

Acknowledgments

Supported by Parke-Davis Corp.

Appendix

Appendix: Study Participants

1. Geoffrey Ahern, M.D., Ph.D. The University of Arizona Tucson, AZ 85724-5023
2. Fred Allen, M.D. Carolina Neurological Clinic, PA Charlotte, NC 28203
3. Piero Antuono, M.D. Froedtert Memorial Lutheran Hospital Milwaukee, WI 53226-3522
4. Jeff Apter, M.D. Princeton Biomedical Research, PA
5. Stephen Asher, M.D. Anderson Plaza Medical Building Boise, ID 83702-6130
6. Nancy Barbas, M.D. University of Michigan Medical Center Ann Arbor, MI 48109-0005
7. James Burke, M.D., Ph.D. Duke University Medical Center Durham, NC 27710-0001
8. Gastone Celesia, M.D. Loyola University Medical Center
9. David J. Coffey, M.D. Dartmouth-Hitchcock Medical Center
10. Cal Cohn, M.D. The Cohn Center Psychiatry Houston, TX 77074
11. Kirk R. Daffner, M.D. Reisa Sperling, M.D. Brigham Behavioral Neurology Group
12. Alan Dengiz, M.D. St. Joseph Mercy Hospital Ann Arbor, MI 48106
13. David Drachman, M.D. University of Massachusetts Medical School
14. Barry Gordon, M.D., Ph.D. Johns Hopkins University
15. Neill Graff-Radford, M.D. Mayo Clinic-Jacksonville
16. Linda Hershey, M.D. BVAMC Buffalo, NY 14215
17. Marc Hertzman, M.D. Crain Towers Glen Burnie, MD 21061
18. Mustafa Husain, M.D. Southwestern Medical Center Dallas, TX 75235-9070
19. William Jagust, M.D. UC Davis Medical Center
20. Jeffrey Kaye, M.D. Oregon Health Sciences University
21. Arifulla Khan, M.D. Northwest Clinical Research Center Kirland, WA 98034
22. Ranga Krishnan, M.D. Duke University Medical Center
23. Dennis McMannus, M.D. SIU School of Medicine Springfield, IL 62794-1413
24. Jacob Mintzer, M.D. Medical University of South Carolina Charleston, South Carolina 29425-0742

25. Jorg Pahl, M.D. Pahl Brain Associates, P.C. Oklahoma City, OK 73120
26. Murray Rosenthal, DO Behavioral Medicine Resources San Diego, CA 92123
27. Carl Sadowsky, M.D. Palm Beach Neurological Group West Palm Beach, FL 33407-2441
28. Frederick Schaerf, M.D. Ft. Meyers, FL 33907
29. Douglas Scharre, M.D. Ohio State University Columbus, OH 43210
30. Rachel Schindler, MD Pfizer Inc. New York, NY 10017
31. Joshua Shua-Haim, M.D. Monmouth Medical Center Long Branch, N 07740
32. Paul Soloman, Ph.D. Memory Disorders Clinic Bennington, VT 05201
33. Steven Targum, M.D. CNS Philadelphia Philadelphia, PA 19106
34. Christopher Van Dyck, M.D. Yale University School of Medicine
35. Troy Williams, M.D. Peoria, AZ 85381
36. William Pendlebury, M.D. University of Vermont

