Published in final edited form as:
Hum Brain Mapp. 2013 November ; 34(11): . doi:10.1002/hbm.22120.

Sample Size Estimates for Well-Powered Cross-Sectional Cortical Thickness Studies

Heath R. Pardoe[1,2], **David F. Abbott**[1,2], **Graeme D. Jackson**[1,2,3,*], and **The Alzheimer's Disease Neuroimaging Initiative**[†]

[1]Brain Research Institute, Florey Neuroscience Institutes, Melbourne Brain Centre, Austin Hospital, Heidelberg, Victoria, Australia

[2]Department of Medicine, The University of Melbourne, Victoria, Australia

[3]Department of Radiology, The University of Melbourne, Victoria, Australia

Abstract

Introduction—Cortical thickness mapping is a widely used method for the analysis of neuroanatomical differences between subject groups. We applied power analysis methods over a range of image processing parameters to derive a model that allows researchers to calculate the number of subjects required to ensure a well-powered cross-sectional cortical thickness study.

Methods—0.9-mm isotropic T_1-weighted 3D MPRAGE MRI scans from 98 controls (53 females, age 29.1 ± 9.7 years) were processed using Freesurfer 5.0. Power analyses were carried out using vertex-wise variance estimates from the coregistered cortical thickness maps, systematically varying processing parameters. A genetic programming approach was used to derive a model describing the relationship between sample size and processing parameters. The model was validated on four Alzheimer's Disease Neuroimaging Initiative control datasets (mean 126.5 subjects/site, age 76.6 ± 5.0 years).

Results—Approximately 50 subjects per group are required to detect a 0.25-mm thickness difference; less than 10 subjects per group are required for differences of 1 mm (two-sided test, 10 mm smoothing, $\alpha = 0.05$). Sample size estimates were heterogeneous over the cortical surface. The model yielded sample size predictions within 2–6% of that determined experimentally using independent data from four other datasets. Fitting parameters of the model to data from each site reduced the estimation error to less than 2%.

Conclusions—The derived model provides a simple tool for researchers to calculate how many subjects should be included in a well-powered cortical thickness analysis.

Keywords

MRI; neuroimaging; study design; power analysis; morphometry; cortical thickness

[*]Correspondence to: Graeme D. Jackson, Brain Research Institute, Florey Neuroscience Institutes, Melbourne Brain Centre, Austin Hospital, 245 Burgundy St, Heidelberg, Victoria 3084, Australia. BRI@brain.org.au.

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu)

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/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Additional Supporting Information may be found in the online version of this article.

INTRODUCTION

Coregistration of cortical thickness maps derived from whole-brain T_1-weighted MRI, and subsequent vertex-wise statistical inference, is a popular method for localizing differences in cortical gray matter between groups of subjects. The method allows the investigator to noninvasively determine how disease or environmental factors are related to neuroanatomy. Researchers and funding bodies use power analyses to provide guidance on how many subjects should be included in a study to have a good chance of detecting a real effect, balancing the probability of a false negative (type II error) with the desire not to waste valuable resources studying more subjects than necessary. The aim of this study was to derive a model that allows researchers to determine the number of subjects per group that should be included in a cross-sectional cortical thickness analysis, as a function of image processing parameters, to ensure the analysis is well-powered.

Failure to detect an existing thickness difference is an example of a type II error. Methods for controlling the probability of making a type II error are collectively known as power analysis. Power is defined as $1 - \beta$, where β is the type II error rate. Traditionally, standard power for a well-designed study is equal to 0.8. There are three factors, described by Cohen [1988], that affect the power of a study. These three factors will be described in the context of cross-sectional cortical thickness studies:

1. Effect size: In cortical thickness studies, the effect size corresponds to cortical thickness differences between groups. A study aiming to detect a larger thickness difference than a similarly designed study aiming to detect a smaller thickness difference will have higher power. In this study, we investigated the detection of hypothetical vertex-wise thickness differences between 0.125 and 1 mm. We have used an unstandardized measure of effect size, the thickness difference in mm, in preference to standardized measures such as Cohen's *d*, because the thickness of the cortex is a property with an easily interpretable physical meaning [Wilkinson, 1999].
2. How well the sample resembles the population: Cortical thickness studies rely on sampling to draw inferences about the population of interest. The more subjects that are randomly sampled, the closer the sample will resemble the population. More included subjects will always improve the power of a study.
3. Type I error rate alpha (α): The type I error rate determines the standard of proof required to declare a thickness difference statistically significant. A typical vertex-wise cortical thickness study may involve a few hundred thousand vertices, and subsequent statistical inferences, over the cortical sheet. These analyses are therefore described as "mass univariate." Mass univariate analyses mean that the standard α level of 0.05 will give an unacceptably high level of vertex-wise type I errors (false positives). The multiple comparisons problem is normally accounted for by lowering α. A lower α constitutes a higher standard of proof; because of the more stringent standard of proof, a vertex-wise cortical thickness study will always be poorly powered relative to a nonmass univariate study with the same number of subjects. The effect of lowering α on the number of subjects required for a well powered analysis was investigated in this study.

There are two further factors specific to cross-sectional cortical thickness analyses that affect the power of a study. The first is the variability of vertex-wise cortical thickness estimates over the cortical sheet. Because a typical cortical thickness analysis is mass univariate, power analyses should be carried out at each vertex. In this way, we can map the number of required subjects for a given set of parameters (thickness difference, power level, α, and variance). Mapping the number of subjects in each group allows us to determine the

spatial variability of the number of required subjects over the cortex. It is possible that some cortical regions will require more subjects for a well powered investigation than other regions.

The second factor specific to cross-sectional cortical thickness analyses is the spatial extent of surface-based smoothing applied to the coregistered cortical thickness maps. Surface-based smoothing is necessary to make the data more normally distributed, improving the validity of the statistical tests used to make inferences. The smoothing also corrects for residual misalignment following coregistration. The spatial extent of the smoothing biases the analysis towards detecting focal thickness differences of the same spatial extent due to the matched filter theorem. We investigated the relationship between spatial smoothing and the required number of subjects for a well-powered study.

In this study we use a genetic programming approach to empirically model the relationship between the number of subjects required per group for a well-powered cross-sectional cortical thickness study and the size of the thickness difference (effect size), the type I error rate, the smoothing filter and the "sidedness" of the statistical test, that is, one- or two-sided tests [Schmidt and Lipson, 2009]. Our model will allow researchers to estimate how many subjects per group they need to scan to detect thickness differences of a given magnitude. The model will be validated on MRI data acquired from different scanners and subject cohorts. The validation procedure will determine how useful the derived model will be for other research groups. Model parameters are provided that allow researchers to tailor the number of required subjects to specific cortical regions. Novel aspects of the study include mapping the number of required subjects over the cortical sheet, and providing a simple equation for calculating the number of subjects per group based on cohort- and image-processing parameters.

