



ORIGINAL ARTICLE

Interdatabase Variability in Cortical Thickness Measurements

M. Ethan MacDonald ^{1,2,3,4}, Rebecca J. Williams^{1,2,3,4}, Nils D. Forkert^{1,2,3,4}, Avery J. L. Berman^{1,2,3,4,6}, Cheryl R. McCreary^{1,2,3,4,5}, Richard Frayne^{1,2,4,5} and G. Bruce Pike^{1,2,3,4,6}

¹Departments of Radiology, University of Calgary, Calgary, Alberta, Canada T2N 4N1, ²Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada T2N 4N1, ³Healthy Brain Aging Lab, University of Calgary, Calgary, Alberta, Canada T2N 4N1, ⁴Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada T2N 4N1, ⁵Seaman Family Magnetic Resonance Research Centre, Foothills Medical Centre, Alberta Health Services, Calgary, Alberta, Canada T2N 4N1 and ⁶Department of Biomedical Engineering, McGill University, Montreal, Quebec, Canada

Address correspondence to Matthew Ethan MacDonald, Healthy Brain Aging Lab, Room 2910F, Health Science Building, University of Calgary, Foothills Campus, 3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1. Email: memacdon@ucalgary.ca  orcid.org/0000-0001-5421-3536

Abstract

The phenomenon of cortical thinning with age has been well established; however, the measured rate of change varies between studies. The source of this variation could be image acquisition techniques including hardware and vendor specific differences. Databases are often consolidated to increase the number of subjects but underlying differences between these datasets could have undesired effects. We explore differences in cerebral cortex thinning between 4 databases, totaling 1382 subjects. We investigate several aspects of these databases, including: 1) differences between databases of cortical thinning rates versus age, 2) correlation of cortical thinning rates between regions for each database, and 3) regression bootstrapping to determine the effect of the number of subjects included. We also examined the effect of different databases on age prediction modeling. Cortical thinning rates were significantly different between databases in all 68 parcellated regions (ANCOVA, $P < 0.001$). Subtle differences were observed in correlation matrices and bootstrapping convergence. Age prediction modeling using a leave-one-out cross-validation approach showed varying prediction performance ($0.64 < R^2 < 0.82$) between databases. When a database was used to calibrate the model and then applied to another database, prediction performance consistently decreased. We conclude that there are indeed differences in the measured cortical thinning rates between these large-scale databases.

Key words: aging, cerebral cortical thickness, magnetic resonance imaging

Introduction

Cerebral cortex thickness is known to change with age, becoming thinner during the aging process. This effect has been shown in single database studies and reliably reproduced (Salat

et al. 2004; Desikan et al. 2006; Marcus et al. 2007; Ericsson et al. 2008; Chen, He, et al. 2011; Gong et al. 2012; Lemaitre et al. 2012). Establishing normal healthy rates of cortical thinning is important to understanding aging physiology. Increased

thinning is also associated with conditions such as Alzheimer's disease (Li et al. 2012; Gaser et al. 2013). Although not specific to a particular disease, divergence from normal healthy cortical thickness may indicate which subjects will go on to develop cognitive impairment (Gaser et al. 2013). Cortical thickness measurements have been used recently for predicting chronological age from 3D structural anatomical images of individual subjects (Franke et al. 2010, 2012, 2013, 2015; Gaser et al. 2013; Valizadeh et al. 2017). Methods used for age prediction vary but all techniques calibrate a model from a larger population of subjects and then use an out-of-sample or leave-one-out cross validation strategy to assert that they can predict age. The accuracy of these predictions could be expected to perform best when the out-of-sample test case is from a database where parameters, such as resolution, scanner configuration, and population demographics are similar. Likewise, if the model is trained using a particular database, and applied to another database it may not perform as well if there are underlying differences between the databases.

Measurement of the cerebral cortex thickness is now well established using magnetic resonance (MR) imaging with spatial resolution on the order of 1 mm^3 and good contrast between gray and white matter. MR imaging does not use ionizing radiation and allows for healthy subjects to be imaged without compromising their safety. Structural T_1 -weighted (T_1w) whole brain images are routinely collected as they provide a source image optimized for investigating various brain structures. There are a variety of available anatomical atlases, algorithms, and software packages for segmenting brain structures (Hildebrand and Ruegsegger 1997; Fischl and Dale 2000; Fischl et al. 2004; Desikan et al. 2006; Tustison et al. 2014).

The signal intensity in T_1w images is impacted by a number of factors and is thus not quantitative with respect to T_1 relaxation; however, the morphological (or structural) information is of sufficient quality to make precise measurements of anatomical changes in the brain such as cortical thickness. In practice, there are large differences in the numerical voxel measurements (i.e., voxel values and contrast) from T_1w imaging, which can depend on sources including the individual subject being imaged, the orientation of their head in the magnet, scanner hardware (magnetic field strength, vendor, software version, receiver coil configuration, etc.), image acquisition parameters (including spatial resolution, flip angle and acquisition timing), and other vendor specific hardware and image reconstruction algorithms (Han et al. 2006; Dickerson et al. 2008). Subject specific changes can also subtly affect cortical thickness measurements, such as hydration, menstrual cycle phase, and activity level (Duning et al. 2005; Sowell et al. 2007; Pletzer et al. 2010; Kempton et al. 2011).

Although there are similarities between T_1w images in different image databases, there may also be important differences. It is generally assumed that structural metrics calculated from T_1w images do not vary considerably between different large-scale studies and thus studies are often combined to yield the largest possible multicenter database for analysis. A few studies have tested for differences between vendor and field strength (Han et al. 2006; Dickerson et al. 2008), and in these studies the parameters and subjects were precisely controlled between scans, keeping sequence parameters and subjects the same as field strength varied. These studies concluded that there were small but significant effects of field strength and vendor on cortical thickness, and that sequence parameters could also result in differences. From these studies, it was concluded that magnet configuration

and sequence parameters should be as similar as possible when performing large populations studies. However, the reality is that many parameters beyond the vendor and field strength will be different, such as the head coil and other hardware, subject specific and image acquisition parameters. Thus, the variation accounted for by field strength alone would be misrepresentative. Our current work aims to build on this observation by explicitly investigating the differences between several large databases.

The overall goal of the current study is to characterize and compare the rate of cortical thinning between multiple large image databases. Four lifespan databases are compared, each with >180 subjects. The cerebral cortex thickness was calculated for all the subjects and parcellated. We determined the rate of change in thickness of the cerebral cortex with age, and examined differences between databases with a regression-based analysis. Statistical analyses were performed to test for differences between databases. The variance in each database was characterized with correlation and bootstrapping analysis. Finally, age prediction was performed between each of the different databases to determine the effect of database on chronological age estimation.