References

1. Fox NC, Freeborough PA. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. *Journal of Magnetic Resonance Imaging* 1997;7:1069–75. [PubMed: 9400851]
2. Fox NC, Scahill RI, Crum WR, Rossor MN. Correlation between rates of brain atrophy and cognitive decline in AD. *Neurology* 1999;52:1687–1689. [PubMed: 10331700]
3. Fox NC, Freeborough PA, Rossor MN. Visualization and quantification of rates of atrophy in Alzheimer's disease. *The Lancet* 1996;348:94–97.
4. Fox NC, Cousens S, Scahill R, et al. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease. *Arch Neurol* 2000;57:339–443. [PubMed: 10714659]
5. Freeborough PA, Fox NC. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Trans on Medical Imaging* 1997;15:623–629.
6. Jack CR Jr, Petersen RC, Xu Y, et al. The rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology* 1998;51:993–999. [PubMed: 9781519]
7. Jack CR Jr, Petersen RC, Xu Y, O'Brien PC, Smith Ge, Ivnik RJ, et al. Rates of Hippocampal Atrophy in Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease. *Neurology* 2000;55:484–489. [PubMed: 10953178]
8. Perry E, Tomlinson B, Blessed G, Bergmann K, Givson P, Perry R. Correlation of cholinergic abnormalities with senile plaques and mental test scores in senile dementia. *BMJ* 1978;2:1457–1459. [PubMed: 719462]
9. Whitehouse PJ, Price DL, Strubel RG, Clark AW, Coyle JT, DeLong MR. Alzheimer's disease and senile dementia: loss of neurons in the basal forebrain. *Science* 1982;215:1237–1239. [PubMed: 7058341]
10. Folstein MF, Folstein SE, McHugh PR. "Mini Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975;12:189–198. [PubMed: 1202204]
11. Felmlee JP, Lanners DM, Rettman DW, Hangiandreou NJ, Jack CRJ. MR imaging quality control measurements taken as part of a Multi-center trial: Initial results. *Radiology* 1997;205:619. [PubMed: 9393512]

12. Jack CR Jr, Petersen RC, O'Brien PC, et al. MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology* 1992;42:183–188. [PubMed: 1734300]
13. Jack CR Jr, Petersen RC, Xu YC, et al. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* 1997;49:786–794. [PubMed: 9305341]
14. Jack CR Jr, Twomey CK, Zinsmeister AR, et al. Anterior temporal lobes and hippocampal formations: normative volumetric measurements for MR images in young adults. *Radiology* 1989;172:549–554. [PubMed: 2748838]
15. Jack CR Jr, Sharbrough FW, Twomey CK, Cascino GD, Hirschorn KA, Marsh WR, et al. Temporal Lobe Seizures: Lateralization with MR volume measurements of hippocampal formation. *Radiology* 1990;175:423–429. [PubMed: 2183282]
16. Killiany RJ, Moss MB, Albert MS, et al. Temporal lobe regions on magnetic resonance imaging identify patients with early Alzheimer's disease. *Arch Neurol* 1993;50:949–954. [PubMed: 8363449]
17. Krasuski JS, Alexander GE, Horwitz B, Daly EM, Murphy DG, Rapoport SI, et al. Volumes of medial temporal lobe structures in patients with Alzheimer's disease and mild cognitive impairment (and in healthy controls). *Biological Psychiatry* 1998;43:60–68. [PubMed: 9442345]
18. Laakso MP, Soininen H, Partanen K, et al. Volumes of hippocampus, amygdala and frontal lobes in the MRI-based diagnosis of early Alzheimer's disease: correlation with memory functions. *J of Neural Transmission* 1995;9:73–86.
19. Lehericy S, Baulac M, Chiras J, et al. Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease. *AJNR* 1994;15:927–937.
20. Pearlson GD, Harris GJ, Powers RE, et al. Quantitative changes in mesial temporal volume, regional cerebral blood flow, and cognition in Alzheimer's disease. *Arch Gen Psychiatry* 1992;49:402–408. [PubMed: 1586276]
21. Laakso MP, Lehtovirta M, Partanen K, Riekkinen PJ, Soininen H. Hippocampus in Alzheimer's disease: A 3-year followup MRI study. *Biol Psychiatry* 2000;47:557–561. [PubMed: 10715362]
22. Fox NC, Warrington EK, Freeborough PA, et al. Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study. *Brain* 1996;119:2001–2007. [PubMed: 9010004]
23. Petersen RC, Jack CR Jr, Xu YC, Waring SC, O'Brien PC, Smith GE, et al. Memory and MRI-based hippocampal volumes in aging and Alzheimer's disease. *Neurology* 2000;54:581–587. [PubMed: 10680786]
24. DeCarli C, Haxby JV, Gillette JA, et al. Longitudinal changes in lateral ventricular volume in patients with dementia of the Alzheimer type. *Neurology* 1993;42:2029–2036. [PubMed: 1407587]

Figure 1. One year temporal horn enlargement

Coronal images obtained at base line (top panel) and 12.5 months later (from an 85 year old woman). The dramatic increase in size of the temporal horns in just over 12 months is visually apparent. The temporal horns increased in volume by 3416 mm^3 (36.6%).

