J Math Imaging Vis (2011) 40: 20–35
DOI 10.1007/s10851-010-0240-4

Statistical Analysis of Cortical Morphometrics Using Pooled Distances Based on Labeled Cortical Distance Maps

E. Ceyhan · M. Hosakere · T. Nishino · J. Alexopoulos · R.D. Todd · K.N. Botteron · M.I. Miller · J.T. Ratnanather

Published online: 24 November 2010

Abstract Neuropsychiatric disorders have been demonstrated to manifest shape differences in cortical structures. Labeled Cortical Distance Mapping (LCDM) is a powerful tool in quantifying such morphometric differences and characterizes the morphometry of the laminar cortical mantle of cortical structures. Specifically, LCDM data are distances of labeled gray matter (GM) voxels with respect to the gray/white matter cortical surface. Volumes and descriptive measures (such as means and variances for each subject) based on LCDM distances provide descriptive summary information on some of the shape characteristics. However, additional morphometrics are contained in the data and their analysis may provide additional clues to underlying differences in cortical characteristics. To use more of this information, we pool (merge) LCDM distances from subjects in the same group. These pooled distances can help detect morphometric differences between groups, but do not provide information about the locations of such differences in the tissue in question. In this article, we check for the influence of the assumption violations on the analysis of pooled LCDM distances. We demonstrate that the classical parametric tests are robust to the non-normality and within sample dependence of LCDM distances and nonparametric tests are robust to within sample dependence of LCDM distances. We specify the types of alternatives for which the tests are more sensitive. We also show that the pooled LCDM distances provide powerful results for group differences in distribution of LCDM distances. As an illustrative example, we use GM in the ventral medial prefrontal cortex (VMPFC) in subjects with major depressive disorder (MDD), subjects at high risk (HR) of MDD, and healthy subjects. Significant morphometric differences were found in VMPFC due to MDD or being at HR. In particular, the analysis indicated that distances in left and right VMPFCs tend to decrease due to MDD or being at HR, possibly as a result of thinning. The methodology can also be applied to other cortical structures.

Keywords Computational anatomy · Depression · Laminar cortical mantle · Morphometry · Ventral medial prefrontal cortex

E. Ceyhan (✉)
Dept. of Mathematics, Koç University, 34450 Sarıyer, Istanbul, Turkey
e-mail: elceyhan@ku.edu.tr

E. Ceyhan · M. Hosakere · M.I. Miller · J.T. Ratnanather
Center for Imaging Science, The Johns Hopkins University, Baltimore, MD 21218, USA

T. Nishino · J. Alexopoulos · K.N. Botteron
Dept. of Psychiatry, Washington University School of Medicine, St. Louis, MO 63110, USA

T. Nishino · K.N. Botteron
Dept. of Radiology, Washington University School of Medicine, St. Louis, MO 63110, USA

R.D. Todd
Dept. of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA

M.I. Miller · J.T. Ratnanather
Institute for Computational Medicine, The Johns Hopkins University, Baltimore, MD 21218, USA

M.I. Miller · J.T. Ratnanather
Dept. of Biomedical Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA

1 Introduction

In the past 15 years, the laminar structure of the neo-cortex has received considerable attention thanks to advances in

high resolution magnetic resonance imaging (MRI) technology and the development of Computational Anatomy (CA) methods (e.g., [3, 7, 13, 15, 17, 19]). Specifically, Labeled Cortical Distance Mapping (LCDM) has been shown to be a powerful tool for structural comparison of cortical thickness characteristics in the cingulate cortex in studies of Alzheimer's disease and schizophrenia [1, 16, 26].

LCDM characterizes the morphometry of the laminar cortical mantle. The term "morphometry" here has two components, the structural formation (like surface and form) of the tissue and scale or size (like volume and surface area). Thus, morphometry refers to all aspects of laminar shape, where "shape" refers to the surface structure, while "size" refers to the scale of the tissue in question. Specifically, LCDM data are distances of labeled gray matter (GM) voxels with respect to the gray/white matter (GM/WM) cortical surface. Hence LCDM distances are *local measures* characterizing the morphometry of the cortical mantle.

In this article, we assess the use of *pooling* of LCDM distances in discriminating between diagnostic groups. In particular we consider LCDM data for the Ventral Medial Prefrontal Cortex (VMPFC) which is implicated in major depressive disorders (MDD) [10–14]. Abnormalities have been demonstrated in structure and function of the prefrontal cortex due to MDD [10, 11]. Other structural imaging studies have largely focused on adult onset MDD, while only few have focused on early onset MDD. Structural deficits in a subregion of the VMPFC, i.e., subgenual prefrontal cortex, have also been associated with early onset of MDD [2].

Previously, we analyzed morphometric measures (i.e., volume and descriptive summary statistics based on LCDM distances such as median, mode, range, and variance) and demonstrated that except for left-right asymmetry and correlation between left and right measures, these variables usually *failed* to discriminate between MDD and healthy groups [5]. This may be due to the fact that the subjects are age-matched female twins, whose VMPFC may be similar in size. This might also be partly due to the small sample size (i.e., number of subjects). On the other hand, by only using a descriptive summary statistic (such as volume or median) of the numerous distances for each person, we essentially lose most of the information provided by LCDM measures. Therefore, we suggest a strategy to avoid such information loss and to more fully utilize the shape or morphometric characteristics contained in the data by using all of the LCDM distances. Along these lines, we pool (i.e., merge) the LCDM distances by condition or group and use the pooled distances to detect morphometric differences. However the pooled distances do not have within sample independence, as the distances of neighboring voxels of each voxel are dependent. Moreover, there is also dependence between distances in left and right VMPFC in each subject, as they belong to the same person. But we demonstrate that within sample dependence does not affect the tests in terms of empirical significance levels (or Type I errors) or power. Throughout the article, we use $\alpha = 0.05$ as the significance level to declare a p-value to be significant.

We describe the acquisition of LCDM distances in Sect. 2.1, the methods we employ in Sects. 2.2 and 2.3, present the analysis of pooled distances in Sect. 3, and investigate the influence of assumption violations in Sect. 4.

2 Methods

2.1 Data Description and Acquisition

A cohort of 34 right-handed young female twin pairs between the ages of 15 and 24 years old were obtained from the Missouri Twin Registry in order to study cortical changes in the VMPFC associated with MDD. The inclusion criteria for affected twin pairs were onset prior to age 16 and the DSM-IV criteria for MDD being greater than duration of 4 weeks. Control twin pairs had no personal or first degree of family history of MDD. Both monozygotic and dizygotic twin pairs were included, of which 14 pairs were controls (Ctrl) and 20 pairs had one twin affected with MDD, their co-twins were designated as the HR group. Three high resolution T1-weighted MPRAGE magnetic resonance scans of each subject in this population were acquired using a Siemens scanner with 1 mm^3 isotropic resolution. Images were then averaged, corrected for intensity inhomogeneity and interpolated to $0.5 \times 0.5 \times 0.5$ mm^3 isotropic voxels. Following [23], a region of interest (ROI) comprising the VMPFC stripped of the basal ganglia, eyes, sinus, cavity, was defined manually and segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) by Bayesian segmentation using the expectation maximization algorithm [12]. A triangulated representation of the cortex at the GM/WM boundary was generated using isocontouring algorithms [12].

