Alzheimer's & Dementia: Translational Research & Clinical Interventions 2 (2016) 177-181

Featured Article

Proof of concept demonstration of optimal composite MRI endpoints for clinical trials

Steven D. Edland[a,b,*], M. Colin Ard[b], Jaiashre Sridhar[c], Derin Cobia[d], Adam Martersteck[c,e], M.-Marsel Mesulam[c,f], Emily J. Rogalski[c]

[a]*Division of Biostatistics, Department of Family Medicine & Public Health, University of California San Diego, La Jolla, CA, USA*
[b]*Department of Neurosciences, University of California San Diego, La Jolla, CA, USA*
[c]*Cognitive Neurology and Alzheimer's Disease Center, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[d]*Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[e]*Department of Radiology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[f]*Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

Abstract

Introduction: Atrophy measures derived from structural MRI are promising outcome measures for early phase clinical trials, especially for rare diseases such as primary progressive aphasia (PPA), where the small available subject pool limits our ability to perform meaningfully powered trials with traditional cognitive and functional outcome measures.

Methods: We investigated a composite atrophy index in 26 PPA participants with longitudinal MRIs separated by 2 years. Rogalski *et al.* [5] previously demonstrated that atrophy of the left perisylvian temporal cortex (PSTC) is a highly sensitive measure of disease progression in this population and a promising endpoint for clinical trials. Using methods described by Ard *et al.* [1], we constructed a composite atrophy index composed of a weighted sum of volumetric measures of 10 regions of interest within the left perisylvian cortex using weights that maximize signal-to-noise and minimize sample size required of trials using the resulting score. Sample size required to detect a fixed percentage slowing in atrophy in a 2-year clinical trial with equal allocation of subjects across arms and 90% power was calculated for the PSTC and optimal composite surrogate biomarker endpoints.

Results: The optimal composite endpoint required 38% fewer subjects to detect the same percent slowing in atrophy than required by the left PSTC endpoint.

Conclusions: Optimal composites can increase the power of clinical trials and increase the probability that smaller trials are informative, an observation especially relevant for PPA but also for related neurodegenerative disorders including Alzheimer's disease.

Keywords: Primary progressive aphasia; PPA; Clinical trial; Sample size; Power calculations; Composite endpoint; Alzheimer's disease; Structural magnetic resonance imaging; MRI; Region of interest

1. Introduction

Clinical trials of interventions to slow the course of chronic progressive neurodegenerative diseases typically use cognitive neuropsychometric and functional (activities of daily living) outcome measures to demonstrate efficacy. Treatment efficacy is difficult to demonstrate with these endpoints, because decline is subtle during the relatively short span of observation of a clinical trial, and because there is substantial random variability in these measures from person to person and from observation to observation within a person. Volumetric magnetic resonance imaging (MRI) on the other hand has been shown to have good signal-to-noise

*Corresponding author. Tel: +1-848-246-1250; Fax: +1-858-246-1287.
E-mail address: sedland@ucsd.edu

http://dx.doi.org/10.1016/j.trci.2016.05.002

properties in this context. For example, for amnestic dementia of the Alzheimer type (AD), a substantial literature has consistently demonstrated that volumetric MRI endpoints could reduce required sample size in AD treatment trials and secondary prevention trials of mild cognitive impairment by 75% or more compared to a standard cognitive function outcome [1]. The need for efficient endpoints is especially critical for rare subtypes of disease where the pool of subjects available for recruitment limits our ability to even perform large-scale phase 3 trials using neuropsychometric and functional outcome measures.

Primary progressive aphasia (PPA) is a clinical dementia syndrome characterized by an initially isolated and progressive decline in language function and is associated with peak atrophy within the left hemisphere perisylvian language network [2–4]. Rogalski *et al.* [5] demonstrated that atrophy of the left perisylvian temporal cortex in particular is a highly sensitive measure of disease progression and a promising endpoint for clinical trials. Using this endpoint, clinical trials as small as 10 participants per arm would have 80% power to detect a 40% slowing of atrophy [5]. Efficiency of trials may be improved beyond these impressive levels by efficient utilization of the richness of data obtained by MRI. Xiong *et al.* [6] proposed that "composite" endpoints calculated as weighted averages of volumetric region of interest (ROI) substructures may outperform simple sums. These methods were operationalized by Ard *et al.* [7], who derived algorithms for determining optimal composite measures that maximize statistical power when used as an endpoint for clinical trials. In this brief communication, we demonstrate the potential utility of composite atrophy measures for clinical trials of neurodegenerative diseases with a prominent atrophy component.