METHODS

Participants and Image Acquisition

Ninety-eight neurologically normal controls (53 females, age 29.1 ± 9.7 years) were included in the study. All participants provided informed consent. Whole brain T_1-weighted 3D MPRAGE MRI was acquired on a 3-T Siemens TIM Trio Scanner. Image acquisition parameters were as follows: TR = 1900 ms, TI = 900 ms, TE = 2.6 ms, flip angle = 9°, voxel resolution = 0.9 mm isotropic.

Four additional control MRI datasets were used to validate the sample size model derived as part of this study. These data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.lo-ni.ucla.edu). The datasets consisted of whole brain T_1-weighted 3D MPRAGE image acquisitions with voxel resolutions of 1 mm by 1 mm in-plane and 1.2-mm slice thickness [Jack Jr et al., 2008]. Two 3 T datasets were analyzed, designated by ADNI as Normal-bl-3.0T and Normal-m12–3.0T, comprising 60 subjects (38 females, mean age 75.2 ± 4.8 years) and 54 subjects (34 females, mean age 76.3 ± 5.0 years), respectively. Two 1.5-T datasets were also analyzed, designated Normal-m06-1.5T and Normal-m24-1.5T, comprising 214 subjects (101 females, mean age 76.6 ± 5.1 years) and 178 subjects (85 females, mean age 78.1 ± 4.9 years), respectively.

Image Processing

Cortical thickness mapping and intersubject coregistration were carried out using the standard Freesurfer 5.0 processing stream [Fischl and Dale, 2000]. Individual cortical thickness maps were coregistered to the supplied "fsaverage" template. Coregistered cortical thickness maps were smoothed using the surface-based smoothing filter supplied with the Freesurfer distribution. Smoothing with spatial extents of 5, 10, 15, 20, and 25 mm full

width at half maximum was applied to each subject. The general linear model was used to estimate and correct for the effects of age and sex at each vertex in the coregistered, smoothed cortical thickness maps. The effect of age and sex correction on the sample size estimates was investigated by conducting a vertex-wise sample size calculation on both corrected and uncorrected thickness maps.

Power Analysis

Standard methods for power analysis, based on normal distribution statistics, were used to calculate the number of subjects required in each group to adequately control for the likelihood of a type II error. Standard deviation was estimated vertex-wise from the coregistered, smoothed cortical thickness maps. Power analyses were conducted for hypothetical effect sizes of 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, and 1 mm. Direct measures of effect size in mm were used in this study, as one of the advantages of cortical thickness mapping is that the technique measures a property of the cortex that has physical meaning. However, given the nonuniform thickness of the cortex, some researchers may prefer to use standardized measures such as percentage change in cortical thickness. An analysis was undertaken in which sample sizes required to detect a hypothetical 10% change in cortical thickness was measured. Power was set at 0.8 and sample size calculations based on a two-sample *T*-test were used. One- and two-sided power analyses were performed. The type I error rate α was set at 0.05, 0.025, 0.005, 0.0025, and 0.0005. Power analyses were carried out using the power. t.test function provided with the statistical software package "R" (http://www.r-project.org/).

The calculated minimum sample size required for adequate power was calculated at each vertex and mapped back onto the fsaverage template. In this way, whole brain maps of minimum sample size were derived using the above parameters. The 95th percentile of the distribution of vertex-wise minimum sample size estimates was used to calculate whole-brain and lobar values. The 95th percentile was chosen to ensure almost complete coverage of the brain or lobe of interest without being biased by the upper 5% of vertices. The PALS-B12 lobar atlas [Van Essen, 2005] provided with the Freesurfer 5.0 distribution was used to estimate the number of subjects required for an adequately powered study on a per-lobe basis. In summary, the effects of changing effect size, smoothing filter, type I error rate (α) and one-sided and two-sided analyses on the minimum sample size for a well-powered cross-sectional cortical thickness analysis, mapped over the whole cortical sheet, were explored in this study. The large number of independent variables (effect size, smoothing filter, type I error rate and "sidedness") made reporting the estimates of minimum number of subjects per group as a function of these variables unwieldy. An empirical formula allowing for the calculation of the minimum number of subjects per group as a function of the independent variables was calculated using the genetic programming approach implemented in the software package "Eureqa" (http://www.eureqa.com, version 0.83 beta, Schmidt and Lipson [2009]). Eureqa was used to search for a function f such that

$$\text{Number of subjects per group} = f(\text{effect size},\ \text{smoothing}, \alpha,\ \text{is.one.sided})$$

where effect size is the thickness difference in mm, smoothing is the extent of the smoothing filter as described above, α is the type I error rate, and is.one.sided is a binary variable set to 1 for a one sided analysis and 0 for a two-sided analysis. The symbolic building blocks used to obtain the solution were constrained to constants, addition, subtraction, multiplication, division, square root, and logarithmic operators. The Eureqa software internally validates the derived model by subdividing the input data into a training set, used to derive an equation

describing the relationship between the explanatory variables, and a test set used to evaluate the derived equation.

The form of the equation derived in the previous section was used to estimate lobar-specific relationships between the number of subjects in each group and the previously described explanatory variables. Constant parameters in the previously derived whole cortex equation were reevaluated on a per-lobe basis using nonlinear least-squares estimation.

The applicability of the derived model to MRI data acquired at different sites was evaluated by comparing whole brain sample sizes estimated using the derived model with experimentally determined sample sizes evaluated from the four ADNI datasets. The mean absolute error, expressed as a percentage of the experimentally determined sample size, was used to evaluate how well the model estimated the sample size. Two sets of constant parameters were used; the first were the default parameters derived using the model fitting procedure described above, and the second utilized constant parameters derived by applying nonlinear least squares estimation of the model parameters using sample size estimates from each ADNI dataset.

Comparison with Previous Studies

A previous study has reported that seven subjects per group are required to detect a change of 0.2 mm in a cross-sectional analysis [Han et al., 2006]. In order to test the comparability of our derived estimates of the number of subjects per group, an analysis was undertaken using the same parameters from the cited study; namely a thickness difference of 10% at each voxel, $\alpha = 0.05$, one-sided analysis, power = 0.9 and a surface-based smoothing kernel with FWHM extent of 6 mm. Due to the use of slightly different parameters to the main body of this study (power = 0.9 and surface smoothing = 6 mm FWHM) these results will be presented separately from the primary analysis. A similar analysis was undertaken to compare estimated numbers with the analysis presented in Lerch and Evans [2005]. In this case, we used a thickness difference of 0.6 mm, $\alpha = 1.222 \times 10^{-4}$ (calculated from the reported adjusted *t*-threshold of 4.67 and 24 degrees of freedom), surface-based smoothing = 30 mm and power = 0.95. It should be noted that the Lerch et al. study used a 3D smoothing filter to obtain an estimate of 25 subjects per group; in our case, as previously noted, we used a surface-based smoothing filter.