Bayesian segmentation [12] automatically segments the tissue via the Expectation-Maximization minimization of Gaussians for the three tissue classes at each voxel. Partial volume i.e. voxels that share mixtures were resolved via a Neymann-Pearson recalibration of the segmentation based on a training set [23]. The threshold between GM and WM was used to generate a triangulated isosurface via the marching tetrahedra algorithm i.e. the mesh is dense. Validation with several VMPFC subvolumes yielded misclassification errors of 0.05–0.10 ($n = 5$) for the segmentation and subvoxel accuracy of the isosurface with 50 percent of the vertices within 0.12–0.28 mm ($n = 14$) from semi-automated contours [23].

LCDM is generated as follows: first, the ROI subvolume is partitioned by a regular lattice of voxels of specific size h, denoted $V(h)$. Every voxel is labeled by tissue type

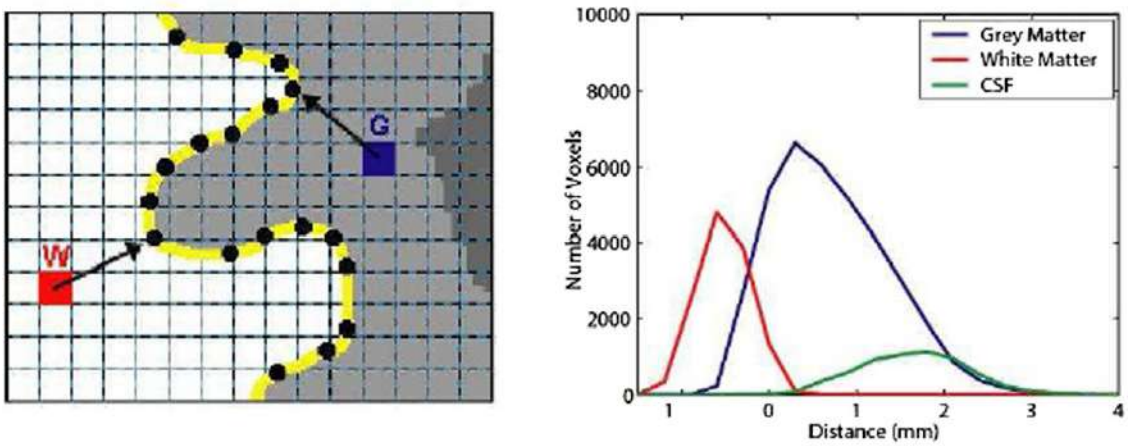

Fig. 1 A two-dimensional illustration of normal distances from a GM and a WM voxel to the GM/WM surface (*left*) and non-normalized histograms of LCDM distances of GM, WM, and CSF tissues (*right*)

as gray matter (GM), white matter (WM), or cerebrospinal fluid (CSF) (see, e.g., [12, 17]). For every GM voxel in the ROI, the distance from the centroid of the voxel to the closest point on GM/WM surface is computed. Let $S(\Delta)$ be the triangulated graph representing the GM/WM surface. An LCDM distance is a set distance function $d : v_i \in V(h) \to d(v_i, S(\Delta))$, the distance between the centroid of voxel v_i and the set $S(\Delta)$; that is, it is the distance from the center of the voxel to the closest vertex on the surface. More precisely,

$$D_i := d(C_M(v_i), S(\Delta)) = \min_{s \in S(\Delta)} \|C_M(v_i) - s\|_2 \qquad (1)$$

where $C_M(\cdot)$ stands for center of mass (or centroid), and $\|\cdot\|_2$ is the usual L_2-norm. We use a signed (or labeled) distance to indicate the location of each voxel with respect to the GM/WM surface. Figure 1 illustrates the computation of distances between labeled voxels and the cortical surface; also shown is the corresponding non-normalized histograms of LCDM distances. Observe that GM tissue comprises most of the cortex, and by construction, while most of GM distances are positive, most of WM distances are negative, and all of CSF distances are positive. Negative distances for some GM close to the GM/WM boundary are possible by construction, because the surface is constructed in such a way that a surface is always intersecting voxels, i.e., partial volume. So some appropriately labeled GM voxels may fall on a side of surface that they should not belong to. However, these mislabeled voxels constitute a small proportion of all voxels and do not affect the overall analysis. Reliability of LCDMs is dependent on GM segmentation and reconstruction of GM/WM surfaces which has been validated for several cortical structures including VMPFC [23], cingulate cortex [24, 26] and planum temporale [22]. Condensing to a single distance value for each vertex on the surface is the next logical step in extending LCDM. This is called Local LCDM or LLCDM and is useful in comparing thickness across multiple subjects for a cortical structure (see [20, 21]).

For the left ROI, let D^L be the set of LCDM distances, D^L_{ijk} be the distance calculated as in (1) and associated with k^{th} voxel in subject j in group i for $k = 1, 2, \ldots, K_{ij}$, $j = 1, 2, \ldots, n_i$ and $i = 1, 2, 3$ (here group 1 is for MDD, 2 is for HR, and 3 is for Ctrl). Thus, $n_1 = 20$, $n_2 = 20$, and $n_3 = 28$. Right distances D^R are denoted similarly as D^R_{ijk}. Based on prior anatomical knowledge (e.g., [14]), cortical thickness of the VMPFC is roughly 6 mm, so we can safely retain distances between -0.5 mm and 5.5 mm so that (potentially) mislabeled GM is excluded from the data. In this particular case for the left and right VMPFC, only 0.16% and 0.14% of distances were below -0.5 mm respectively; similarly, only 0.22% and 0.07% of distances were above 5.5 mm, respectively.