2. Methods

Study subjects and imaging techniques have been previously described [5]. Briefly, study subjects included 26 individuals with a root diagnosis of PPA [2–4] (8 PPA logopenic, 10 PPA agrammatic, 8 PPA semantic). For the purposes of this article, the three clinical subtypes of PPA were combined to insure sufficient sample size to estimate parameters required for calculating weights. Hence, this analysis is best interpreted as a proof-of-concept demonstration of the potential utility of composite volumetric measures, rather than derivation of an endpoint appropriate for use in a clinical trial. Mean age at baseline was 63.7 years (SD = 6.7), 58% were women, mean Boston Naming Test score was 39.5 (SD = 20.9), and mean Western Aphasia Battery Aphasia Quotient score was 86.8 (SD = 8.0). All subjects received a baseline structural MRI and follow-up MRI approximately 2 years later (mean interval 2.0 years).

Structural MRIs were processed using the cross-sectional [8] and longitudinal [9] pipelines from FreeSurfer, version 5.1.0. Ten regions of interest (ROIs) within the left perisylvian temporal cortex region (Fig. 1) taken from the automated Desikan–Killiany cortical parcellation atlas were the components of a composite outcome measure [10]. The composite was calculated as the optimally weighted sum of these ROIs using weights that maximize the signal-to-noise ratio of rate of change on the composite, as previously described [7]. Clinical trail endpoints with high signal-to-noise ratio, also called the mean to standard deviation ratio (MSDR), are more sensitive to treatment effects and optimize the power of a trial. We used relative efficiency to compare the performance of different outcome measure, where relative efficiency is defined as the ratio of sample size required for trials using the respective outcomes calculated using the standard formula for a two-sample *t* test:

$$n/arm=2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\sigma_d^2\Big/\Delta^2$$

where σ_d^2 is the within group variance of the outcome measure being compared across treatments, in this case, the change from baseline to 2-year follow-up, Δ is the treatment effect size under the alternative, and $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the usual quantiles of the standard normal distribution, with α

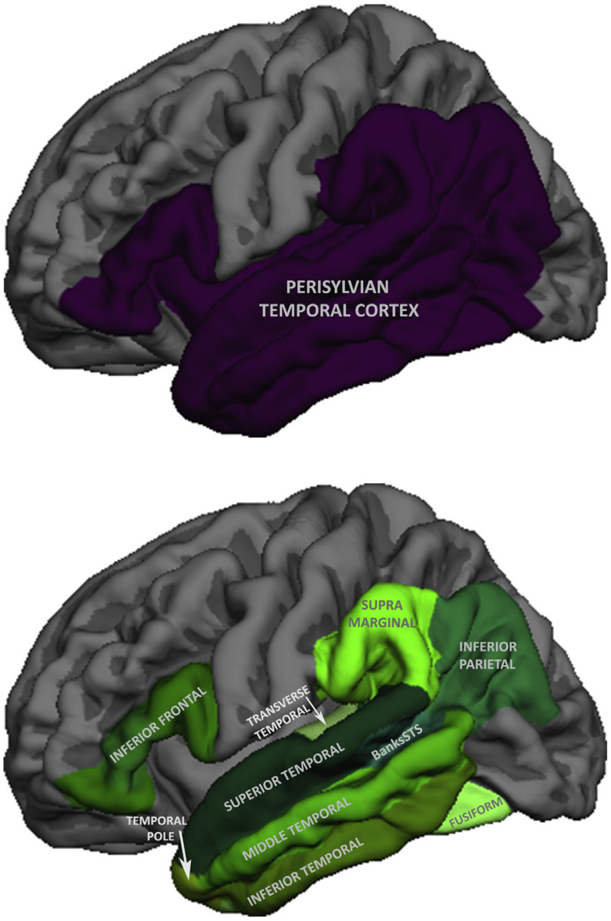

Fig. 1. Regions of interest used to examine longitudinal cortical atrophy in PPA. Top: The perisylvian temporal cortex region of interest defined in Rogalski *et al.* [5]. Bottom: the regions of interest used to create the composite outcome measure.

equal to the type I error rate of a two-sided test, typically set to 0.05, and $(1-\beta)$ equal to the power of the trial, typically set to 0.8 or 0.9.