RESULTS

In the following summary of results, the default parameters used are a thickness difference of 0.25 mm, spatial smoothing of 10 mm, a type I error rate of 0.05 and two-sided analyses, except where indicated. Mapping the distribution of the number of subjects per group required for a well-powered cross-sectional cortical thickness analysis reveals considerable heterogeneity over the cortical surface (Fig. 1). Regions such as the anterior temporal lobe, insula, and supra-marginal gyrus require considerably more subjects than other cortical regions in order to reliably detect a cortical thickness difference of the same magnitude. A map of the number of subjects per group required to detect a 10% change in cortical thickness, with other parameters held at their default values, was generated and is provided as Supporting Information Figure 1. Correcting for age and sex revealed a modest but consistent reduction in sample size estimates over the cortex. For the default parameters listed above, the average reduction in subjects per group for a well-powered analysis was 1.35 subjects (Supporting Information Fig. 2).

Using the PALS-B12 lobar atlas allows the estimation of the minimum number of required subjects on a per-lobe basis (Fig. 2). The per-lobe analysis indicates that, in order to cover 95% of each lobe, the frontal, parietal and occipital lobes require ~30 subjects, whereas the

temporal lobe requires approximately fifty subjects to detect a 0.25-mm thickness difference (10 mm spatial smoothing, α = 0.05, two-sided analysis). In order to cover 95% of limbic structures, 234 subjects would need to be included to detect a 0.25-mm thickness difference. In our study limbic structures refer to the medial surface of each hemisphere of the cortex, primarily encompassing the cingulate gyrus and parahippocampal gyrus. These results indicate that cortical thickness measurements in the limbic structures have a high variance, and cross-sectional comparisons of cortical thickness in the limbic structures will be underpowered relative to the other cortical lobes.

The empirical relationship between the minimum number of subjects per group and effect size, smoothing, α and one- or two-sided tests is described by Eq. (1):

$$N = \frac{k_1.smooth.\log(\alpha)}{k_2.\theta^2} + k_3.\log(\alpha) + \frac{k_4 + k_5.\log(\alpha) + k_6.p}{\theta^2(k_2 + smooth.)} \quad (1)$$

where N is the number of subjects per group, θ is the thickness difference in mm, *smoothing* is the extent of the surface-based smoothing kernel in mm, α is the type I error rate and p is a binary variable with $p = 1$ for one-sided analyses and $p = 0$ for two-sided analyses. Whole-brain and per-lobe estimates of the parameter values $k_{1..6}$ are provided in Table I. The mean absolute error of the difference between the number of subjects per group predicted by the model and those derived from the imaging dataset is less than one subject for each cortical region, indicating that the model is an excellent fit.

The dataset used to derive the above equation, and an implementation of the equation in the software language R, are provided at http://www.brain.org.au/software/cortex/power. The nonlinear fitting procedure was unable to obtain usable parameter estimates for the limbic structures using the model presented in this study. If the reader is interested in estimates of the number of subjects per group required for the limbic structures, refer to the data-set provided at the link above.

The minimum number of subjects required to reliably detect a given thickness difference over the cortical surface is reduced as the thickness difference increases (Fig. 3). In order to reliably detect a cortical thickness change of 0.25 mm over 95% of the entire cortical surface, around 60 subjects are required in each group with surface-based smoothing of 10-mm FWHM. For a thickness difference of 1 mm, less than 10 subjects are required in each group. The spatial extent of the smoothing filter has a strong effect on the number of subjects required for a well-powered analysis, with the number of subjects required to detect 0.25 mm ranging from greater than 160 when no smoothing is applied to around 20 subjects when a large smoothing filter of 25 mm is applied (Fig. 4).

A more stringent statistical threshold corresponds to a lower α level. We investigated the effect of applying a more stringent threshold as this approach is the standard technique for controlling excessive false positives associated with the multiple comparisons problem. Each order of magnitude decrease in α, for example from 0.05 to 0.005, requires approximately 40 more subjects to achieve the same level of power (Fig. 5).

Comparing model-derived sample size estimates with those calculated using control data acquired from different sites (and age range of the control subjects) indicates that the mean percentage error varies from 1.96 to 6.28% (Table II). Re-evaluating the constant parameters by fitting the derived model to site-specific data using a nonlinear least squares estimation approach reduces the mean percentage error to between 1.35 and 1.72%.

By applying the parameters reported in a previous study to our cohort ([Han et al., 2006] effect size 0.2 mm, 6 mm FWHM smoothing kernel, power = 0.9, and α = 0.05), we found

that the number of subjects required in each group for a well-powered cortical thickness analysis to cover 95% of the cortex is 121 subjects per group. For coverage of 50% of the cortex, which would be obtained if the average of the vertex-wise across-subjects standard deviation was used in the power calculation (as per Han et al 2006), 27 subjects would be required per group. Following parameters reported in Lerch et al. [2005] (effect size of 0.6 mm, 30 mm FWHM surface-based smoothing kernel, power = 0.95, and $\alpha = 1.22 \times 10^{-4}$, calculated from a *t*-threshold of 4.67), 14 subjects per group would be required to detect a thickness difference of 0.6 mm over 95% of the cortical surface.

DISCUSSION

We have derived sample-size estimates over the surface of the cortical sheet for the detection of cortical thickness differences between two groups of equal number and variance, assuming normally distributed test statistics. The primary outcome of this study is a simple equation that allows researchers to estimate the number of subjects per group required for a well-powered cross-sectional cortical thickness study as a function of study-specific parameters, including the thickness difference to be detected, the applied level of smoothing, and the type I error rate. The number of subjects required per group for a well-powered cross-sectional cortical thickness analysis is heterogeneous over the surface of the cortical sheet. The heterogeneous distribution may be due to natural variability in cortical thickness over the cortical surface, acquisition-based variability, or difficulty modeling the cortical surface in regions of high topological complexity. Brain regions that require a low number of subjects for a reasonably powered study include the central sulcus, the sylvian fissure, and the calcarine and parieto-occipital fissures. The low variance in these regions is most likely because these cortical folds are consistent across individuals, and across-subject registration in Freesurfer is based on aligning cortical folding patterns.