2.2 Pooling LCDM Distances by Group

Although the descriptive measures such as mean, median, and variance of LCDM distances are *global measures* regarding the morphometry of VMPFC, they are summary statistics (such as volume or median), so they tend to oversimplify the data since instead of a large number of LCDM distances per subject, we will have two (e.g., one mean value for left, one for right VMPFC) measures for each subject [5]. Hence we lose most of the information conveyed by the LCDM distances. A solution to this problem is using all the LCDM distances in our analysis. So we *pool* LCDM distances of subjects from the same group and thereby obtain yet another *global measure* of morphometry. That is, we pool the LCDM distances for all left MDD VMPFCs, those for all left HR VMPFCs, and those for all left Ctrl

Table 1 The sample sizes (n), means, medians, and standard deviations (SD) of the pooled LCDM distances (in mm) for left and right VMPFCs categorized by group

Group	Left VMPFCs				Right VMPFCs			
	n	mean	median	SD	n	mean	median	SD
MDD	238937	1.62	1.46	1.13	170534	1.63	1.49	1.10
HR	228224	1.61	1.46	1.11	216978	1.59	1.46	1.08
Ctrl	308498	1.66	1.50	1.14	293479	1.66	1.53	1.12
Overall	775659	1.63	1.48	1.13	680991	1.63	1.50	1.10

VMPFCs. Likewise, we pool the right VMPFC LCDM distances. Thus, for left VMPFCs

$$D_i^L = \{D_{i\ell}^L : \ell = 1, 2, \ldots, N_i\} = \bigcup_{j=1}^{n_i} D_{ijk}^L \tag{2}$$

where $D_{i\ell}^L$ is the ℓ^{th} distance in group i and $N_i = \sum_{j=1}^{n_i} K_{ij}$ is the number of distances (i.e., GM voxels) in group i for $i = 1, 2, 3$ (group 1 is for MDD, group 2 for HR, and group 3 for Ctrl). Similarly, we denote the right pooled distances as D_i^R. Furthermore, we denote the overall (i.e., groups combined) pooled left and right distances as $D^L = \bigcup_{i=1}^3 D_i^L$ and $D^R = \bigcup_{i=1}^3 D_i^R$, respectively. See Table 1 for the corresponding sample sizes, means, and standard deviations of the pooled LCDM distances, overall and for each group. For pooling the LCDM distances, the most crucial assumption is that the subjects with the same diagnosis have similar VMPFC in morphometry, which is reasonable in practice. By pooling, the most common characteristics of the VMPFC specific to a diagnostic group are emphasized, while the differences at the individual (i.e., subject) level are downplayed. Furthermore, the pooled distances will be more powerful in detecting the differences between LCDM distances (hence differences in morphometry).

2.3 Statistical Tests

There is an inherent dependence between LCDM distances of voxels to the gray matter/white matter boundary due to spatial correlation at the level of individual subjects. When we pool the LCDM distances by group, this spatial dependence is not removed. That is, pooling neither creates nor removes the inherent dependence of the distances, as it only ignores the subject information. We compare the distributions and central measures (e.g., means) of the LCDM distances using various statistical tests. In particular, we consider Kruskal-Wallis (K-W) test for omnibus multi-group comparison of the LCDM distributions and ANOVA F-tests for omnibus multi-group comparison of the LCDM means. For k groups the null hypothesis for K-W test is $H_0 : F_1 = F_2 = \cdots = F_k$ where F_i is the distribution function of group i and the null hypothesis for ANOVA F-test is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i is the mean of group i, for $i = 1, 2, \ldots, k$. For comparison of distributions of LCDM distances of pairs of groups, we apply Wilcoxon rank sum test and Kolmogorov-Smirnov (K-S) tests; and for comparisons of means of pairs of LCDM distance groups, we apply Welch's t-test (see [8] for more detail on these tests). For pairwise comparisons, Wilcoxon rank sum test is done as a post hoc test after a significant K-W test, because Wilcoxon rank sum and K-W tests are both variants of the same test for multiple or two group comparison. K-S test is performed to determine the stochastic ordering. Wilcoxon rank sum test (also called the Mann–Whitney U test) is a non-parametric test for assessing whether two independent samples of observations have similar values. It is based on the sum of the ranks of the two independent samples, when the samples are pooled together. Under the null hypothesis, it is assumed that the distributions of both groups are equal, i.e., $H_0 : F_1 = F_2$. In other words, the probability of an observation from the first population being larger than the one from the second population is the same as the probability of an observation from the second population being larger than the first. For two groups, the K-S test is a non-parametric test based on the estimated maximum difference between the cumulative distributions of the two groups. Under the null hypothesis, it is assumed that the distributions are equal, i.e., $H_0 : F_1 = F_2$. Welch's t-test is an extension of the usual Student's t-test and is intended for use with two samples having (possibly) unequal variances. The null hypothesis for Welch's t-test is $H_0 : \mu_1 = \mu_2$.

Wilcoxon and t-tests imply an ordering in a location parameter such as mean or median. Stochastic ordering, if present, can be deduced from the direction of the alternative, together with the graph of the cumulative distribution functions (cdfs). However, we can also use Kolmogorov-Smirnov (K-S) tests for $H_0 : F_1 = F_2$. Although Wilcoxon rank sum and K-S tests have the same null hypothesis, Wilcoxon test gives an overall distribution comparison based on the rankings of the observations, while K-S test compares the cdfs of the observations at values where the maximum differences between cdfs occur. Hence Wilcoxon test can be significant for only one of the one-sided alternatives, while K-S test yields p-values that are not complementary for the one-sided alternatives (i.e., they don't add

up to 1). Hence, p-values can be significant for both or none of the directional alternatives. This results from the fact that, the order of the cdfs F_1 and F_2 can be different at different distance values (plotted on the horizontal axis). Moreover, if p-value based on K-S test is significant for only one-sided alternative, then we can also deduce stochastic ordering. The p-values being insignificant or significant for both one-sided alternatives imply lack of stochastic ordering. But the first case implies that equality of the distributions is retained, while the latter implies that the distributions are different. Although K-S test does not provide the actual values where the significant differences between cdfs occur, it is more informative and suggestive of distributional differences compared to Wilcoxon rank sum test. Furthermore, different cdf orderings at different values can be masked by the Wilcoxon rank sum test. Hence K-S test is more informative compared to Wilcoxon rank sum test.

We perform the omnibus multi-group tests before the pairwise comparison tests, because if a multi-group test is not significant, there is no need to perform the pairwise tests. For example, if K-W test is not significant, then the distributions of the LCDM distances of groups are not significantly different, hence Wilcoxon rank sum test on each pair of groups is redundant. On the other hand, if a multi-group test yields a significant result, it only means that there are some significant differences between the groups, but does not indicate which groups are different. To determine the pairs that have significant difference, we have to perform the pairwise comparison methods. Among the tests we consider, K-W and ANOVA F-tests are omnibus tests, and Wilcoxon rank sum and Welch's t-tests are for commonly used multiple comparison procedures after obtaining a significant omnibus test result. Rejecting an omnibus test for k groups suggest that there are differences between some pair(s) of the groups, and to determine which pair(s) exhibit significant differences, $k(k-1)/2$ pairwise comparisons are needed. Hence, for large k values, an omnibus test might save a great deal of time and energy since after an insignificant omnibus test, there is no need for the pairwise tests. For small k values, one might do an omnibus test followed by pairwise tests, or just the pairwise tests directly. However, for even $k = 4$, we need 6 pairwise tests, and this might still be too many pairwise tests, if omnibus test were insignificant.

For the nonparametric tests (K-W, Wilcoxon rank sum, and K-S tests) only within sample independence is violated, but for the parametric tests (ANOVA F-tests and t-test), the assumptions of normality (i.e., Gaussianity) and within sample independence are violated. See [4] for a complete list of assumptions for each of these tests. However, we investigate the influence of assumption violations on both nonparametric and parametric tests in Sect. 4 by an extensive Monte Carlo simulation study where we find the effect of assumption violations is negligible and we conjecture that this is due to the fact that the correlation structure is similar for each person (hence for each group). Moreover, our analysis does not concern inference for single populations but comparison of multiple populations. Given the difficulty to develop a method that accounts for spatial correlations, we ignore this type of spatial dependence henceforth.