Methods

Image Data

T_1w structural anatomical scans from 4 large databases were gathered: 1) the Calgary Normative Study (CNS) Database (Tsang et al. 2017), 2) the Open Access Series of Imaging Studies (OASIS) database (Marcus et al. 2007; OASIS 2017), 3) the Information eXtraction of Images (IXI) database (Ericsson et al. 2008; IXI 2017), and 4) the Dallas Lifespan Brain Study (DLBS) database (DLBS 2017). The first database was collected locally and the other 3 are open source databases and used widely among the research community. Reporting of acquisition parameters varied between databases. T_1w imaging used a 3D magnetization prepared rapid gradient echo (MPRAGE) (Mugler and Brookeman 1990) and had imaging parameters listed as either: inversion time (TI)/repetition time (TR)/echo time (TE)/flip angle (α), or TR/TE/ α , respectively, as reported by the database repository. Briefly:

- The CNS database was collected in Calgary, Alberta (Tsang et al. 2017). All subjects in this study were healthy subjects who tested above 26 on the Montreal Cognitive Assessment (Nasreddine et al. 2005) and self-reported no neurological disease or psychiatric illness. This database consisted of 188 subjects (71 males, 117 females, aged 18–87 years). The data was collected on a 3 T General Electric (GE) scanner and the image resolution was 1 mm isotropic. The acquisition parameters were TI/TR/TE/ α of 650 ms/5.84 ms/2.36 ms/8°.
- The OASIS database was collected in St. Louis, MI (Marcus et al. 2007; OASIS 2017). All subjects included from this study were right handed, had a normal mini-mental state exam scoring (Tombaugh and McIntyre 1992) and were otherwise healthy. A total of 316 subjects were included (119 males, 197 females, aged 18–96 years). Acquisitions were performed on a 1.5 T Siemens scanner. These T_1w images had a 1 mm \times 1 mm \times 1.25 mm resolution. Postprocessing on this data was performed and included facial feature removal to anonymize subjects prior to open access release. The acquisition parameters were TI/TR/TE/ α of 200 ms/9.7 ms/4 ms/10°.
- The IXI database was collected in London, UK (IXI 2017). Images in this database were collected on 3 MR scanners

(Phillips 3 T, Phillips 1.5 T, or GE 1.5 T). A total of 563 subjects were included (250 males, 313 females, aged 20–86 years). The images in this database had a resolution of 0.94 mm × 0.94 mm × 1.2 mm. On the Phillips 3 T the parameters were TR/TE/α 9.6 ms/4.6 ms/8°, on the Phillips 1.5 T the parameters were 9.8 ms/4.6 ms/8°. The acquisition parameters for the GE scanner were not reported.

- The DLBS database was collected in Dallas, TX (DLBS 2017). Images were acquired on a Phillips 3 T. A total of 315 subjects had image data in this database (117 males, 198 females, aged 20–89 years). Subjects had normal mini-mental state exam (>24/30). Image resolution was 1 mm isotropic and the scan parameters TR/TE/α were 8.3 ms/3.7 ms/12°.

A summary of these databases is provided in Table 1.

Image Processing

All T1w images were processed using FreeSurfer (version 5.3.0). The cerebral cortex was segmented and parcellated into 34 regions per hemisphere using the Desikan–Killiany (Desikan et al. 2006) atlas. Processing was performed by running the ReconAll function with the Qcache feature, all processing was performed on a parallel computer cluster (SGI Altix XE 1300 cluster with 316 compute nodes each with two 6-core Intel Xeon X5650 2.66 GHz processors and 48 GB of memory). Briefly, the FreeSurfer pipeline for segmentation and cortical parcellation involves skull stripping (Ségonne et al. 2004), registration to Talairach space, intensity normalization (Sled et al. 1998), white matter segmentation, tessellation of the gray matter/white matter boundary, and automated topology correction (Fischl et al. 2001; Ségonne et al. 2007). The tessellated surface is used to define the white matter and pial boundaries, which is required for the calculation of cortical thickness (Dale and Sereno 1993; Dale et al. 1999; Fischl and Dale 2000). Following surface inflation (Fischl et al. 1999), cortical parcellation with respect to gyral and sulcal structure is performed (Fischl et al. 2004; Desikan et al. 2006). Cortical thickness is calculated as the closest distance from the gray/white boundary to the gray-cerebral spinal fluid boundary at each vertex on the tessellated white matter surface (Fischl and Dale 2000). Average thickness was calculated for each parcellated region and for each hemisphere, resulting in 70 measurements per subject.

Quality assurance of the FreeSurfer processing was performed by visually inspecting all the segmentations and excluding subjects with clear errors. After this, a statistical analysis of the histograms, scatterplots, and correlations of each region for each database was used to identify potential outliers. A total of 14 subjects were excluded from the analysis: 3 from the CNS database, 4 from the OASIS database, 3 from the IXI database, and 4 from the DLBS.

Table 1 Summary of ages and magnet type each database

Database	Age range (years)	Number of participants	Scanner
CNS	18–87	188 (71 M)	3 T GE
OASIS	18–96	316 (119 M)	1.5 T Siemens
IXI	20–86	563 (250 M)	1.5 T and 3 T Phillips, 1.5 T GE
DLBS	20–89	315 (117 M)	3 T Phillips

Analysis

Cortical surface maps were calculated using the general linear model (GLM) utility in FreeSurfer with 2 regressors: database as a factor (CNS, OASIS, IXI, DLBS) and age as a covariate. All subjects were first registered to a standard surface. A surface map of the log P-value (from the GLM) was rendered. The GLM analysis allowed for both positive and negative scale representing where the databases were similar and dissimilar, respectively.

The cortical thickness measurements were then exported to Matlab (R2016b) for the following statistical analyses. Thickness measurements from the 70 parcellated regions were regressed versus age using linear regression. Laterality index, $(L - R)/(L + R)$, was computed for each and region and database. Parameters associated with each regression (the slope, the intercept and the coefficient of determination (R^2)) were plotted on a multivariable polar plot. Pearson correlation was used to determine the significance of the linear relationship between cortical thickness and age for each region in each database, this was not done to compare between databases. Assumptions required for analysis of covariance (ANCOVA) tests were then verified, which included transforming the cortical thickness data to residuals by removing the regression line component, then Jarque–Bera tests for normality were performed on each region in each database, and Levene’s tests for homogeneity of variance between databases were performed for each region. When testing ANCOVA assumptions, thresholds of $P < 0.05$ and $P < 0.01$ were counted.

An independent-samples one-way ANCOVA with the 4-level factor of database (CNS, OASIS, IXI, DLBS) and age as the covariate was performed on each region separately. To test for the differences in both the intercept and the slope a separate lines model was used for the ANCOVA instead of the more conventional parallel lines model. The ANCOVA tests examined cortical thickness differences between databases. A threshold of $P < 0.001$ was considered significant for the ANCOVA (a conservative threshold was selected to compensate for the large number of subjects and tests). Due to the high number of comparisons, a Bonferroni multiple comparisons corrected P value was also considered ($P < 0.001/70 = 1.43(10^{-5})$).

The variance of the data was characterized with correlation and bootstrapping. Correlation matrices were used to evaluate relationships between regions within each database (Chen, He, et al. 2011) and for the consolidated database. Box’s M test was used to test for differences between covariance matrices between the databases. A bootstrap analysis was performed to determine the number of subjects per decade of life (20–29, 30–39, etc.) required to establish consistent correlations. Using 10 000 permutations of the subject order, the number of subjects included per decade was increased from 3 to 20. If there were fewer than twenty subjects per decade in a database, then no new subjects were included for that decade when the maximum was reached. The standard deviation of the calculated slope, intercept and R^2 were calculated from the set of permutations. The standard deviation of each parameter was plotted against the number of subjects per decade. The standard deviation converged to a given tolerance (95% of the minimum) of each regression coefficient, which indicated the deviation of that parameter with resampling. This can be considered the intradatabase variation of the regression coefficient. For the consolidated database, the number of subjects included was computed from 3 to 30 subjects, since more were available.

Age estimation using a multiple regression prediction model was performed and a calibration was calculated using each database. A leave-one-out validation strategy was performed

on each database and for the consolidated database. A model calibrated with each database was applied to the other databases in a full out-of-sample validation strategy. The R^2 and mean absolute difference (MAD) in age were used as a performance metrics to determine the quality of prediction.

Results

Histograms of the age distribution and gender are shown for each database and for the consolidated database in Figure 1.

Figure 2 shows a statistical map of the brain surface generated from the FreeSurfer GLM utility, in which we tested for cortical thickness differences between databases. With this

analysis, local regions of larger and smaller differences are revealed. There is lower P -values of the temporal and occipital lobes and lower in the frontal and parietal lobes. Similar regressions are seen between the right and left hemispheres.