The variability of sample size estimates across the cortex has important implications for the interpretation of the results of previous studies, as well as future study planning. If a study has reported a significant thickness difference in the frontal lobes, for example, it is possible that the sample sizes used in the study did not provide enough power for the detection of the same cortical thickness difference in the temporal lobes. Conversely, given the similarity in required sample sizes in the frontal lobe and parietal lobes (and even less in the occipital lobes, Fig. 2), the researcher could be more confident that the absence of a similar effect in the parietal and occipital lobes is “real” and that they have not made a type II error in these regions.

The validation of our derived model against sample size estimates calculated from additional control datasets from the ADNI study provide supportive evidence that the derived model can provide useful guidance for prospective studies carried out at other sites and over different age ranges. The low percentage difference between the model and empirically derived sample size estimates (final column in Table II, mean less than 1.72% for all four cohorts) suggests that the derived model appropriately describes the relationship between image processing parameters and sample size. If no control data is available at the site, and the image acquisition parameters are reasonably similar to the parameters in this study (and appropriate for cortical thickness mapping), we recommend using the mean percentage errors and standard deviations provided in Table II to modify sample size estimates to ensure a well powered study is carried out over the whole cortex. A conservative approach would be to calculate a sample size using Eq. (1), then adding 9.98% (= 6.28 + 2 × 1.85, mean + 2 × SD) to the estimated value to account for across site variability. If control data are available, the methods described in this article could be used to derive more appropriate sample size estimates that are likely to be lower than the estimates derived from the approach just described. We have provided a software package that allows researchers to

estimate the number of subjects required for a cortical thickness study at http://www.brain.org.au/software/cortex/power. The provided software can calculate sample sizes whether preliminary control data are available or not. We have not investigated pediatric populations in this study. These groups have previously documented nonlinear age related thickness changes [Sowell et al., 2004; Shaw et al., 2006]. In these circumstances our estimates may be too low; for a more accurate estimate we recommend carrying out a vertex-wise power analysis on preliminary data using the routines provided.

It should be noted that our sample size calculations are based on Freesurfer-derived cortical thickness estimates. The validation of cortical thickness estimates derived from Freesurfer mean that we are confident that thickness estimates are representative of the actual cortical thickness over the cortical sheet [Rosas et al., 2002; Kuperberg et al., 2003]. The same statement cannot be made about alternative methods for mapping cortical thickness that have not necessarily been subject to the same level of scrutiny. Therefore we recommend against applying the sample size estimates from our derived model to non-Freesurfer-based cortical thickness studies. However we do recommend undertaking a vertex-wise or voxel-wise power analysis, as described in this study, for alternative cortical thickness mapping methods. The software package provided at http://www.brain.org.au/software/cortex/power could be easily modified for this task. Similarly, the use of lower quality MRI scans than that used in this study, whether a lower spatial resolution or poorer contrast, will mean that more subjects will be required to achieve an adequately powered study. The authors of Freesurfer provide guidelines on minimum standards for data quality (available at http://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferBeginnersGuide at the time of publication). Most modern clinical MRI scanners are capable of acquiring 3D structural T_1-weighted MRI scans that meet these standards. A potential future application of the techniques presented in this article is to provide a direct quantitative estimate of the benefit of improvements in MR image acquisition and analysis for cross-sectional studies.

Both one- and two-sided analyses are presented as the choice depends on whether the investigator has a prior hypothesis regarding the direction of the cortical change. For example the majority of studies in neurodegenerative disorders such as Alzheimer's disease report focal cortical thickness reductions in the patient group [Lerch et al., 2005]. Normal aging in adults is also associated with global cortical thickness reduction [Salat et al., 2004]. However cortical thickness increase may be associated with the patient group, such as focal cortical dysplasia [Bernasconi et al., 2001], carriers of genes associated with schizophrenia susceptibility [Cerasa et al., 2011], and autism [Hardan et al., 2006]. If a researcher wishes to restrict themselves to looking for cortical thickness changes in one direction (e.g., thickness decrease in the patient group), the estimated subject group numbers based on one-sided analyses may be used. For research studies with no prior hypotheses as to the direction of the thickness change, subject group numbers based on the two-sided analyses are appropriate. Given that cortical thickness mapping is a fairly new technique, we recommend that any prospective studies should base their sample size estimates on a two-sided analysis.

By accounting for spatial variability, our method allows the number of subjects required to be better tailored to the particular cortical region of interest for a prospective study. The number of required subjects per group is also affected by the magnitude of the thickness difference the researcher hopes to detect. Surprisingly, many published research papers do not report the magnitude of the detected cortical thickness differences between groups. Presumably this is due to the conventional approach of displaying a map of supra-threshold *P*-values to indicate regions in which statistically significant differences in thickness exist, rather than a map of the effect size. Although the aim of this article is not to comprehensively document reported cortical thickness differences in the literature, some examples include reported cortical thickness differences between 0.1 and 0.6 mm [Lazar et

al., 2005; Sowell et al., 2008; Acosta et al., 2009; Kubota et al., 2010; Tunnard et al., 2011; Wallace et al., 2010; Cerasa et al., 2011]. One would hope that as the field matures the practice of providing information on the magnitude of reported cortical thickness differences between subject groups will become commonplace.

The results of this study indicate that as the applied surface smoothing increases, the number of subjects required per group decreases. The authors recommend against interpreting this finding as an argument in favor of using a large smoothing filter, as the analysis is maximally sensitized to detecting focal abnormalities with the same spatial extent as the filter, by the matched filter theorem. One should therefore tailor the smoothing filter size to the expected extent of the cortical abnormality whenever practical. The spatial extent of the smoothing filter is unlikely to perfectly match the hypothetical abnormality. This means that a thickness difference will be averaged with some cortex in which there is no substantive difference in thickness. For this reason, the magnitude of the estimated cortical thickness differences, particularly in an exploratory study, may be an underestimate of the true thickness difference.

Finally, we investigated the effect of a more stringent type I error rate on the number of subjects required per group for a well-powered study. More stringent type I error rates, in the form of a lower *P*-value, are applied to adjust for the increased incidence of false positive findings in mass univariate analyses. Although the application of a large number of statistical tests constitutes a mass univariate approach, there is a certain level of spatial dependence of cortical thickness. For example, sensory cortex is consistently thinner than temporal lobe cortex. The use of smoothing also increases the spatial dependence of vertex-wise estimates. For this reason, traditional methods for threshold adjustment based on multiple tests, such as the Bonferroni method, are overly conservative. Typical methods for adjusting the threshold in cortical thickness analyses are false positive rate threshold adjustment [Genovese et al., 2002], permutation-based methods [Sowell et al., 2004], and cluster-based thresholding. Regardless of the type of threshold adjustment, a more stringent α level is used as a threshold for statistical significance.