In the analysis of the pooled distances, we apply classical parametric and nonparametric tests to detect the differences in LCDM distances due to diagnostic group factors. Such differences will imply morphometric changes (if any) due to the particular disease in question. K-W test provides an overall test of distributional equality for multiple groups. That is, if K-W test yields a significant p-value, then we conclude that LCDM distances are different in distribution for at least two groups, but it does not indicate which pair or pairs of groups exhibit differences. To find out which pairs exhibit significant distributional differences, we apply Wilcoxon rank sum test for each pair of LCDM groups. On the other hand, K-S test is only applicable to compare the distributions of two LCDM groups. Similarly, if an ANOVA F-test yields a significant p-value, it implies that the mean LCDM distances are different for at least two LCDM distance groups. To find out which pairs exhibit significant mean differences, we apply Welch's t-test for each pair of LCDM groups. The p-values for the t-test and Wilcoxon rank sum test are complementary, in the sense that p-values for the one-sided alternatives add up to 1 and can be significant for only one of the one-sided alternatives. Hence, Wilcoxon test provides an overall distributional comparison for two LCDM groups. On the other hand, p-values for K-S test are not complementary, as they do not add up to 1 for the one-sided alternatives. For example, one might have significant p-values for both of the one-sided alternatives, which implies that at a particular distance value, a group's empirical cumulative distribution function (ecdf) is significantly larger, while at another distance value the other group's ecdf is larger. Wilcoxon test (together with the ecdf plots) and K-S tests (either with p-values for both one-sided tests or with the ecdf plots) might provide the stochastic ordering (if present) of pooled distances.

3 Analysis of Pooled LCDM Distances

First we test for any distributional differences between the LCDM distances of the three diagnostic groups by K-W test and apply the ANOVA F-tests (with or without assuming homogeneity of variances (HOV)) for the equality of the means of the left and right LCDM distances of the three groups. The null hypothesis for these tests are provided in Sect. 2.3 (see also [4]).

The left and right pooled distances for each group are significantly non-normal (i.e., their distributions are significantly different from a Gaussian distribution) where based

on Lilliefor's test of normality $p < 0.0001$ for each test (see, e.g., [25]), due to the heavy right skew of the densities. This skew is biologically reasonable since most of the gray matter voxels will be expected to be near the GM/WM surface. Moreover, HOV is rejected ($p < 0.0001$ for both left and right pooled distances based on B-F test). Hence nonparametric tests of group comparisons would be more appropriate for this data. However, our Monte Carlo simulation results (see Sect. 4) suggest that both parametric and nonparametric tests are appropriate, with each being more sensitive for different alternatives.

The hypothesis of equality of the distributions of the pooled distances can be attributed to the similarity in the VMPFC shapes for all groups, but not vice versa (i.e., the equality of the distributions does not necessarily imply morphometric similarity, but only similarity in the distance structure of GM tissue with respect to the GM/WM surface). Notice that LCDM distances analyzed in this fashion provide morphometric information, on cortical mantle thickness and shape because the comparison is done on the ranking of distances (for K-W test) and means of the distances (for ANOVA F-tests) with respect to the GM/WM surface. For example, suppose two VMPFC tissues are composed of 100 and 1000 voxels of similar proportional distances, and then the test will detect no difference, although the morphometry is obviously different. Hence, as long as the voxels are at a similar distance from the GM/WM surface, their abundance will not influence the test results. That is, these tests are "independent of sample density" of LCDM distances.

The resulting p-values are presented in Table 2. Observe that there are significant differences between the LCDMs of the three groups, i.e., the distributions (and hence the means) of the LCDM distances for at least two groups are significantly different. Hence we conclude that there are significant morphometric differences in both left and right VMPFCs of at least two of the diagnostic groups in question. Hence, we perform pairwise comparisons by Wilcoxon rank sum test and Welch's t-test for left (and right) distances, using Holm's correction for multiple comparisons. In fact, we could start with pairwise tests directly, since we have only three diagnostic groups. However, for completeness and generality, we follow the more conventional path with an omnibus multi-group test followed by pairwise tests. The simultaneous hypotheses for Wilcoxon tests for left pooled LCDM distances are

$$H_{0,1}: F_1^L = F_2^L, \qquad H_{0,2}: F_1^L = F_3^L, \qquad H_{0,3}: F_2^L = F_3^L. \tag{3}$$

The less-than alternative for pairwise Wilcoxon tests is then

$$H_{a,1}: F_1^L > F_2^L, \qquad H_{a,2}: F_1^L > F_3^L, \qquad H_{a,3}: F_2^L > F_3^L. \tag{4}$$

Notice that if, for example, MDD left distances tend to be smaller than HR left distances, then the corresponding distribution functions have the opposite order, i.e., $F_1^L > F_2^L$. Hence the left sided (i.e., less than) alternative for LCDM distances implies that MDD pooled distances tend to be smaller than Ctrl pooled distances, and HR pooled distances tend to be smaller than Ctrl pooled distances and MDD pooled distances tend to be smaller than HR pooled distances. The greater than alternatives are similar except the inequalities should be reversed. Then we adjust these p-values for simultaneous comparisons by Holm's correction method for each alternative. We perform a similar analysis for right pooled distances.

The null hypotheses for pairwise t-tests are similar to the ones provided in (3) and (4) with F being replaced by μ and the inequalities reversed.

We present the p-values in Table 3. Observe that the distributions of LCDM distances for MDD and HR groups are not significantly different for both left and right VMPFCs (p-values based on Wilcoxon rank sum test are 0.3022 and 0.0776, respectively). On the other hand, mean LCDM distances for MDD subjects are significantly larger than that for HR subjects for both left and right VMPFCs (p-values based on t-test are 0.0383 and 0.0041, respectively). This seemingly contradictory situation occurs since the LCDM distances are highly skewed right. The LCDM distances for both MDD and HR left VMPFCs tend to be significantly smaller than those of Ctrl left VMPFCs. The same holds for the right VMPFCs also.

Stochastic ordering of the distances could be deduced from the direction of the alternative, together with the graph of the cdfs. See Fig. 5 for the cdf plots of the pooled distances. Although K-S test do not provide the actual distance values where the significant differences between cdfs occur, it is more informative and suggestive of distributional differences than Wilcoxon tests. Furthermore, different cdf orderings at different distance values are masked by the Wilcoxon test in MDD and HR left distances. The associated p-values are presented in Table 4 where tests for the alternatives are adjusted by Holm's correction method. Observe that the cdf of Ctrl-left distances is significantly smaller than those of MDD and HR-left distances. Furthermore, the cdfs of MDD and HR-left distances are significantly different from each other, with both sides being significant, which suggests that the order of cdf comparisons changes at different distance values. Thus, we conclude that MDD-left $<^{ST}$ Ctrl-left and HR-left $<^{ST}$ Ctrl-left where $<^{ST}$ stands for "stochastically smaller than". That is, it is more likely for MDD- or HR-left distances to be smaller compared to Ctrl-left distances.