Figure 2. Annual change in ADAS-Cog vs. temporal horn

Scatter plot of the annualized percent change in ADAS-Cog score vs. annualized percent change in temporal horn volume.

Table 1
Baseline Demographic, Cognitive, and MRI Data by Treatment Group

	Treated (N = 188)	Placebo (N = 174)	Median of all (n = 362)
Women, N (%)	111 (59%)	105 (60%)	
Age at first MRI (years)	75 (50, 92)	73 (51, 90)	74
Education ¹	4 (1, 6)	4 (1, 6)	4
ADAS-Cog ²	24 (6, 54)	25 (7, 57)	24
MMSE ³	21 (10, 27)	22 (10, 26)	21
GDS ⁴	5 (2, 6)	5 (2, 6)	5
Raw right Temporal Horn mm ³	2258 (583, 7937)	2002 (388, 8049)	2085
Raw left Temporal Horn mm ³	2137 (639, 8942)	1937 (397, 8664)	2006
Raw total Temporal Horn mm ³	4377 (1359, 16119)	3877 (785, 12265)	4140
Raw right hippocampus mm ³	2207 ± 424	2267 ± 433	2223
Raw left hippocampus mm ³	2100 ± 429	2134 ± 434	2102
Raw total hippocampus mm ³	4307 ± 820	4401 ± 839	4326
Hippocampal W-score	-2.09 (-3.42, 0.93)	-1.78 (-3.34, 1.68)	-1.89
Percentile (from W score) ⁵	0.02	0.04	0.03

Table values are mean and sd for hippocampal volume data; median and range for others

¹ 1= Eighth Grade or less; 2= some high school; 3=high school graduate; 4=some college; 5= college graduate; 6=any post graduate work

² Alzheimer's Disease Assessment Scale-Cognitive sub-scale (ADAS-Cog): eight missing in treated group, nine missing in placebo group

³ Mini-Mental State Exam: one missing in treated group, one missing in placebo group

⁴ Global Deterioration Scale: one missing in placebo group

⁵ Percentile ranking of hippocampal W score relative to normal elderly subjects

Table 2
Annual Change from Baseline in Behavioral/Cognitive and MRI Variables

	Raw Change Treated (N = 100)	Raw Change Placebo (N = 92)	Overall Median Raw Change (N = 192)	Overall Median Annual Percent Change (N=192)	Percent Decliners ⁴
ADAS-Cog ¹	4.8 (−10.7, 25.6)	3.5 (−21.5, 19.9)	4.1 (−21.5, 25.6)	16.4 (−59.9, 152.9)	60.4
MMSE	−2.1 (−16.2, 6.4)	−1.1 (−18.1, 7.2)	−1.9 (−18.1, 7.2)	−8.3 (−181.1, 48.6)	66.2
GDS ³	0 (−1.3, 2.4)	0 (−2.4, 2.6)	0 (−2.4, 2.6)	0.0 (−47.6, 95.4)	38.5
Total hippocampal mm ³	−221 (−665, 19)	−220 (−674, −7)	−221 (−674, 19)	−4.9 (−15.2, 0.5)	99.0
Total Temporal Horn volume mm ³	658 (−576, 3241)	497 (−623, 2541)	616 (−623, 3241)	16.1 (−13.1, 53.5)	85.4
Duration between 2 scans in months	12.3 (9, 14)	12.2 (9, 15)	12.2 (9, 15)		

Values in table represent median and range.