The most commonly used method for adjusting the significance threshold in cross-sectional cortical thickness analyses, false discovery rate, requires a distribution of *P*-values. Because we do not carry out vertex-wise comparisons with a patient group in this study, we do not have a *P*-value distribution and so cannot specify any single adjusted *P*-value threshold. Furthermore, the false discovery rate threshold is inversely related to the spatial extent of the hypothetical thickness difference [Genovese et al., 2002]. It is unlikely the researcher will know the spatial extent in advance, and so a reasonable value for a lower α must be inferred from previous studies. It is difficult to get an idea of adjusted thresholds from the literature, as vertex-wise cortical thickness analysis studies often report the threshold as "$P < 0.05$ adjusted for multiple comparisons" with the adjustment method of choice, without reporting the actual adjusted *P*-value threshold. However some example adjusted thresholds from the literature are $P = 0.005$ [Kubota et al., 2010] and $P = 0.00035$ [Lazar et al., 2005], suggesting that [0.05, 0.0005] is a reasonable interval for adjusted thresholds. For vertex-wise cross-sectional cortical thickness studies to achieve a level of power equivalent to a well-powered "single" univariate analysis, the researcher would need to include up to 140 subjects per group to detect a difference of 0.25 mm when the type I error rate is set at 0.0005, depending on other factors explored in this study such as smoothing filter and location on the cortical sheet.

Previous cortical thickness studies that provided estimates of sample sizes based on power analyses did not adopt the approach of explicitly mapping the number of required subjects over the cortical sheet and did not provide a method for estimating the number of subjects

per group based on thickness difference, effect size and other image processing parameters [Han et al., 2006]. With regard to the Han et al study, our estimate of 27 subjects per group contrasts with the reported seven subjects per group required to detect a 10% thickness reduction in each voxel. We believe that this difference is primarily due to differences in the definition of standard deviation used for power calculations in the two studies. The Han et al study substituted vertex-wise estimates of the mean value of the absolute differences as a measure of the standard deviation. In our study we used the commonly accepted definition of standard deviation, that is, the square root of the variance, using a denominator of $n - 1$. These gave contrasting estimates of the standard deviation of 0.12 mm [Han et al., 2006] compared with an average vertex-wise across-subject standard deviation of 0.36 mm for age-corrected, 6 mm FWHM smoothed data calculated from the cohort presented in this article. It is probable that the different standard deviation estimates are the reason behind our considerably higher estimates in the number of subjects per group required to detect a thickness difference of 0.2 mm with an $\alpha = 0.05$. The Lerch study reports that a group size of 25 subjects can detect a 0.6 mm thickness change after smoothing with a 30-mm FWHM surface-based diffusion smoothing kernel [Lerch and Evans, 2005]. Analysis of our dataset gives a lower estimate of 14 subjects per group using similar processing parameters. The difference may reflect differences in the cortical thickness mapping and coregistration techniques utilized in the Lerch et al. [2005] study; in particular the Lerch et al. study did not use a surface-based registration method. Our study also had the advantage of a considerably larger cohort (98 subjects compared with 25 subjects in the Lerch study and 15 subjects in the Han study) which may improve vertex-wise estimation of the variance of cortical thickness.

In this study we have limited ourselves to power analyses of cross-sectional comparisons of cortical thickness between two subject groups. Sample size estimates for alternative methods of investigation have not been explored, such as correlating cortical thickness changes against a continuous variable, or longitudinal studies in which thickness changes are tracked across multiple time points in a group of individuals. Additional metrics that are derived from the cortical modeling procedure carried out by Freesurfer, such as volume and curvature measures, may also benefit from power analyses. In pathologies where more than one measure is affected, it would be useful to determine which parameter has the most statistical power for detecting changes in the patient group. Future research will focus on applying methods for sample size estimation to these modes of analysis.

In summary, we have provided a comprehensive investigation of how many subjects should be included to ensure a well-powered cross-sectional cortical thickness study. We have investigated the spatial variability of the number of subjects per group over the cortical surface, and provided estimates of this number on a per-lobe level. Finally, we have provided a simple equation that provides estimates of sample size as a function of standard study specific parameters. The outcomes of this study will allow researchers planning prospective studies to recruit and image appropriately sized cohorts to detect regional cortical thickness differences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract grant sponsor: National Institutes of Health; contract grant number: NIH-NINDS R37-31146; Contract grant sponsor: Victorian Life Sciences Computation Initiative (VLSCI; Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government); Contract grant number: VR0056; Contract grant sponsor: Victorian Government's Operational Infrastructure Support Program; Contract grant sponsor: Scobie nd McKinnon

Trust; Contract grant sponsor: NHMRC program; Contract grant number: 628952; Contract grant sponsor: Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health); Contract grant number: U01 AG024904.

References

Acosta O, Bourgeat P, Zuluaga MA, Fripp J, Salvado O, Ourselin S. Automated voxel-based 3D cortical thickness measurement in a combined Lagrangian-Eulerian PDE approach using partial volume maps. Med Image Anal. 2009; 13:730–743. [PubMed: 19648050]

Bernasconi A, Antel SB, Collins DL, Bernasconi N, Olivier A, Dubeau F, Pike GB, Andermann F, Arnold DL. Texture analysis and morphological processing of magnetic resonance imaging assist detection of focal cortical dysplasia in extra-temporal partial epilepsy. Ann Neurol. 2001; 49:770–775. [PubMed: 11409429]

Cerasa A, Quattrone A, Gioia MC, Tarantino P, Annesi G, Assogna F, Caltagirone C, De Luca V, Spalletta G. Dysbindin C-A-T haplotype is associated with thicker medial orbitofrontal cortex in healthy population. Neuroimage. 2011; 55:508–513. [PubMed: 21184829]

Cohen, J. Statistical Power Analysis for the Behavioural Sciences. New Jersey: Lawrence Erlbaum Associates; 1988.

Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci USA. 2000; 97:11050–11055. [PubMed: 10984517]

Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage. 2002; 15:870–878. [PubMed: 11906227]

Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R, Maguire P, Rosas D, Makris N, Dale A, Dickerson B, Fischl B. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. Neuroimage. 2006; 32:180–194. [PubMed: 16651008]

Hardan AY, Muddasani S, Vemulapalli M, Keshavan MS, Minshew NJ. An MRI study of increased cortical thickness in autism. Am J Psychiatry. 2006; 163:1290–1292. [PubMed: 16816240]

Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging. 2008; 27:685–691. [PubMed: 18302232]

Kubota M, Miyata J, Yoshida H, Hirao K, Fujiwara H, Kawada R, Fujimoto S, Tanaka Y, Sasamoto A, Sawamoto N, Fukuyama H, Murai T. Age-related cortical thinning in schizophrenia. Schizophr Res. 2010; 125:21–29. [PubMed: 21036016]

Kuperberg GR, Broome MR, McGuire PK, David AS, Eddy M, Ozawa F, Goff D, West WC, Williams SC, van der Kouwe AJ, Salat DH, Dale AM, Fischl B. Regionally localized thinning of the cerebral cortex in schizophrenia. Arch Gen Psychiatry. 2003; 60:878–888. [PubMed: 12963669]

Lazar SW, Kerr CE, Wasserman RH, Gray JR, Greve DN, Tread-way MT, McGarvey M, Quinn BT, Dusek JA, Benson H, et al. Meditation experience is associated with increased cortical thickness. Neuroreport. 2005; 16:1893–1893. [PubMed: 16272874]

Lerch JP, Evans AC. Cortical thickness analysis examined through power analysis and a population simulation. Neuro-image. 2005; 24:163–173. [PubMed: 15588607]

Lerch JP, Pruessner JC, Zijdenbos A, Hampel H, Teipel SJ, Evans AC. Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. Cereb Cortex. 2005; 15:995–1001. [PubMed: 15537673]

Rosas HD, Liu AK, Hersch S, Glessner M, Ferrante RJ, Salat DH, van der Kouwe A, Jenkins BG, Dale AM, Fischl B. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology. 2002; 58:695–701. [PubMed: 11889230]

Salat DH, Buckner RL, Snyder AZ, Greve DN, Desikan RS, Busa E, Morris JC, Dale AM, Fischl B. Thinning of the cerebral cortex in aging. Cereb Cortex. 2004; 14:721–730. [PubMed: 15054051]

Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science. 2009; 324:81–81. [PubMed: 19342586]

Shaw P, Greenstein D, Lerch J, Clasen L, Lenroot R, Gogtay N, Evans A, Rapoport J, Giedd J. Intellectual ability and cortical development in children and adolescents. Nature. 2006; 440:676–679. [PubMed: 16572172]

Sowell ER, Kan E, Yoshii J, Thompson PM, Bansal R, Xu D, Toga AW, Peterson BS. Thinning of sensorimotor cortices in children with Tourette syndrome. Nat Neurosci. 2008; 11:637–639. [PubMed: 18488025]

Sowell ER, Thompson PM, Leonard CM, Welcome SE, Kan E, Toga AW. Longitudinal mapping of cortical thickness and brain growth in normal children. J Neurosci. 2004; 24:8223–8231. [PubMed: 15385605]

Tunnard C, Whitehead D, Hurt C, Wahlund L, Mecocci P, Tsolaki M, Vellas B, Spenger C, Kloszewska I, Soininen H, Lovestone S, Simmons A. Apathy and cortical atrophy in Alzheimer's disease. Int J Geriatr Psychiatry. 2010; 26:741–748. [PubMed: 20872914]

Wallace GL, Dankner N, Kenworthy L, Giedd JN, Martin A. Age-related temporal and parietal cortical thinning in autism spectrum disorders. Brain. 2010; 133(Pt 12):3745–3754. [PubMed: 20926367]

Wilkinson L. Statistical methods in psychology journals: Guidelines and explanations. American psychologist. 1999; 54:594–594.

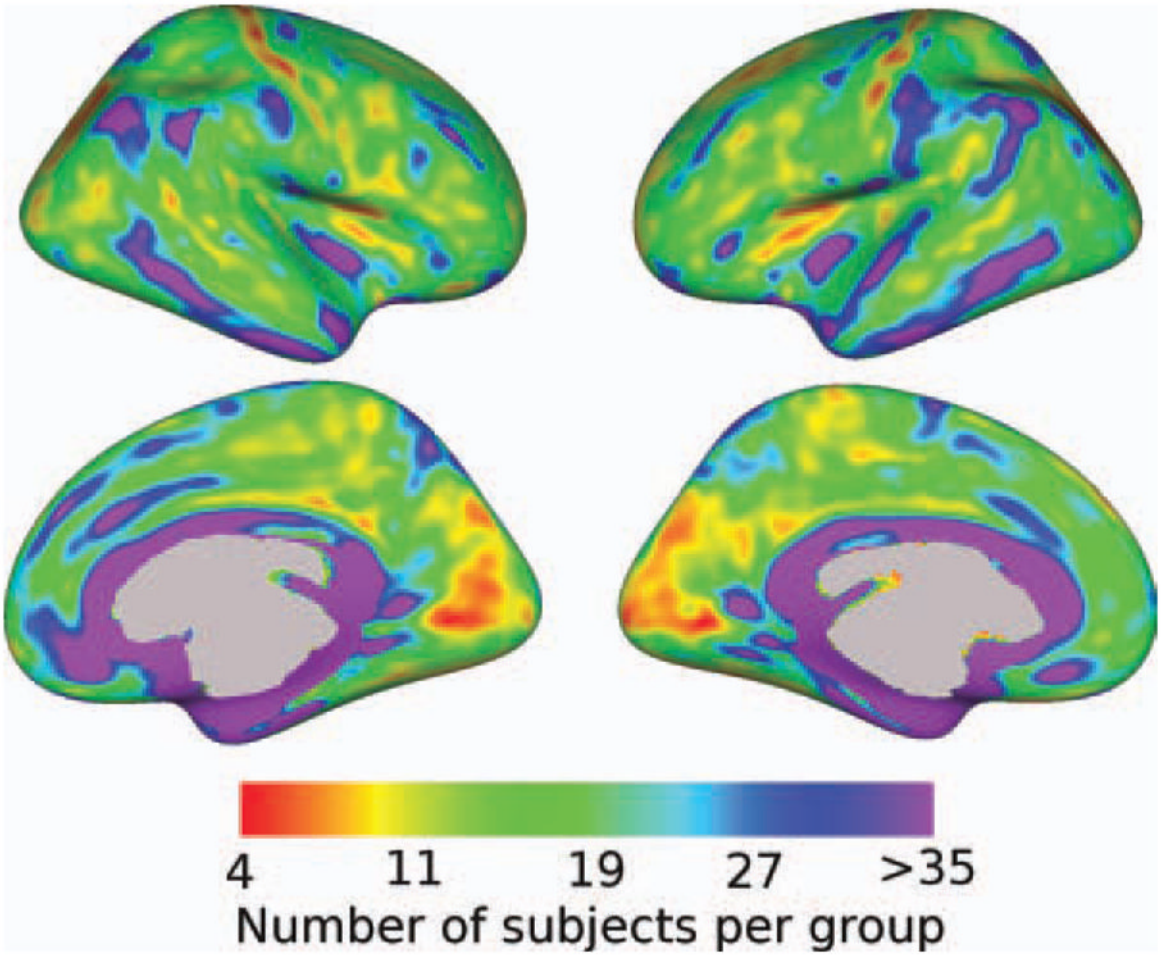

Figure 1.
Inflated surface view of the number of subjects required per group to detect a thickness difference of 0.25 mm. Lateral view in the top row, medial view on the bottom row. Standard deviation was estimated from 98 neurologically normal controls. 10 mm FWHM surface smoothing, power = 0.8, type I error rate = 0.05, two-sided analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