The cdf of MDD-right distances is significantly smaller than HR-right distances which implies HR-right $<^{ST}$ MDD-right. But K-S test yields significant result for both types of one-sided alternative for MDD-right, Ctrl-right and HR-right, Ctrl-right and MDD-left and HR-left pairs (see Table 4

Table 2 The p-values for the multi-group comparisons of the pooled LCDM distances by K-W test, ANOVA F-tests with and without HOV. p_{KW}: p-value for K-W test, p_{F_1}, p_{F_2}: p-values for ANOVA F-tests with and without HOV, respectively

Multi-group comparisons of the pooled distances	
Left	Right
$p_{KW} < 0.0001$, $p_{F_1} < 0.0001$, $p_{F_2} < 0.0001$	$p_{KW} < 0.0001$, $p_{F_1} < 0.0001$, $p_{F_2} < 0.0001$

Table 3 The p-values for the simultaneous pairwise comparisons of the pooled distances by Wilcoxon rank sum test and the t-test. The p-values are adjusted by Holm's correction method ($g(\ell)$: first group is greater (less) than the second group)

Pair	With Wilcoxon rank sum test		With t-test	
	Left	Right	Left	Right
MDD, HR	0.3022 (ℓ)	0.0776 (g)	0.0383 (g)	0.0041 (g)
MDD, Ctrl	<0.0001 (ℓ)	<0.0001 (ℓ)	<0.0001 (ℓ)	<0.0001 (ℓ)
HR, Ctrl	<0.0001 (ℓ)	<0.0001 (ℓ)	<0.0001 (ℓ)	<0.0001 (ℓ)

Table 4 The p-values for the cdf comparisons (overall and by group) of the pooled LCDM distances. The p-values for each type of alternative are adjusted by Holm's correction method

p-values for cdf comparisons						
Pair	Left			Right		
	2-sided	$1^{st} < 2^{nd}$	$1^{st} > 2^{nd}$	2-sided	$1^{st} < 2^{nd}$	$1^{st} > 2^{nd}$
MDD, HR	<0.0001	<0.0001	0.0073	0.0316	0.0158	0.6017
MDD, Ctrl	<0.0001	0.5362	<0.0001	<0.0001	0.0069	<0.0001
HR, Ctrl	<0.0001	0.4170	<0.0001	<0.0001	0.0043	<0.0001

and Fig. 5). This implies, for example, the cdfs of MDD-right and Ctrl-right distances are different, but the differences between the cdfs of the groups change over the distance values; that is, for small distances, the order of cdfs for right distances is Ctrl < MDD < HR, which is the order for the proportion of voxels with smaller distances to the total number of voxels. Hence there is no stochastic ordering between them. That is, the proportion of voxels with smaller distances is largest for HR subjects and smallest for Ctrl subjects. For large distances the order of cdfs for right distances is HR < MDD < Ctrl, which can be interpreted similarly. This result indicates the cortical thinning for HR and MDD subjects compared to Ctrl subjects in the right VMPFC.

4 The Influence of Assumption Violations: A Monte Carlo Analysis

In this section, we investigate the influence of the assumption violations due to the spatial correlation and non-normality inherent in the LCDM distances on the tests. The most crucial step in a Monte Carlo simulation is being able to generate distances resembling those of LCDM distances of GM in VMPFCs; i.e., simulating the true randomness in LCDM distances.

For illustrative purposes, we choose the left VMPFC of HR subject 1. Recall that the LCDM distances for left VMPFC of HR subject 1 are denoted as D_{21}^L. We rearrange the distances, D_{21}^L, so that first stack of distances is in the interval $I_0 := [-1, 0.5]$ mm, the second stack of distances is in $I_1 := (0.5, 1.0]$ mm, the third stack of distances is in $I_2 := (1.0, 1.5]$ mm, and so on (until the last stack of distances is in $I_{11} := (5.5, 6.0]$ mm). Let ν_i be the number of distances that fall in I_i, i.e., $\nu_i = |D_{21}^L \cap I_i|$, for $i = 0, 1, 2, \ldots, 11$. Hence $\nu = (\nu_0, \nu_1, \ldots, \nu_{11}) = (2059, 1898, 1764, 1670, 1492, 1268, 814, 417, 142, 81, 61, 16)$. Then we merge these stacks into one group, (by appending $D_{21}^L \cap I_{i+1}$ to $D_{21}^L \cap I_i$ for $i = 1, 2, \ldots, 10$). See Fig. 2, where the left graph is for the stacked distances and the right graph is for distances sorted in ascending order.

A possible Monte Carlo simulation for these distances can be performed as follows. We independently generate n numbers in $\{0, 1, 2, \ldots, 11\}$ proportional to the above frequencies, ν_i, with replacement, i.e., with the discrete probability mass function $P_N(N_j = i) = \nu_i/11659$ for $i = 1, 2, \ldots, 11$ and $j = 1, 2, \ldots, n$. So, $P_N(N_j = i) = \nu_{p,i}$ where

$$(\nu_{p,0}, \nu_{p,1}, \ldots, \nu_{p,11}) = \nu_p$$

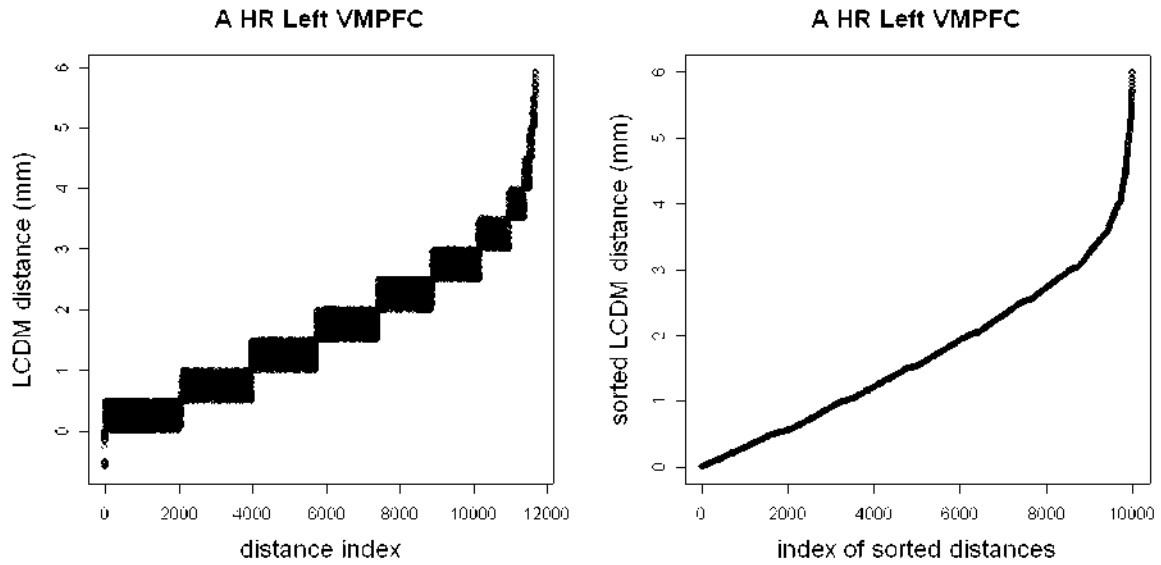

Fig. 2 Plots of the LCDM distances for the left VMPFC of HR subject 1. The *left plot* is for the distances stacked for intervals of size 0.5 mm and the *right plot* is for the sorted distances