Indexing two trial outcome measures to be compared as A and B, the relative efficiency of outcome A to outcome B is defined as

$$\frac{2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\sigma_{dA}^2\Big/\Delta_A^2}{2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\sigma_{dB}^2\Big/\Delta_B^2}\times 100\%$$

Let μ_A and μ_B represent the mean change under placebo for outcome measures A and B, respectively. We can express effect sizes as proportional slowing of mean rate of change. For example, to power for a treatment effect that slows atrophy by 25%, we would set $\Delta_A = 0.25 \times \mu_A$ and $\Delta_B = 0.25 \times \mu_B$. Expressing effect sizes in this way, the proportions and the terms involving α and β drop out of the relative efficiency formula, leading to a simple function of MSDRs of the two instruments being compared:

$$\frac{\sigma_{dA}^2/\mu_A^2}{\sigma_{dB}^2/\mu_B^2}\times 100\% = \frac{MSDR_B^2}{MSDR_A^2}\times 100\%$$

For example, a relative efficiency of 50% means a trial using instrument A would require half as many subjects as a trial using instrument B to detect the same percent slowing in rate of decline.

3. Results

Mean rate of decline and person-to-person variability in rate of decline are summarized in the Table 1. The MSDR for the referent total left perisylvian temporal cortical volume endpoint is 3.90. The MSDR for the component subregions of the left perisylvian temporal cortex range from 1.58 to 3.36, consistently below the MSDR of the left perisylvian temporal cortex, meaning that components individually would be less sensitive to atrophy than the full perisylvian ROI. The MSDR of the composite atrophy measure is 4.95, over 25% larger than the left perisylvian temporal cortex MSDR. In terms of relative efficiency, the optimal composite endpoint requires 38% fewer subjects than the total perisylvian temporal cortex volume measure to detect the same percent slowing in atrophy. For example, with 90% power, only 15 subjects per arm would be required to detect a 25% slowing in rate of progression in the composite outcome measure assuming the distribution of decline observed in our pilot study.

For comparison, we also report relative efficiency and sample size projections for clinical trials using total cortical volume or specific neuropsychometric instruments as the primary outcome measure. Relative to the full perisylvian temporal cortex ROI, the total cortical volume endpoint would require 40% more subjects, and the neuropsychometric outcomes would require more than ten times more subjects per arm (Table 1).

4. Discussion

We have described a relatively intuitive and accessible volumetric composite index defined simply as the (optimally) weighted sum of ROI volumes. In our example, the resulting outcome measure substantially improved the efficiency of clinical trials in PPA, reducing required sample size relative to the total perisylvian cortical volume outcome by 38%. Other summaries of MRI data for this purpose have been proposed. Less intuitive and accessible perhaps are mathematically derived atrophy indexes, for example, the weighted average of vertices summarizing ventricular morphometry [11]. In the other extreme, using the single

Table 1
Test characteristics and relative efficiency of various potential clinical trial outcomes measures