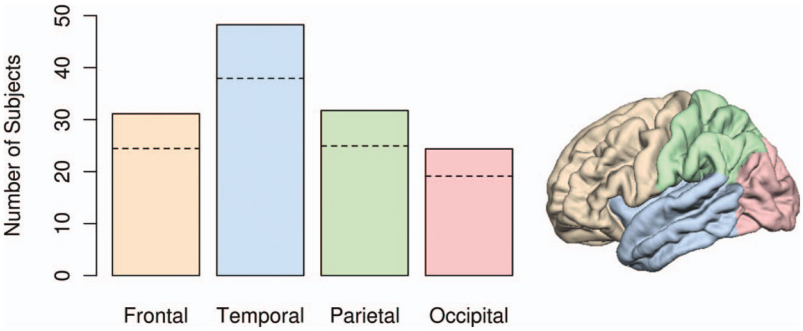

Figure 2.
Number of subjects required per group to detect a change of 0.25 mm over 95% of each major lobe after 10 mm FWHM surface smoothing. Regions were derived from the PALS-B12 lobar atlas. The limbic lobe required 234 subjects per group and was omitted from the plot. Power = 0.8, type I error rate = 0.05, two-sided analysis. The dashed horizontal lines indicate the equivalent one-sided analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

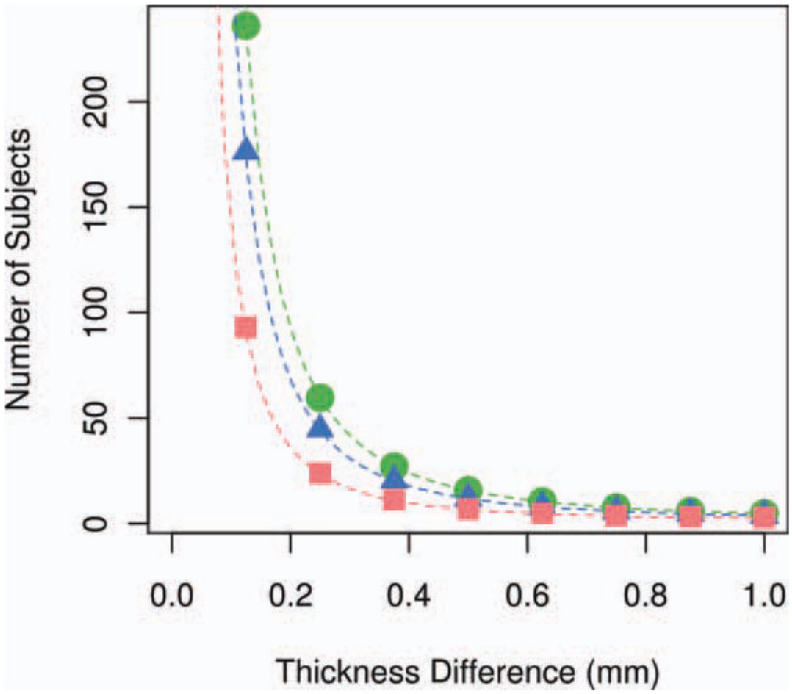

Figure 3.
Number of subjects required to detect given effect size over 95% of the entire cortical surface (green circles), temporal lobe (blue triangles), and occipital lobe (pink squares). The dashed lines show the predicted number of subjects per group according to Eq. (1). 10 mm surface-based smoothing, power = 0.8, type I error rate = 0.05, two-sided analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

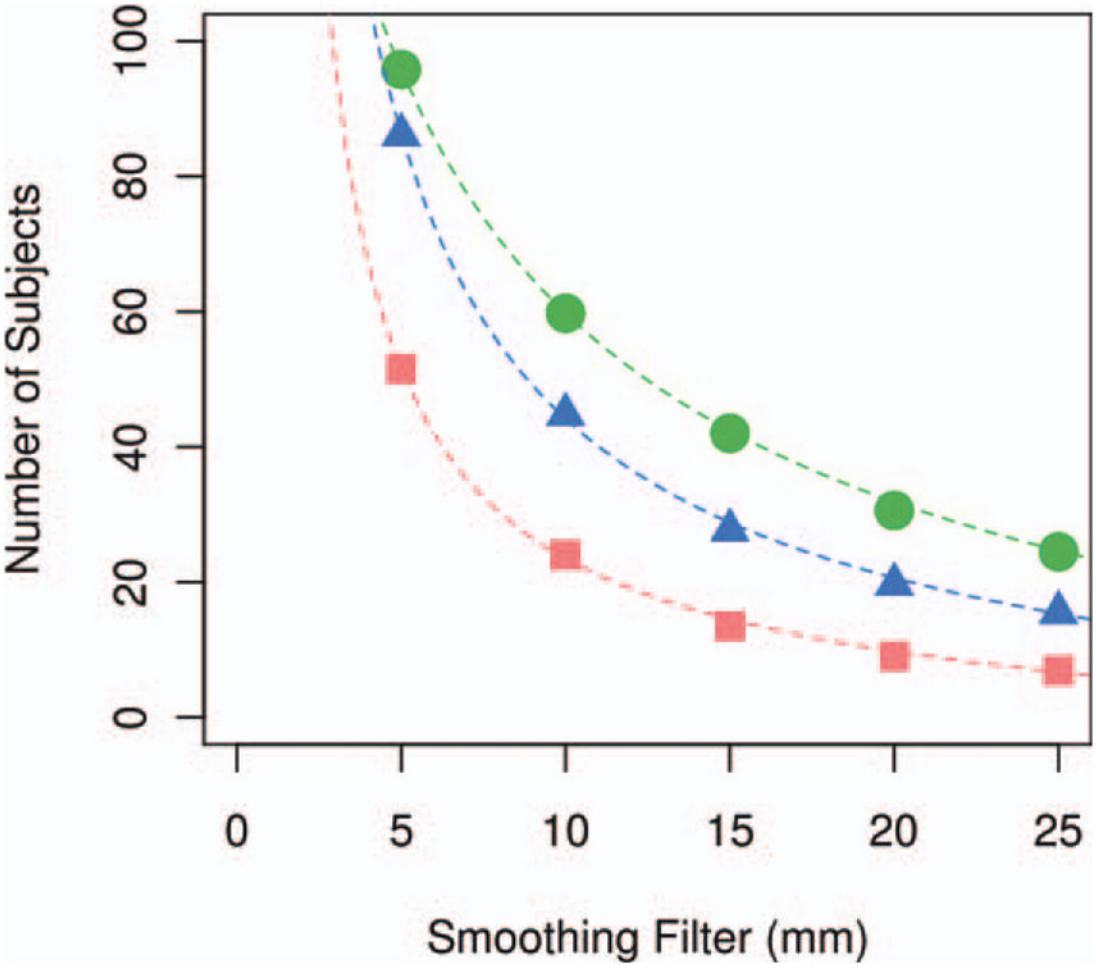

Figure 4.