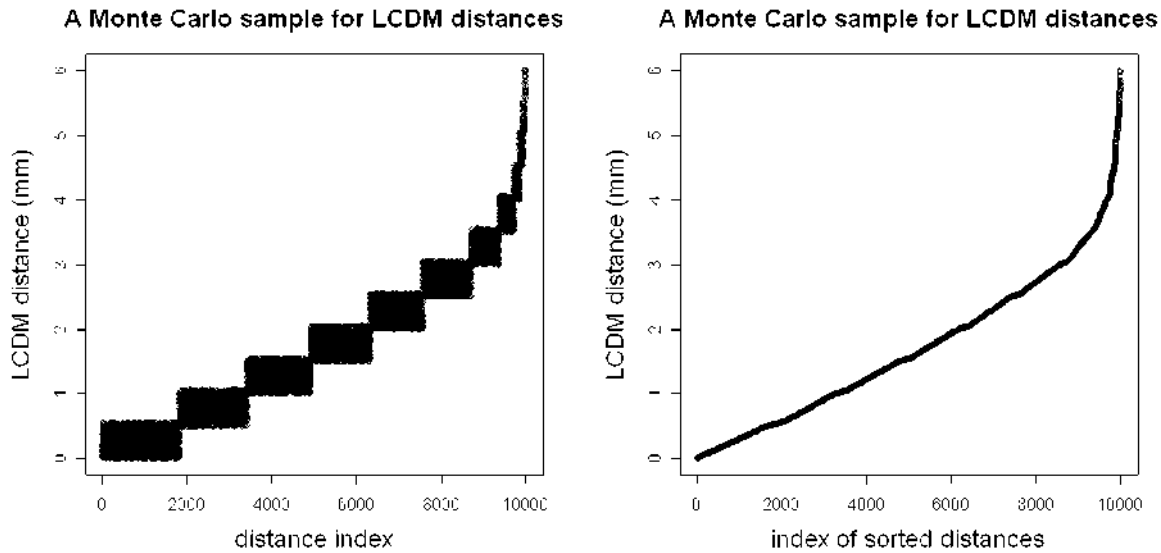

Fig. 3 Plots of the data values generated by Monte Carlo simulation to resemble LCDM distances. The *left plot* is for the distances stacked for intervals of size 0.5 and the *right plot* is for the sorted distances

$$= (0.177, 0.163, 0.151, 0.143, 0.126, 0.109, 0.070, 0.036, 0.012, 0.007, 0.005, 0.001). \quad (5)$$

Let n_i be the frequency of i among the n generated numbers from $\{0, 1, 2, \ldots, 11\}$ with distribution P_N, for $i = 1, 2, \ldots, 11$. Hence $n = \sum_{i=0}^{11} n_i$. Then we generate as many $U(0, 1)$ numbers for each $i \in \{0, 1, 2, \ldots, 11\}$ as i occurs in the generated sample of 1000 numbers and add these uniform numbers to i. That is, we generate $U_{ik} \sim \mathcal{U}(0, 1)$ for $k = 1, 2, \ldots, n_i$ for each i. Then we divide each distance by 2 to make the range of generated distances $[0, 6.0]$ mm which is the range of D_{21}^L, so the desired distance values are $d_{ik} = (i + U_{ik})/2$. Hence the set of simulated distances is

$$D_{mc} = \{d_{ik} = (i + U_{ik})/2 : U_{ik} \sim \mathcal{U}(0, 1) \text{ for } k = 1, \ldots, N_i \text{ and } N_i \sim P_N \text{ for } i = 0, 1, 2, \ldots, 11\}. \quad (6)$$

A sample of the distances generated in this fashion is plotted in Fig. 3 where the left plot is for the distances as they are generated at each bin (stack) of size 0.5 mm, the right plot is for the distances sorted in ascending order. Com-

Table 5 Estimated significance levels and proportions of agreement between the tests based on Monte Carlo simulations of distances with three groups, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ with sizes n_x, n_y, and n_z, respectively, with $N_{mc} = 10000$ Monte Carlo replicates. $\hat{\alpha}_{KW}$ is the empirical size estimate for K-W test, $\hat{\alpha}_{F_1}$, $\hat{\alpha}_{F_2}$ are for ANOVA F-tests with and without HOV, respectively; $\hat{\alpha}_{KW,F_1}$, $\hat{\alpha}_{KW,F_2}$, and $\hat{\alpha}_{F_1,F_2}$ are the values of proportion of agreement between the indicated tests in the subscripts. The empirical sizes in the same row with the same superscript are not significantly different from each other

(n_x, n_y, n_z)	Empirical size			Prop. of agreement		
	$\hat{\alpha}_{KW}$	$\hat{\alpha}_{F_1}$	$\hat{\alpha}_{F_2}$	$\hat{\alpha}_{KW,F_1}$	$\hat{\alpha}_{KW,F_2}$	$\hat{\alpha}_{F_1,F_2}$
(1000, 1000, 1000)	0.0511^{a}	0.0508^{a}	0.0506^{a}	0.0417^{l}	0.0419^{l}	$0.0499^{\approx}$
(5000, 5000, 10000)	0.0495^{a}	0.0498^{a}	0.0497^{a}	0.0386^{l}	0.0386^{l}	$0.0491^{\approx}$
(5000, 7500, 10000)	0.0480^{a}	$0.0451^{a,<}$	$0.0449^{a,<}$	0.0368^{l}	0.0369^{l}	$0.0446^{\approx}$
(10000, 10000, 10000)	0.0483^{a}	0.0483^{a}	0.0480^{a}	0.0392^{l}	0.0392^{l}	$0.0477^{\approx}$

($^{>}$ ($^{<}$) Empirical size is significantly larger (smaller) than 0.05; i.e. method is liberal (conservative)
(^{l} ($^{\approx}$) The proportion of agreement (not) significantly less than the minimum of the empirical sizes

paring Fig. 2 and Fig. 3, we observe that the Monte Carlo scheme described above generates distances that resemble LCDM distances for left VMPFC of HR subject 1. Therefore the distances generated in this fashion together with modification of some parameters such as $\nu_{p,i}$ would resemble the distances of VMPFCs from real subjects. That is, when such parameters are modified in the Monte Carlo scheme described above, the differences in the LCDM distances could simulate the morphometric differences between real subjects.