For the 68 parcellated regional measurements and 2 cortical thickness measurements averaged for each hemispheres (70 measurements total), the Pearson correlation tests of cortical thickness versus age were significant in all regions and databases except: bilateral temporal pole, entorhinal, left (L)-medial orbitofrontal, L-cuneus, L-caudal anterior cingulate, right (R)-insula, R-rostral anterior cingulate, R-pericalcarine, and R-inferior temporal. The Jarque-Bera tests of normality indicated that the residual data was mostly normally distributed

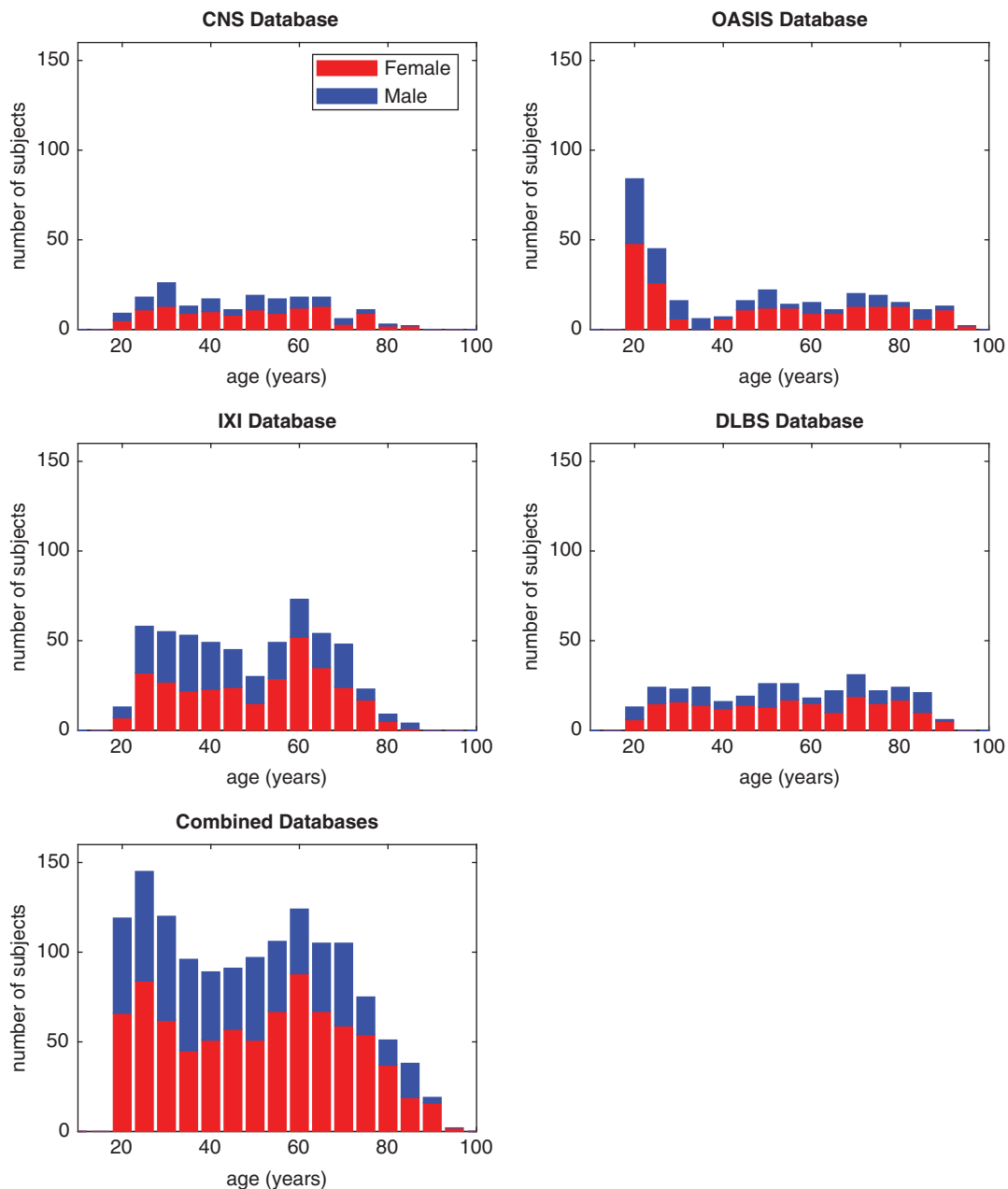


Figure 1. Distribution of age and sex among the 4 databases. Sample sizes per decade are mostly uniform over the lifespan, with 2 primary exceptions: 1) the OASIS database, which has more younger subjects and 2) all the database tend to have fewer subjects over 75. The impact of the latter exception is more pronounced when looking at the combined database histogram. In general, databases included more female subjects (approximately 2/3 female and 1/3 male participants).

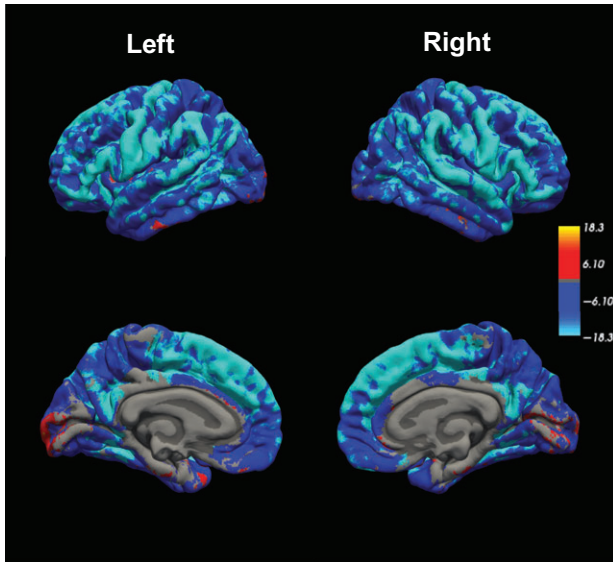


Figure 2. Surface statistical maps from the GLM analysis. These maps show the log(P-value) found when testing for differences using a general linear model with database as a factor and age as a covariate. Almost all regions of the cortex indicated statistically significant differences between databases, consistent with the ANCOVA testing. These images show regions of greater statistical significance (dark blue), and regions of lower statistical significance (teal). A minority of the surface showed regions where the databases were similar (red), or do not show statistical significance between databases (no color overlay).

(19 of 280 were significant at $P < 0.01$ and 8 of 280 tests were significant at $P < 0.05$). The Levene's test for homogeneity of variance between databases were also statistically significant for some regions, indicate that for these regions, variances were not equal (13 of 70 regions were significant at $P < 0.01$ and 6 of 70 at $P < 0.05$). However, this was not a concern as such assumption violations affect the Type I error rate and thus the possibility of false-positive results. Here, our use of large sample sizes and conservative Bonferroni multiple comparisons correction ($P < 0.001/70 = 1.43(10^{-5})$) ensured that the statistical tests were robust to minor violations. The ANCOVA test results were consistent with the surface map.

Figure 3 shows a polar plot of the correlation parameters (slope, intercept, and R^2) for each region and database, and for the consolidated database. A left-right symmetry was observed, as well as differences in the regressions between databases. Figure 4 shows the laterality indices. Example linear regressions for the R-banks superior temporal sulcus and the R-parahippocampal gyrus, are shown in Figure 5. These regions highlight instances where the slopes are similar and dissimilar, respectively. In both cases the ANCOVA test showed statistical significance.

The correlation matrices are shown in Figure 6. These matrices reveal how related each cortical region's thickness is to the other regions' thickness. The principal observation from this figure is the consistency of regions of higher correlation between databases indicated by the red circles in Figure 6. However, there are bands (red arrows) in which the correlation is lower in the Calgary Normative and OASIS databases. Box's M test indicated that there was a difference between all covariance matrices ($P < 0.001$). In Figure 7, the plots from the bootstrapping are displayed. The trends are similar in each of the plots and the standard deviation of the regression coefficients are reduced as the number of subjects included is increased.