The effect of applied smoothing filter on the number of subjects required to detect a vertex-wise thickness difference of 0.25 mm over 95% of the whole cortex (green circles), temporal lobes (blue triangles), and occipital lobes (pink squares). The analysis is maximally sensitized to detecting focal abnormalities with the same spatial extent as the smoothing filter; the use of a more extensive smoothing filter does not necessarily increase the power of the study. The dashed lines show the predicted number of subjects per group according to Eq. (1). Power = 0.8, type I error rate = 0.05, two-sided analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

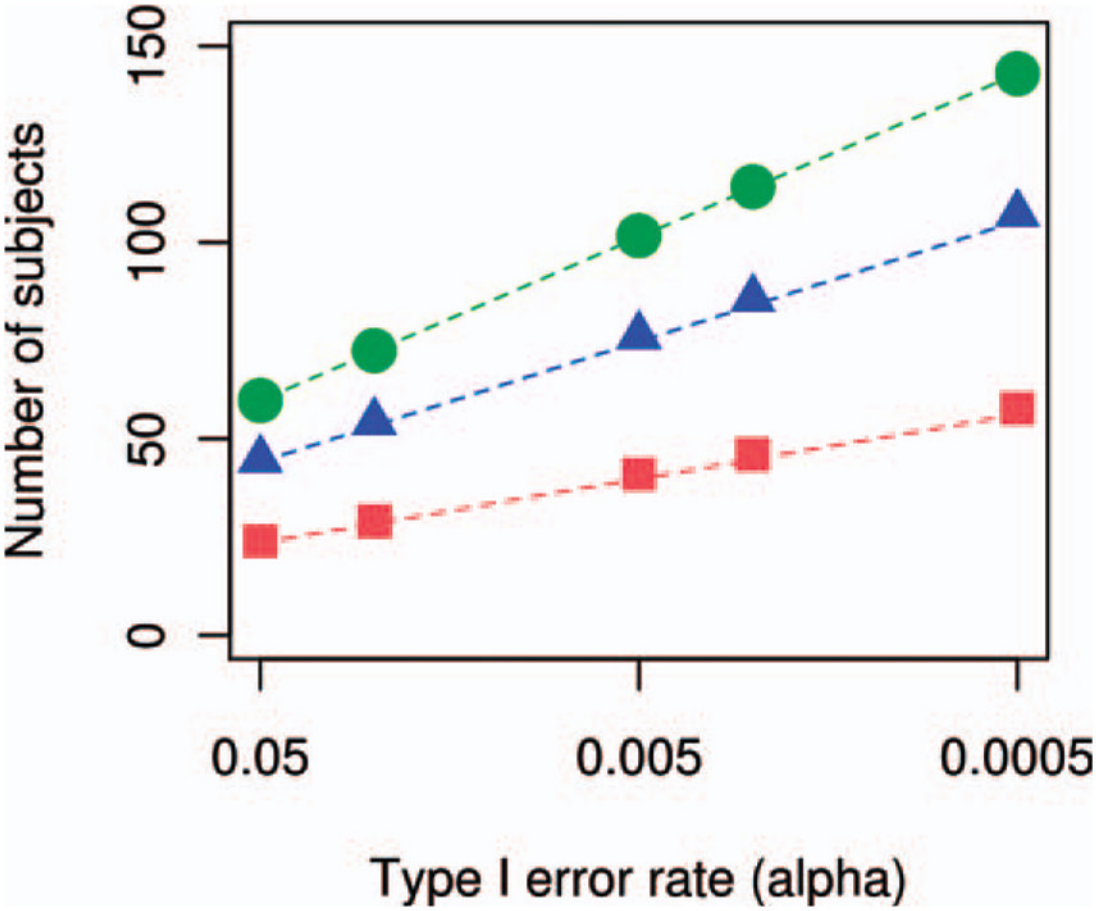

Figure 5.
The effect of more stringent significance threshold (type I error rate α) on the number of subjects required to detect a thickness difference of 0.25 mm over 95% of the cortical surface (green circles), temporal lobe (blue triangles), and occipital lobes (pink squares). The dashed lines represent subject numbers as predicted by Eq. (1). 10 mm FWHM surface smoothing, power = 0.8, two-sided analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE I

Parameter estimates to calculate the number of subjects required per group over the whole cortex and each of the four major lobes, according to Eq. (1)

Cortical region	k_1	k_2	k_3	k_4	k_5	k_6	Mean absolute error
Whole cortex	1.71×10^{-02}	3.71	0.37	4.80	15.81	10.75	0.46
Frontal lobes	-1.94×10^{-03}	−1.10	0.37	1.53	4.84	3.32	0.45
Temporal lobes	1.41×10^{-03}	0.41	0.37	2.76	8.78	5.92	0.78
Parietal lobes	-1.78×10^{-04}	−0.05	0.37	1.82	5.92	3.94	0.68
Occipital lobes	-1.92×10^{-03}	−0.65	0.37	1.28	4.21	2.81	0.69

The final column indicates that the derived equation fits the calculated number of subjects per group with a mean error of less than one subject for each brain region, over all the parameters explored in this study.

TABLE II

Validation of the derived model using MRI data acquired from different sites and age ranges of subjects

Cohort	Default parameters (% mean error ± SD)	Cohort specific parameters (% mean error ± SD)
Normal-bl-3.0T	2.16 ± 1.48	1.52 ± 1.38
Normal-m12-3.0T	1.96 ± 1.59	1.72 ± 1.52
Normal-m06-1.5T	4.31 ± 1.7	1.35 ± 1.28
Normal-m24-1.5T	6.28 ± 1.85	1.51 ± 1.29

Mean absolute percentage errors suggest that the derived model is appropriate for similar quality MRI data acquired at different sites. The second column uses the values of k_1–k_6 from the original model Eq. (1). The third column uses new values of k_1–k_6 estimated by fitting Eq. (1) to cortical thickness measurements from each of the four ADNI cohorts.