4.1 Simulation of Distances that Resemble LCDM Distances

In our Monte Carlo study, we generate three samples $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ with sizes n_x, n_y, and n_z, respectively, and set $n_x = n_y = n_z = 10000$. Each sample is generated similar to the procedure described above. For example, sample $\mathcal{X}$ is generated as follows: First we generate

$$N_X = \{J \sim P_X, J = 1, \ldots, n_x\}, \tag{7}$$

where $P_X(J = i) = \nu_i^x / \sum_{i=0}^{12} \nu_i^x$ with ν_i^x is the i^{th} entry in $\nu_x = (\nu_0^x, \nu_1^x, \ldots, \nu_{12}^x)$ and is also the i^{th} value after the entries $|\nu_i - \eta_x|$ are sorted in descending order for $i = 0, 1, 2, \ldots, 11$ and $\nu_{12}^x = 11659 - \sum_{i=0}^{11} |\nu_i - \eta_x|$. Let n_i^x be the frequency of i among the n_x generated numbers from P_X. Then we generate $U_{ik} \sim U(0, r_x)$ for $k = 1, \ldots, n_i^x$ for each i. Hence the set of simulated distances for set $\mathcal{X}$ is

$$D_{mc}^{\mathcal{X}} = \{(i + U_{ik})/2 : U_{ik} \sim \mathcal{U}(0, r_x) \text{ for } i = 0, 1, \ldots, 12 \text{ and } k = 1, \ldots, N_X\}. \tag{8}$$

Samples $\mathcal{Y}$ and $\mathcal{Z}$ are generated similarly with parameter subscripts in (7) and (8) are modified accordingly.

4.2 Empirical Size Estimates for the Multi-Sample Case

For the null hypothesis of multi-sample case which states the equality of the distributions of LCDM distances, we generate three samples $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ with the below parameters:

$$H_0 : r_x = r_y = r_z = 1.0 \quad \text{and} \quad \eta_x = \eta_y = \eta_z = 0. \tag{9}$$

Notice that each sample is generated so as to resemble those of the left VMPFC of HR subject 1 up to scale. This is done without loss of generality, since any other VMPFC can either be obtained by a rescaling of the generated distances or by modifying the parameters. So for example, for sample $\mathcal{X}$, $P_X(X_j = i) = \nu_{p,i}$ with $\nu_{p,i}$ being the i^{th} entry in ν_p in (5) and the set of simulated distances for set $\mathcal{X}$ is as in (8) with $r_x = 1.0$ and $\eta_x = 0$.

We repeat this sample generation procedure $N_{mc} = 10000$ times. We count the number of times the null hypothesis is rejected at $\alpha = 0.05$ level for K-W test of distributional equality and ANOVA F-tests (with and without HOV) of equality of mean distances. The ratio of the number of significant results by each test to N_{mc} yields the estimated significance levels under H_o. The estimated significance levels for various values of n_x, n_y, and n_z are provided in Table 5, where $\hat{\alpha}_{BF}$ is the empirical size estimate for K-W test, $\hat{\alpha}_{F_1}$ is for ANOVA F-test with HOV, and $\hat{\alpha}_{F_2}$ is for ANOVA F-test without HOV. Furthermore, $\hat{\alpha}_{KW,F_1}$ is the proportion of agreement between K-W and ANOVA F-test with HOV, i.e., the number of times out of 10000 Monte Carlo replicates both KW and ANOVA F-test with HOV reject the null hypothesis. Similarly, $\hat{\alpha}_{KW,F_2}$ is the proportion of agreement between K-W and ANOVA F-test without HOV, and $\hat{\alpha}_{F_1,F_2}$ is the proportion of agreement between ANOVA F-test with HOV and ANOVA F-test without HOV. Using the asymptotic normality of the proportions, we test the equality of the empirical size estimates with 0.05, and compare the empirical sizes pairwise. We observe that the K-W test is at the

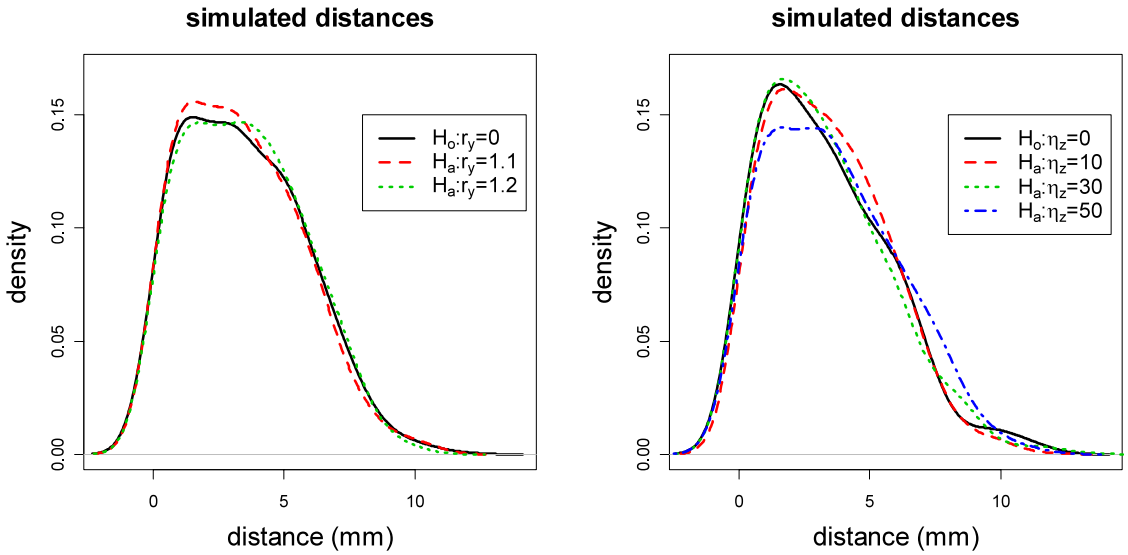

Fig. 4 Plots of the kernel density estimates of the Monte Carlo simulated LCDM distances under the null case and alternatives with $\eta_z = 0$ and $r_y \in \{1.1, 1.2\}$ (*left*); null case and alternatives with $r_y = 1.0$ and $\eta_z \in \{10, 30, 50\}$ (*right*). For the parameters r_y and η_z, see Sect. 4