The average number of subjects per decade required to reach 30% of the standard deviation with 20 subjects of the slopes was 15.1, 12.1, 11.4, and 12.3, for the CNS, OASIS, IXI, and DLBS databases respectively. The other parameters followed this trend. The convergence of the IXI database was faster compared with the other, which implies fewer outliers in the IXI database. The subtle deviation in the cross-correlation matrix and bootstrapping indicates similar qualities in the variance of each database.

Table 2 shows a summary of the age predictions from the multiple regression model. The R^2 values in the table range between 0.647 and 0.818, and the MAD values ranged from 8.74 to 14.00 years. Of particular interest is the lower R^2 of the leave-one-out cross validation for the CNS database, 0.647, this can likely be explained by the fewer subjects available for calibration. When inspecting Table 1, it can be noted that the R^2 is generally lower when an alternative database is used for calibration rather than the leave-one-out cross validation. The MAD shows a similar trend whereby the best predictions are achieved when the calibration is performed with the leave-one-out validation rather than applying a model calibrated from another database (a.k.a., full out-of-sample validation). The MAD metric appears to be a better indicator of prediction accuracy than R^2 . To illustrate this point, the OASIS database, which had weaker correlations (Fig. 3), had higher R^2 suggesting that it was well predicted, although the MAD indicated that it was the poorest intrapredictive database with leave-one-out cross-correlation. In all cases, MAD increased when an alternative database was used for calibration. When consolidating databases, a reasonable predictive power remained; this is similar to other recent reported findings of R^2 using many databases and multiple linear regression modeling (Valizadeh et al. 2017).

Discussion

The main finding of this study is that there are significant differences in the estimated cortical thinning rates between these large databases. A total of 1382 subjects were included from 4 databases, which when combined increased statistical power, but showed small yet statistically significant differences between databases. The fact that the differences are mostly small and that correlation matrix and bootstrapping results are similar, despite considerable acquisition protocol differences, gives a high level of confidence to the overall cortical thinning trends. The differences are, however, larger than the asymptotic standard deviation observed in the bootstrapping plots, indicating that the differences are practical and important.

There are several assumptions required prior to performing ANCOVA tests, including linearity with the covariate, and normality and homogeneity of the residuals. Tests of linearity, normality, and homogeneity of variance were performed, and the assumptions were satisfied in the majority of tests. Often a parallel lines model is used, but in this study a model allowing for different slopes was used and, thus, consistency of slopes was not tested for. Post hoc tests are also often done after finding statistical significance in a multifactorial analysis, but this would result in an additional 840 statistical tests. The findings of these post hoc tests would determine regional cortical thickness differences between individual databases, however, this was not the aim of the current study.

Inclusion of sex as a cofactor when performing the statistical tests was also incorporated (results are not shown). Specifically, a multivariate ANCOVA (MANCOVA) test similar to

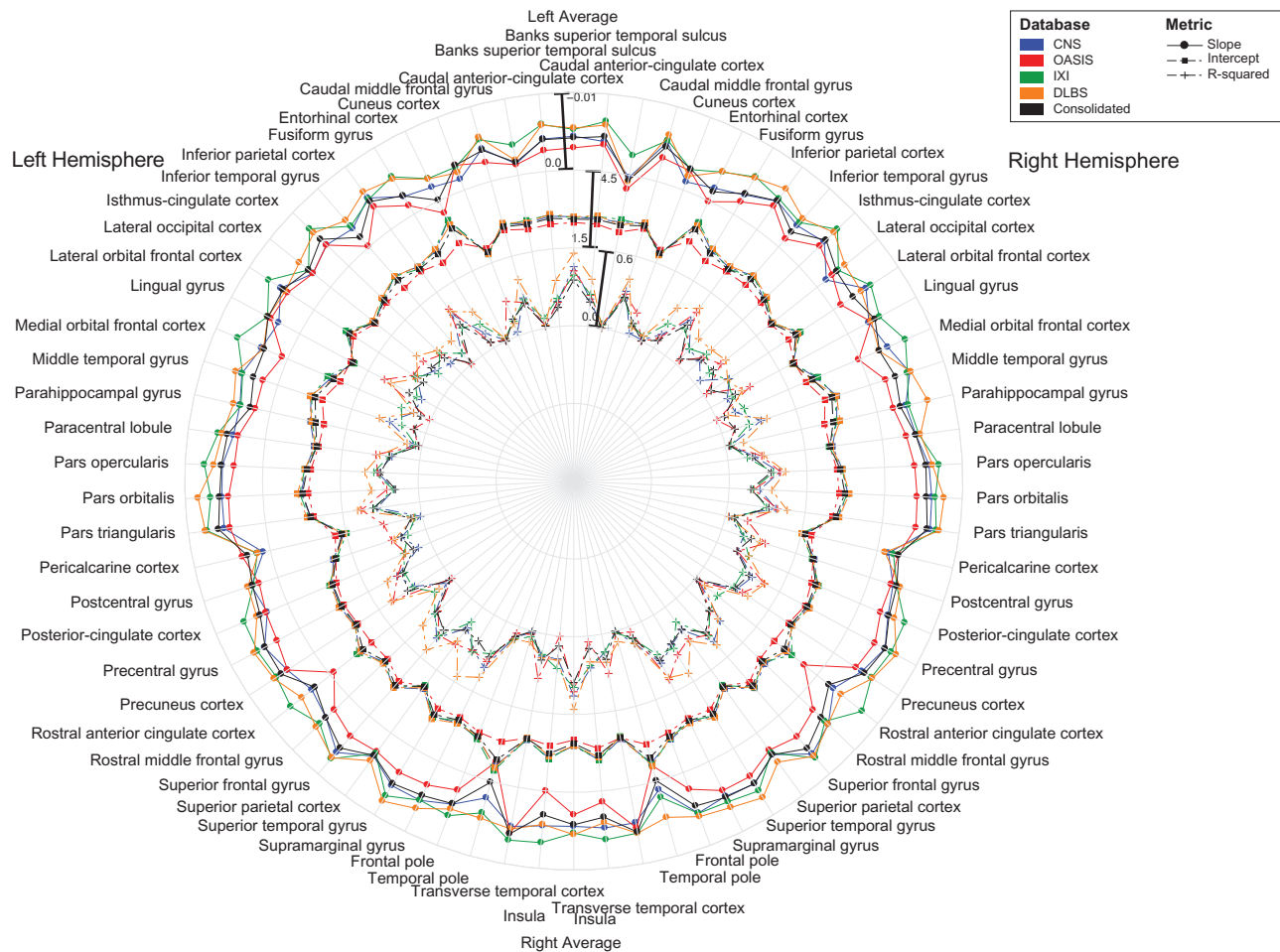


Figure 3. Polar plot with correlation metrics for each region. This polar plot illustrates the slope, intercept and R^2 for each database in each region. The outer ring plots the slope, the middle ring plots the intercept and the innermost ring plots the R^2 . Databases are separated by color and markers separate the metrics (see legend). The ranges of values are indicated by the scale bars on each of the 3 rings; the range of slopes is from 0.0 mm/year to -0.01 mm/year, the range of intercepts is from 1.5 to 4.5 mm, and the range of R^2 is from 0.0 to 0.6.

the ANCOVA test was also performed for each region including sex as a cofactor. Since the databases have similar sex ratios, it did not impact the results and all the regions had the same level of significance, so only the ANCOVA results are reported.

The present findings are in agreement with another study where consistency between databases of similar regions was found, including the superior frontal gyrus, the middle frontal gyrus, and the parssotriangularis (Fjell et al. 2009). This study also investigated the inconsistencies of cross sectional cortical thinning between databases, although their databases were smaller and the analysis and comparison was different, there were regional similarities of consistency and inconsistency between databases.