Measure	FreeSurfer region of interest	Mean 2 year decline	Standard deviation of 2-year decline	MSDR	Relative efficiency	N/arm to detect 25% slowing*
Boston Naming Test (maximum score: 60)		15.7	14.2	1.10	11.08	277/arm
Western aphasia battery -revised, aphasia quotient (maximum score: 100)		22.9	18.2	1.26	8.60	215/arm
Total cortical volume (mm^3)	rh.cortex + lh.cortex	29,397	10,526	2.79	1.40	35/arm
Left perisylvian temporal cortex (mm^3)		7811	2004	3.90	1.00	25/arm
Components, optimal composite						
Left superior temporal gyrus (mm^3)	lh.superiortemporal	1041	310	3.36		
Left middle temporal gyrus (mm^3)	lh.middletemporal	1128	403	2.79		
Left inferior temporal gyrus (mm^3)	lh.inferiortemporal	955	297	3.22		
Left banks, sup. temp. sulcus (mm^3)	lh.bankssts	214	71	3.00		
Left fusiform gyrus (mm^3)	lh.fusiform	810	287	2.82		
Left transverse temporal gyrus (mm^3)	lh.transversetemporal	80	50	1.58		
Left temporal pole (mm^3)	lh.temporalpole	180	102	1.77		
Left inferior frontal gyrus (mm^3)	†multiple regions	768	345	2.23		
Left inferior parietal gyrus (mm^3)	lh.inferiorparietal	1075	417	2.58		
Left Supramarginal gyrus (mm^3)	lh.supramarginal	880	352	2.50		
Optimal composite index		252	51	4.93	0.62	15/arm

*Two-year clinical trial comparing change baseline to year 2 in treatment versus control, two-sided test, $\alpha = 0.05$, power = 90%.
†lh.parsopercularis + lh.parstriangularis + lh.parsorbitalis.

ROI most sensitive to disease, as example the perisylvian temporal cortex in PPA [5] or the hippocampus [1] or frontal lobe [11] in AD is perfectly intuitive and accessible. Determining the relative efficiency of these various approaches, as we have demonstrated here, will be a useful tool for clinicians weighting the tradeoffs of accessibility versus power when selecting endpoints for clinical trials.

There are limitations to this report. The relatively small sample size in this cohort and lack of information about the underlying neuropathology required pooling of etiologically disparate disease entities for the purpose of demonstrating the optimal compositing methodology. Hence, sample size estimates from this report are only for illustrative purposes. We emphasize that meaningful estimation of optimal weighting parameters will require substantially larger, representative samples than used in this proof-of-concept demonstration. To simplify presentation, we ignored the influence of "normal" age-associated atrophy. Age-associated atrophy may not respond to treatment, and ignoring the influence of age-associated atrophy may lead to under estimation of detectible effect size and overstatement of power. This is a substantial concern for typical amnestic AD [12,13]; but is less of an issue for PPA, where onset age is typically $<$65 years, and there is a rapid rate of disease-associated atrophy relative to normal aging. Thus age-associated atrophy is likely to have a negligible effect on the relative efficiency calculations that are the focus of this manuscript. Finally, an implicit assumption of the optimal compositing method is that treatment slows the rate of atrophy proportionally in all ROIs. This is a plausible assumption, but one that cannot be formally tested until an effective treatment is identified.

Biomarker endpoints have clear limitations. There is no guarantee that treatments positively affecting biomarkers will have corresponding effects on cognitive and functional outcomes, and biomarkers will have been validated as surrogates for clinical endpoints before they will be approved as primary endpoints for phase 3 clinical trials [14]. However, surrogate endpoints including volumetric MRI are currently being used in phase 2 trials, to demonstrate target engagement, and to guide the choice of compounds to move forward to phase 3 [15]. To this end, volumetric endpoints are certainly suggested for diseases like PPA with a prominent atrophy component.

Clinical trials of chronic progressive disease are prohibitively expensive. In AD research, this has limited our ability to test new treatments and find a cure for the disease. For less common phenotypes such as the subtypes of PPA, the need for more efficient endpoints is even more pressing because the available participant pool for clinical trials is limited. Every participant enrolled in a clinical trial is a precious resource, and methods to optimally use all information obtained from participants enrolled in clinical trials and increase the probability that effective treatments are identified should be fully investigated. Optimal weighting maximizes signal-to-noise of endpoints and statistical efficiency of trials. To our knowledge, this is the first meaningful application of optimal weighting to volumetric MRI measures. The real world context is need for more cost-effective and informative clinical trials to speed the development of treatments for neurodegenerative diseases. Primary progressive aphasia, a relatively rare (limited subject pool) disease with prominent atrophy component, is perhaps the perfect laboratory for investigating the performance of alternative surrogate volumetric MRI biomarker endpoints for clinical trials.