² MMSE: six missing in placebo group

¹ Alzheimer's Disease Assessment Scale-Cognitive sub-scale (ADAS-Cog): twelve missing in treated group, fourteen missing in placebo group

³ Global Deterioration Scale: eleven missing in treated group, eight missing in placebo group

⁴ Proportion (%) of individuals who declined over time on a particular measure

Table 3

Factors That Influence Change in MRI Volumes

	Change in Hippocampal Volume				Change in Temporal Horn Volume			
	Beta	SE	Part Corr	P-value	Beta	SE	Part Corr	P-value
Age	-0.00	0.07	-0.00	NS	-0.19	0.06	-0.21	.004
Gender	2.20	8.33	0.02	NS	7.20	7.30	0.07	NS
Education	0.09	0.08	0.08	NS	0.05	0.07	0.05	NS
Baseline Volume ¹	-0.15	0.07	-0.15	.045	0.53	0.07	0.51	<0.001

*rank transformation was used for all data

A multivariate model with annualized change in hippocampal volume as the dependent variable, and with age at baseline, gender, education, and hippocampal W score at baseline as independent variables. The beta (slope), standard error, Spearman's partial correlation, and p value associated with each of the independent variables in the model appears in the Table. A similar multivariate model was constructed with change in temporal horn volume as the dependent variable.

Table 4
Correlation Between Annual Percent Change In Behavioral/Cognitive Variables and Hippocampal Volume Change

	Univariate Models		Multivariate Models	
	Baseline Hippocampal W score	Annual raw change of hippocampus	Baseline Hippocampal W score	Annual raw change of hippocampus
ADAS-Cog	−0.09 (NS)	−0.02 (NS)	−0.10 (NS)	−0.04 (NS)
MMSE	0.06 (NS)	0.07 (NS)	0.03 (NS)	0.11 (NS)
GDS	−0.13 (NS)	0.11 (NS)	−0.22 (0.005)	0.04 (NS)

Univariate analyses were performed with baseline hippocampal W score as the independent variable and annualized percent change in behavioral/cognitive performance as the dependent variable. Likewise, univariate models with annualized raw change in hippocampal volume (mm^3) as the independent variable and annualized percent change in behavioral/cognitive performance as the dependent variables were performed. The Spearman correlations (p value) resulting from each of these univariate analyses appear under the columns labeled Univariate models. A series of multivariate models were constructed with annualized percent change in behavioral/cognitive performance as the dependent variable, and with baseline hippocampal W score, annual raw change in hippocampal volume (mm^3), age, gender, and education as independent variables. The partial Spearman correlations (p value) for the baseline hippocampal W score and the annualized raw change in hippocampal volume for each of these models appears under the columns labeled Multivariate models.

Table 5
Correlation Between Annual Percent Change in Behavioral/Cognitive Variables and Temporal Horn Volume Change

	Univariate Models		Multivariate Models	
	Baseline temporal horn volume	Annual raw change temporal horn volume	Baseline temporal horn volume	Annual raw change temporal horn volume
ADAS-Cog	0.07 (NS)	0.27 (<0.001)	−0.04 (NS)	0.36 (<0.001)
MMSE	−0.22 (0.003)	−0.34 (<0.001)	−0.02 (NS)	−0.23 (0.002)
GDS	0.04 (NS)	0.16 (0.039)	0.05 (NS)	0.22 (0.005)

Univariate analyses were performed with baseline temporal horn volume (mm^3) as the independent variable and annualized percent change in behavioral/cognitive performance as the dependent variable. Likewise, univariate models with annualized raw change in temporal horn volume (mm^3) as the independent variable and annualized percent change in behavioral/cognitive performance as the dependent variables were performed. The Spearman correlations (p value) resulting from each of these univariate analyses appear under the columns labeled Univariate models.

A series of multivariate models were constructed with annualized percent change in behavioral/cognitive performance as the dependent variable, and with baseline temporal horn volume (mm^3), annual raw change in temporal horn volume (mm^3), age, gender, and education as independent variables. The partial Spearman correlations (p value) for the baseline temporal horn volume and the annualized raw change in temporal horn volume for each of these models appears under the columns labeled Multivariate models.