Many studies compare a disease group with a healthy group to determine if there are abnormalities of cortical thickness. When a statistically significant difference observed between these 2 groups is reported, it is possible this reported difference is similar to that found between 2 healthy groups. Thus, reporting increased cortical thinning with that disease state might be misrepresentative. To avoid this, it is important to ensure that the acquisition is the same between the healthy and diseased groups and perform bootstrapping to assert that the difference is greater than that found when resampling the individual groups.

There are a number of software packages available for calculating cortical thickness. There may be some differences in the cortical thickness measurements depending on the software used (Hildebrand and Rügsegger 1997; Lerch and Evans 2005; Tustison et al. 2014). In this work, we have chosen to use only FreeSurfer, because we believe this software tool is robust and is also widely accepted. Metrics other than cortical thickness could also be utilized, such as volume or curvature. As our processing methodology was consistent and quality control was performed, we can conclude that deviations found between these databases did not arise from the processing per se and must be a result of the acquisition (e.g., field strength, voxel size) or the studied populations (4 different groups with different age and ethnic distributions). Regions of known susceptibility may also vary more between databases.

Quality assurance of the segmentations and cortical thickness measurements are difficult when dealing with a large cohort. FreeSurfer allows for the adjustment of segmentations by insertion of control points on a subject-by-subject basis. For example, if each region was inspected visually for 5 min, the total time to inspect all segmentations would take over 7800 h (1382 subjects \times 68 regions per region \times 5 min per region). This time does not include the time required to make modifications and reprocess, and then recheck the subjects. It would also

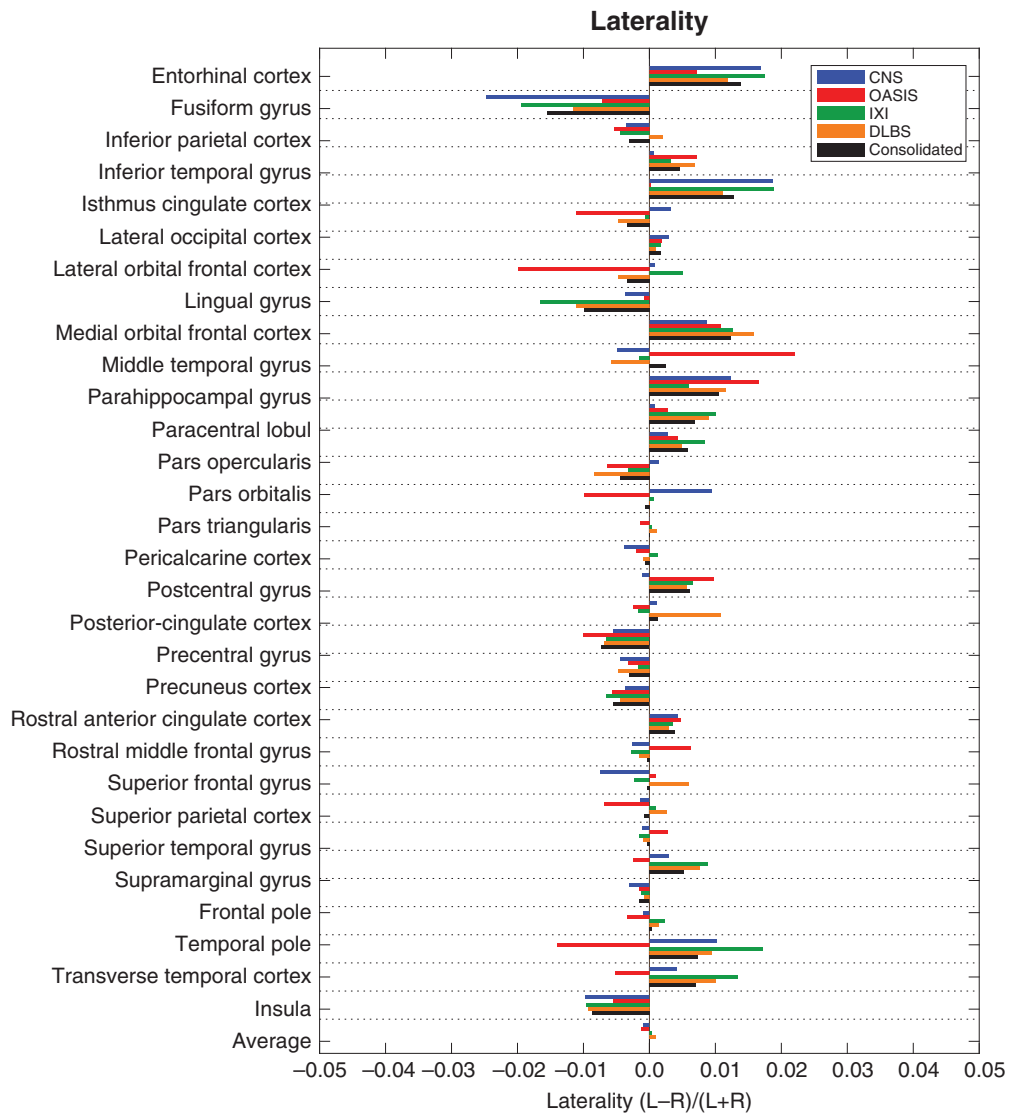


Figure 4. Laterality index computed for each region and each database.

introduce a subjective evaluation into the study that would be difficult for other groups to reproduce this study. Rejecting quality assurance of the segmentations entirely is also negligent, as real errors could exist and skew the results. As described in the methods section, we applied 2 quality assurance approaches: 1) all subjects were visually inspected for errors and 2) all measurements were statistically checked for outliers. We removed a total of 14 subjects from the analysis, consistent with error rates from other healthy normal databases (Smith et al. 2015; Potvin et al. 2016, 2017a,b, 2017a,b). This small amount of the rejected data had no influence on the findings of this study, as inclusion or exclusion did not impact the findings. Overall, the FreeSurfer algorithm provided robust cortical thickness measurements.

The OASIS database was the oldest of the databases used and had a number of deviations with respect to the other databases. In particular, the resolution was slightly lower ($1\text{ mm} \times 1\text{ mm} \times 1.25\text{ mm}$ vs. $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$) and there were 2–4 T1w image volumes for each subject and processed images calculated from the multiple repeated images to yield effective

higher SNR images and skull stripped images. To minimize difference between databases, only the first of the T1w volumetric source images were used in our analysis of the OASIS data.

There are many factors that are known to affect cortical brain thickness, such as sex, hydration, menstrual cycle phase, and handedness (Duning et al. 2005; Sowell et al. 2007; Pletzer et al. 2010; Kempton et al. 2011). Sex distribution was similar between databases (Fig. 1). These parameters, however, were not considered in this analysis since this information was not always reported in the databases, potentially limiting the scope for comparison. Testing for differences in resolution and field strength, flip angle, SNR, scanner manufacturer, or neuro-psych test were also not performed, as we cannot control for those factors independently of database; too many parameters were changing simultaneously between databases to ascribe those to individual sources to the variation. There are several studies that attempt to determine the effect of these parameters when combining large numbers of databases (Potvin et al. 2016, 2017a,b, 2017a,b), but omit testing for changes in database, hence motivating the current study.