Acknowledgments

This project was supported by National Institute on Aging grant AG047580, Optimizing Outcome Measures for Clinical Trials in Pre-Clinical Alzheimer's Disease (SDE, MCA); National Institute on Aging grant AG005131, University of California, San Diego Shiley Marcos Alzheimer's Disease Research Center (SDE, MCA); National Institute of Health grant UL1TR001442, University of California San Diego Clinical and Translational Research Center (SDE); National Institute on Deafness and Other Communication Disorders grant DC008552 (MM); National Institute on Aging grant AG13854 Northwestern University Alzheimer Disease Center (MM); National Institute of Neurological Disorders and Stroke grant NS075075 (MM).

RESEARCH IN CONTEXT

1. Systematic review: Rogalski et al. [5] previously demonstrated that atrophy of the left perisylvian temporal cortex is a highly sensitive measure of disease progression in primary progressive aphasia and a promising endpoint for clinical trials. Using methods described by Ard et al. [1], we constructed a composite atrophy index composed of an optimally weighted linear combination of focal volumetric measures from 10 ROIs within the left perisylvian temporal cortex. Optimal weighting maximizes signal-to-noise and statistical efficiency of clinical trials, and in this application e.g., reduced sample size requirements by 38%.

2. Interpretation: Optimal composite outcome measures show promise as a way improved efficiency of trials. More cost-effective and informative clinical trials would speed the development of treatments for neurodegenerative diseases.

3. Future directions: This proof of concept analysis demonstrated the potential utility of composite volumetric measures to improve the efficiency of trials. We will need larger datasets representative of future clinical trials to more definitively establish the utility of these methods.

References

[1] Ard MC, Edland SD. Power calculations for clinical trials in Alzheimer's disease. J Alzheimers Dis 2011;26 Suppl 3:369–77.

[2] Mesulam MM. Primary progressive aphasia–a language-based dementia. N Engl J Med 2003;349:1535–42.

[3] Mesulam M, Weintraub S. Primary progressive aphasia and kindred disorders. Handb Clin Neurol 2008;89:573–87.

[4] Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. Neurology 2011;76:1006–14.

[5] Rogalski E, Cobia D, Martersteck A, Rademaker A, Wieneke C, Weintraub S, et al. Asymmetry of cortical decline in subtypes of primary progressive aphasia. Neurology 2014;83:1184–91.

[6] Xiong C, van Belle G, Chen K, Tian L, Luo J, Gao F, et al. Combining multiple markers to improve the longitudinal rate of progression-application to clinical trials on the early stage of Alzheimer's disease. Stat Biopharm Res 2013;5.

[7] Ard MC, Raghavan N, Edland SD. Optimal composite scores for longitudinal clinical trials under the linear mixed effects model. Pharm Stat 2015;14:418–26.

[8] Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 1999;9:179–94.

[9] Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 2012;61:1402–18.

[10] Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31:968–80.

[11] Gutman BA, Wang Y, Yanovsky I, Hua X, Toga AW, Jack CR Jr, et al. Empowering imaging biomarkers of Alzheimer's disease. Neurobiol Aging 2015;36 Suppl 1:S69–80.

[12] McEvoy LK, Edland SD, Holland D, Hagler DJ, Roddey JC, Fennema-Notestine C, et al. Neuroimaging enrichment strategy for Secondary Prevention Trials in Alzheimer Disease. Alzheimer Dis Assoc Disord 2010;24:269–77.

[13] Edland SD, ed. Which MRI Measure is Best for Alzheimer's Disease Prevention Trials: Statistical Considerations of Power and Sample Size. *2009 Joint Statistical Meetings Proceedings*. Alexandria, VA: American Statistical Association; 2009:4996-9.

[14] Vellas B, Andrieu S, Sampaio C, Coley N, Wilcock G, European Task Force Group. Endpoints for trials in Alzheimer's disease: a European task force consensus. Lancet Neurol 2008;7:436–50.

[15] Moulder KL, Snider BJ, Mills SL, Buckles VD, Santacruz AM, Bateman RJ, et al. Dominantly inherited Alzheimer network: facilitating research and clinical trials. Alzheimers Res Ther 2013;5:48.