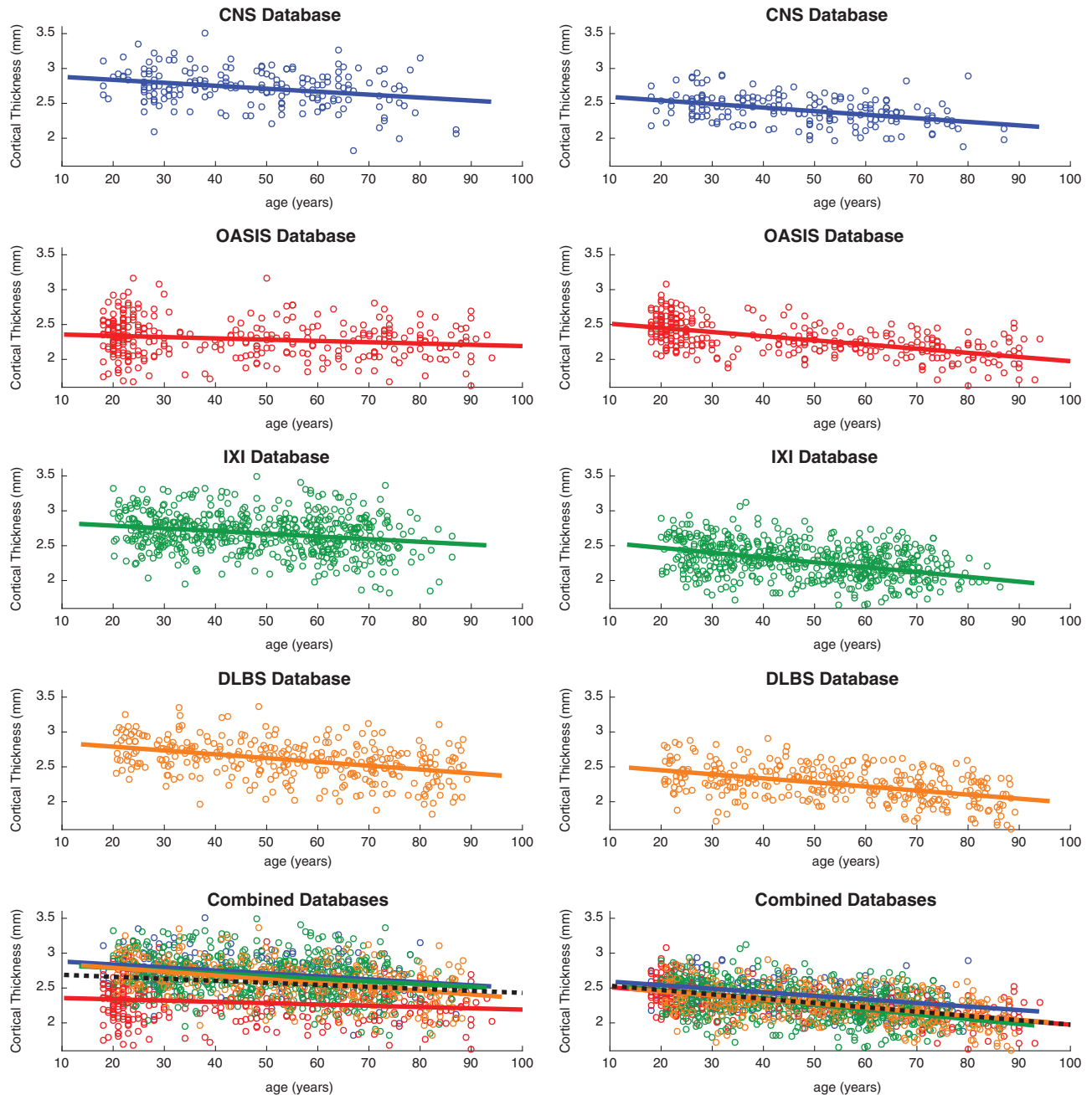


Figure 5. Example correlations of cortical thickness versus age. The first 4 rows of plots correspond to the 4 independent databases and are color coded accordingly, and the fifth row shows their overlap and consolidated correlation from all the databases (black dashed line). Data from 2 regions (R-banks superior temporal sulcus, left; R-parahippocampal gyrus, right) are shown here to illustrate differences between databases. The left column illustrates a case with large divergence between databases and the right column, a case with small divergence.

On average, the R^2 between all regions in all databases was 0.24, indicating that only a fraction (24%) of the cortical thickness variation can be accounted for by age. This is consistent with some other reports (Tamnes et al. 2010; Chen, Rosas, et al. 2011; Lemaitre et al. 2012; Tustison et al. 2014). The regional average R^2 were: 0.20, 0.23, 0.20, and 0.30 for the CNS, OASIS, IXI, and DLBS databases, respectively. Including other cofactors such as sex, age, handedness, have been shown to improve the R^2 slightly. Longitudinal analysis would improve the R^2 (Fjell et al. 2014), which suggests that a large portion of the remaining variance is attributable to individual variation. Furthermore, R^2 could

also be increased by selection of a more sophisticated model, such as localized regression (Fjell et al. 2014; Zhou et al. 2015).

The bootstrap results (Fig. 7) indicate the minimum number of subjects required to establish a consistent regression of cortical thickness versus age. We found that the total number of subjects required being on the order of 15–20 per decade. The minimum was similar between databases showing that this phenomenon is consistent; though, including more subjects would always yield more consistent regression model fits.

Symmetry of cortical thickness between hemispheres is widely assumed. In some studies only the bilateral average

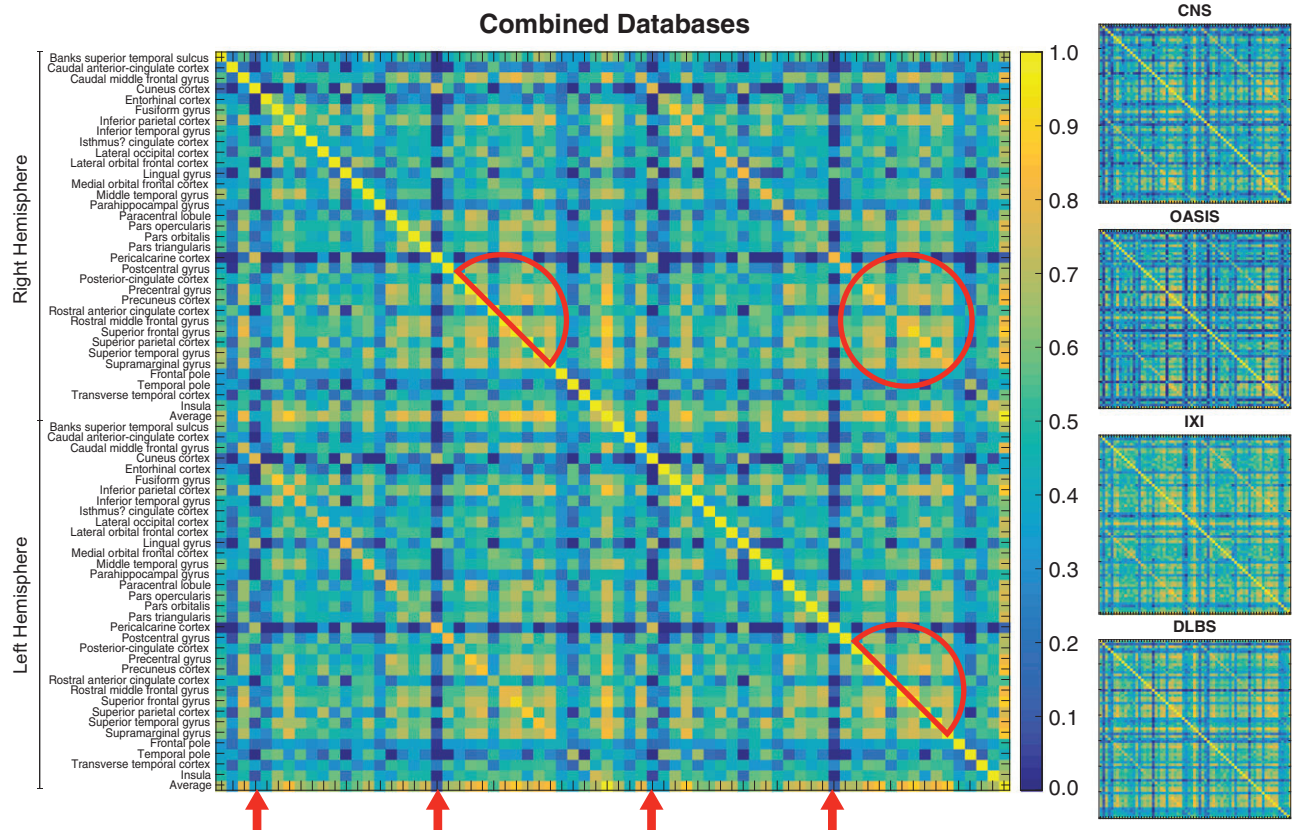


Figure 6. Correlation matrices. The large matrix (left) is the correlation matrix from the consolidated database. On the left-hand side of this matrix is the ordering of the regions. The left-hemisphere follow the right-hemisphere measurements. To the right of the large matrix are 4 plots showing the correlation matrix for each independent database. The regions encircled in red on the large matrix indicate regions of consistently high correlation between the databases, the red arrows point to bands, which are inconsistent between databases.

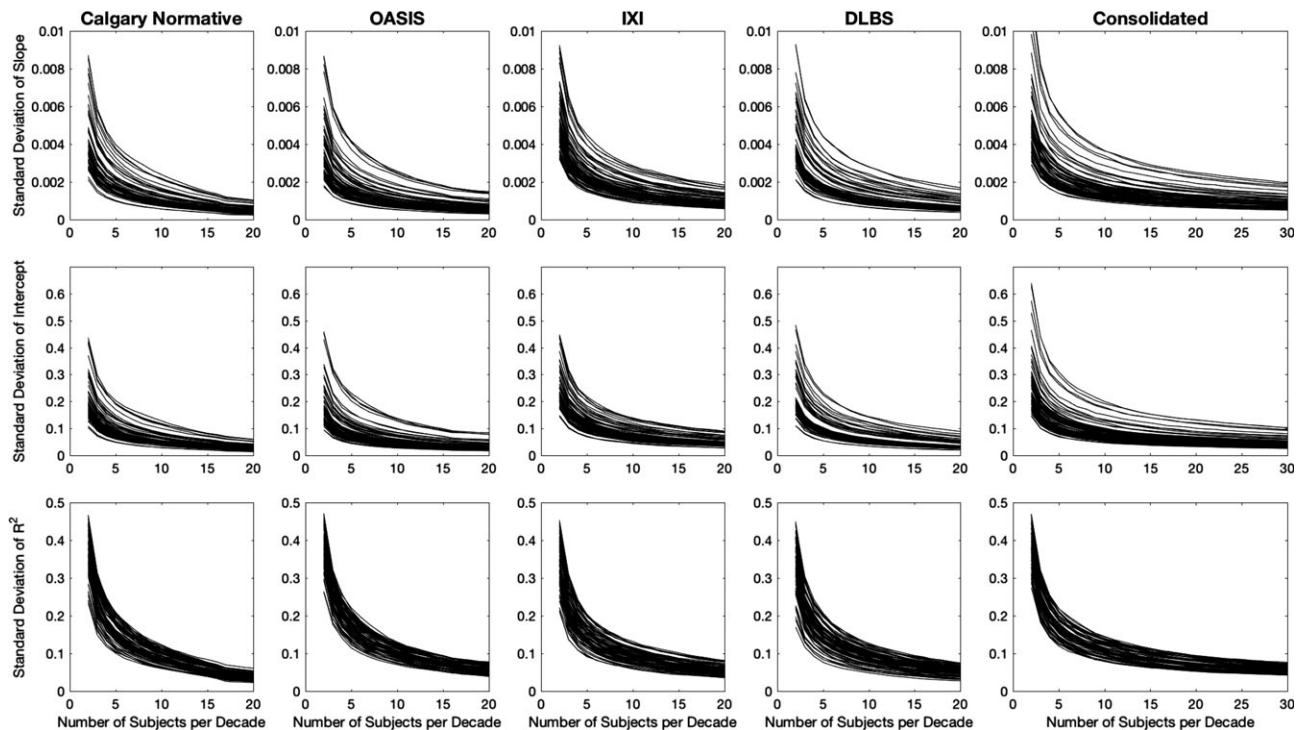


Figure 7. Bootstrapping analysis. Each graph plots the standard deviation of the slope (top row), intercept (middle row) and R^2 value (bottom row) versus number of subjects per decade. A total of 70 individual lines were calculated using 10 000 ordered permutations for each number of subjects per decade. As the number of subjects per decade increased, the standard deviation of the regression parameters decreased and the regression result was more consistent with the inclusion of more data.

Table 2 Age prediction using multiple independent linear regression models. Table shows regression of various prediction analyses. The database used for calibration is listed in the first column. A leave-one-out cross validation was used first, the R^2 and mean absolute difference (MAD) in age between the predicted and chronological age are reported for each set of predictions. After cross validation, the model was calibrated with each database and then applied sequentially to all of the other databases and the R^2 and MAD shown in the remaining columns. Cells of the table are grayed out when the measurement should not be performed, as the test data would be the same as the calibration data

Calibration database	Leave one out		Test database							
			CNS		OASIS		IXI		DLBS	
	R^2	MAD	R^2	MAD	R^2	MAD	R^2	MAD	R^2	MAD
CNS	0.647	10.95			0.726	13.19	0.660	11.37	0.738	11.80
OASIS	0.815	11.51	0.702	10.24			0.690	13.41	0.752	12.12
IXI	0.746	8.74	0.699	9.81	0.773	14.00			0.749	11.46
DLBS	0.818	9.43	0.695	10.51	0.705	23.17	0.676	11.40		
Consolidated	0.784	9.80								

MAD = mean absolute difference (years).

cortical thickness is reported (Lemaitre et al. 2012). Although, Luders et al. (2006) have shown asymmetry. We did not average between hemispheres, and some bilateral symmetry can be observed in Figures 2–4 and 6. Figure 4 shows that although the laterality indices are small, they are nonzero and can be different between databases.

The implication of the age prediction portion of this study suggests that models should have enough subjects to accurately predict the age, and that data acquired in a fashion similar to the subject being predicted should be included in the calibration. A decrease in prediction outcome metrics was observed when using a different database for the calibration than testing. However, when all data were consolidated, the predictive power remained high. We used a simple multiple regression model, but higher-level models, such as neural networks, support vector machines, or random forest regression could also be used (Valizadeh et al. 2017).

Conclusion

We conclude that the cortical thinning relationships for each of the 4 databases considered are statistically different. The differences are small for most brain regions, yet was statistically significant. An observable difference exists between the correlation matrices and bootstrap responses, but these changes are subtle, implying consistency of the variance between databases. Differences detected between diseased and healthy groups should be greater than the differences observed here between healthy groups. Deviations in acquisitions and/or subject population characteristics should be minimized when amalgamating databases. Combining individual databases into large cohorts should be done with caution.

Funding

N.D.F. is supported by the Canada Research Chairs program. R.F. is the Hopewell Professor of Brain Imaging. G.B.P. is the Campus Alberta Innovation Program (CAIP) Chair of Healthy Brain Aging. The authors would like to thank the University of Calgary, Alberta Health Services and Compute Canada (Westgrid) for providing computational resources for these experiments. The authors would also like to thank those responsible for providing the open source databases used. Financial contributions from the Canadian Institutes for Health Research (CIHR) and the Campus Alberta Innovation Program

(CAIP) are also acknowledged. R.J.W. held a Post-doctoral Fellowship from the NSERC-CREATE I3T program. *Conflict of Interest:* None declared.

References

- Chen ZJ, He Y, Rosa-Neto P, Gong G, Evans AC. 2011. Age-related alterations in the modular organization of structural cortical network by using cortical thickness from MRI. *Neuroimage*. 56:235–245.
- Chen JJ, Rosas HD, Salat DH. 2011. Age-associated reductions in cerebral blood flow are independent from regional atrophy. *Neuroimage*. 55:468–478.
- Dale AM, Fischl B, Sereno MI. 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*. 9:179–194.
- Dale AM, Sereno MI. 1993. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *J Cogn Neurosci*. 5: 162–176.
- Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, et al. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 31:968–980.
- Dickerson BC, Fenstermacher E, Salat DH, Wolk DA, Maguire RP, Desikan R, Pacheco J, Quinn BT, Van der Kouwe A, Greve DN, et al. 2008. Detection of cortical thickness correlates of cognitive performance: reliability across MRI scan sessions, scanners, and field strengths. *Neuroimage*. 39:10–18.
- DLBS. 2017. Dallas Lifespan Brain Study Webpage. Available from: URL http://fcon_1000.projects.nitrc.org/indi/retro/dlbs.html.
- Duning T, Kloska S, Steinsträter O, Kugel H, Heindel W, Knecht S. 2005. Dehydration confounds the assessment of brain atrophy. *Neurology*. 64:548–550.
- Ericsson A, Aljabar P, Rueckert D. 2008. Construction of a patient-specific atlas of the brain: application to normal aging. 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, p. 480–483.
- Fischl B, Dale AM. 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci USA*. 97:11050–11055.
- Fischl B, Liu A, Dale AM. 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct

- models of the human cerebral cortex. *IEEE Trans Med Imaging*. 20:70–80.
- Fischl B, Sereno MI, Dale AM. 1999. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*. 9:195–207.
- Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, et al. 2004. Automatically parcellating the human cerebral cortex. *Cereb Cortex*. 14:11–22.
- Fjell AM, Westlye LT, Amlien I, Espeseth T, Reinvang I, Raz N, Agartz I, Salat DH, Greve DN, Fischl B, et al. 2009. High consistency of regional cortical thinning in aging across multiple samples. *Cereb Cortex*. 19:2001–2012.
- Fjell AM, Westlye LT, Grydeland H, Amlien I, Espeseth T, Reinvang I, Raz N, Dale AM, Walhovd KB. 2014. Accelerating cortical thinning: unique to dementia or universal in aging? *Cereb Cortex*. 24:919–934.
- Franke K, Gaser C, Manor B, Novak V. 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Front Aging Neurosci*. 5:90.
- Franke K, Hagemann G, Schleussner E, Gaser C. 2015. Changes of individual BrainAGE during the course of the menstrual cycle. *Neuroimage*. 115:1–6.
- Franke K, Luders E, May A, Wilke M, Gaser C. 2012. Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. *Neuroimage*. 63:1305–1312.
- Franke K, Ziegler G, Klöppel S, Gaser C. 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *Neuroimage*. 50:883–892.
- Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H, Alzheimer's Disease Neuroimaging Initiative. 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PLoS One*. 8:e67346.
- Gong G, He Y, Chen ZJ, Evans AC. 2012. Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex. *Neuroimage*. 59:1239–1248.
- Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R, et al. 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage*. 32:180–194.
- Hildebrand T, Rügsegger P. 1997. A new method for the model-independent assessment of thickness in three-dimensional images. *J Microsc*. 185:67–75.
- IXI. 2017. Information eXtraction of Images Database Website. Available from: URL <http://brain-development.org/ixi-dataset/>.
- Kempton MJ, Ettinger U, Foster R, Williams SCR, Calvert GA, Hampshire A, Zelaya FO, O'Gorman RL, McMorris T, Owen AM, et al. 2011. Dehydration affects brain structure and function in healthy adolescents. *Hum Brain Mapp*. 32:71–79.
- Lemaitre H, Goldman AL, Sambataro F, Verchinski BA, Meyer-Lindenberg A, Weinberger DR, Mattay VS. 2012. Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiol Aging*. 33:617.e611–617.e619.
- Lerch JP, Evans AC. 2005. Cortical thickness analysis examined through power analysis and a population simulation. *Neuroimage*. 24:163–173.
- Li Y, Wang Y, Wu G, Shi F, Zhou L, Lin W, Shen D. 2012. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiol Aging*. 33:427.e415–427.e430.
- Luders E, Narr KL, Thompson PM, Rex DE, Jancke L, Toga AW. 2006. Hemispheric asymmetries in cortical thickness. *Cereb Cortex*. 16:1232–1238.
- Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci*. 19:1498–1507.
- Mugler JP, Brookeman JR. 1990. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magn Reson Med*. 15:152–157.
- Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 53:695–699.
- OASIS. 2017. Open Access Series of Imaging Studies Webpage. Available from: URL <http://www.oasis-brains.org/>.
- Pletzer B, Kronbichler M, Aichhorn M, Bergmann J, Ladurner G, Kerschbaum HH. 2010. Menstrual cycle and hormonal contraceptive use modulate human brain structure. *Brain Res*. 1348:55–62.
- Potvin O, Dieumegarde L, Duchesne S. 2017a. Freesurfer cortical normative data for adults using Desikan-Killiany-Tourville and ex vivo protocols. *Neuroimage*. 156:43–64.
- Potvin O, Dieumegarde L, Duchesne S. 2017b. Normative morphometric data for cerebral cortical areas over the lifetime of the adult human brain. *Neuroimage*. 156:315–339.
- Potvin O, Mouiha A, Dieumegarde L, Duchesne S. 2016. Normative data for subcortical regional volumes over the lifetime of the adult human brain. *Neuroimage*. 137:9–20.
- Salat DH, Buckner RL, Snyder AZ, Greve DN, Desikan RSR, Busa E, Morris JC, Dale AM, Fischl B. 2004. Thinning of the cerebral cortex in aging. *Cereb Cortex*. 14:721–730.
- Sled JG, Zijdenbos AP, Evans AC. 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging*. 17:87–97.
- Smith EE, O'Donnell M, Dagenais G, Lear SA, Wielgosz A, Sharma M, Poirier P, Stotts G, Black SE, Strother S, et al. 2015. Early cerebral small vessel disease and brain volume, cognition, and gait. *Ann Neurol*. 77:251–261.
- Sowell ER, Peterson BS, Kan E, Woods RP, Yoshii J, Bansal R, Xu D, Zhu H, Thompson PM, Toga AW. 2007. Sex differences in cortical thickness mapped in 176 healthy individuals between 7 and 87 years of age. *Cereb Cortex*. 17:1550–1560.
- Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B. 2004. A hybrid approach to the skull stripping problem in MRI. *Neuroimage*. 22:1060–1075.
- Ségonne F, Pacheco J, Fischl B. 2007. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans Med Imaging*. 26:518–529.
- Tamnes CK, Østby Y, Fjell AM, Westlye LT, Due-Tønnessen P, Walhovd KB. 2010. Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. *Cereb Cortex*. 20:534–548.
- Tombaugh TN, McIntyre NJ. 1992. The mini-mental state examination: a comprehensive review. *J Am Geriatr Soc*. 40:922–935.
- Tsang A, Lebel CA, Bray SL, Goodyear BG, Hafeez M, Sotero RC, McCreary CR, Frayne R. 2017. White matter structural connectivity is not correlated to cortical resting-state functional

- connectivity over the healthy adult lifespan. *Front Aging Neurosci.* 9:144.
- Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, Kandel BM, van Strien N, Stone JR, Gee JC, et al. 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage.* 99:166–179.
- Valizadeh SA, Hänggi J, Mérillat S, Jäncke L. 2017. Age prediction on the basis of brain anatomical measures. *Hum Brain Mapp.* 38:997–1008.
- Zhou D, Lebel C, Treit S, Evans A, Beaulieu C. 2015. Accelerated longitudinal cortical thinning in adolescence. *Neuroimage.* 104:138–145.