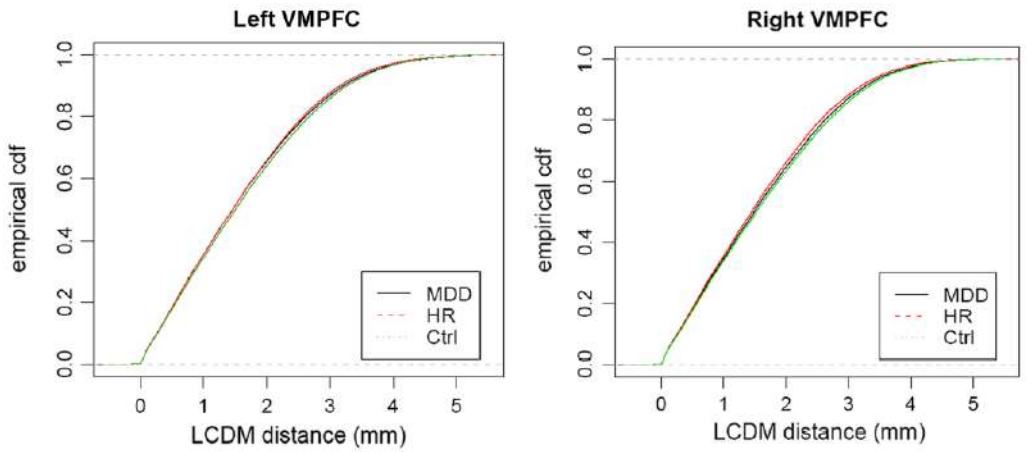

Fig. 5 Empirical cdfs of the pooled LCDM distances when extreme subjects are removed for the left and right VMPFCs

desired significance level, while ANOVA F-tests with and without HOV are at the desired level or slightly conservative. Notice also that under H_o, the tests tend to be more conservative as the sample sizes increase. Hence, if the distances are not that different; i.e., the frequency of distances for each bin and the distances for each bin are identically distributed for each group, the inherent spatial correlation does not seem to influence the significance levels. Moreover, we observe that for LCDM distances K-W and ANOVA with HOV tests have significantly different rejection (hence acceptance) regions, because the proportion of agreement for these tests, $\hat{\alpha}_{KW,F_1}$ is significantly smaller than the minimum of $\hat{\alpha}_{KW}$ and $\hat{\alpha}_{F_1}$, $\min(\hat{\alpha}_{KW}, \hat{\alpha}_{F_1})$. Similarly, K-W and ANOVA without HOV tests have significantly different rejection (hence acceptance) regions because, the proportion of agreement for these tests, $\hat{\alpha}_{KW,F_2}$ is significantly smaller than $\min(\hat{\alpha}_{KW}, \hat{\alpha}_{F_2})$. However, ANOVA with and without HOV tests have about the same rejection (hence acceptance) regions because, the proportion of agreement for these tests, $\hat{\alpha}_{F_1,F_2}$ is not significantly different from $\min(\hat{\alpha}_{F_1}, \hat{\alpha}_{F_2})$. This mainly results from the fact that K-W and ANOVA with HOV tests test different hypotheses, and so do the K-W and ANOVA without HOV tests. But, ANOVA with and without HOV tests basically test the same hypotheses.

4.3 Empirical Power Estimates for Multi-Sample Case

For the alternative hypothesis, we generate sample $\mathcal{X}$ as in the null case, so $D_{mc}^{\mathcal{X}}$ is as in (8). We consider various r_y and η_y values for sample $\mathcal{Y}$ and various r_z and η_z values for

Table 6 The power estimates based on Monte Carlo simulation of distances with three groups, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ with sizes n_x, n_y, and n_z, respectively, with $N_{mc} = 10000$ Monte Carlo replicates. $\hat{\beta}_{KW}$ is the empirical power estimate for K-W test, $\hat{\beta}_{F_1}$ and $\hat{\beta}_{F_2}$ are for ANOVA F-tests with and without HOV, respectively. The superscripts of the power estimates in the same row are labeled in increasing order of significance. That is, the power estimates with the same superscript are not significantly different from each other; while power estimate with label [a] is significantly smaller than the estimate labeled with [b], and so on

(n_x, n_y, n_z)	$\hat{\beta}_{KW}$	$\hat{\beta}_{F_1}$	$\hat{\beta}_{F_2}$
$(r_y, r_z) = (1.1, 1.0)$; $(\eta_y, \eta_z) = (0, 0)$			
(1000, 1000, 1000)	0.0778[a]	0.0770[a]	0.0768 [a]
(5000, 5000, 10000)	0.2281[a]	0.2137[b]	0.2114 [b]
(5000, 10000, 5000)	0.2936[a]	0.2731[b]	0.2745 [b]
(5000, 10000, 7500)	0.3244[a]	0.2939[b]	0.2947 [b]
(10000, 10000, 10000)	0.3900[a]	0.3564[b]	0.3559 [b]
$(r_y, r_z) = (1.1, 1.2)$; $(\eta_y, \eta_z) = (0, 0)$			
(1000, 1000, 1000)	0.1396[a]	0.1316[ab]	0.1313[b]
(5000, 5000, 10000)	0.6725[a]	0.6315[b]	0.6317[b]
(10000, 5000, 5000)	0.6651[a]	0.6262[b]	0.6253[b]
(5000, 10000, 5000)	0.5296[a]	0.4828[b]	0.4828[b]
(10000, 10000, 10000)	0.8410[a]	0.8050[b]	0.8050[b]
$(r_y, r_z) = (1.0, 1.0)$; $(\eta_y, \eta_z) = (10, 0)$			
(1000, 1000, 1000)	0.0574[b]	0.0728[a]	0.0721[a]
(5000, 5000, 10000)	0.0767[b]	0.1930[a]	0.1854[a]
(5000, 10000, 5000)	0.0884[b]	0.2341[a]	0.2381[a]
(5000, 7500, 10000)	0.0832[b]	0.2415[a]	0.2360[a]
(5000, 10000, 7500)	0.0878[b]	0.2571[a]	0.2584[a]
(10000, 10000, 10000)	0.1006[b]	0.3127[a]	0.3061[a]
$(r_y, r_z) = (1.0, 1.0)$; $(\eta_y, \eta_z) = (10, 30)$			
(1000, 1000, 1000)	0.0963[b]	0.1519[a]	0.1512[a]
(5000, 5000, 10000)	0.3986[b]	0.7436[a]	0.7537[a]
(10000, 5000, 5000)	0.3556[b]	0.7175[a]	0.7071[a]
(5000, 10000, 5000)	0.2908[b]	0.5826[a]	0.5831[a]
(5000, 7500, 10000)	0.4191[b]	0.7578[a]	0.7627[a]
(10000, 7500, 5000)	0.3652[b]	0.7229[a]	0.7147[a]
(10000, 5000, 7500)	0.4554[b]	0.8260[a]	0.8226[a]
(7500, 5000, 10000)	0.4739[b]	0.8331[a]	0.8363[a]
(7500, 10000, 5000)	0.3421[b]	0.6743[a]	0.6702[a]
(5000, 10000, 7500)	0.3752[b]	0.6938[a]	0.6983[a]
(10000, 10000, 10000)	0.5352[b]	0.8842[a]	0.8835[a]

sample $\mathcal{Z}$. The five alternative cases we consider are

$$(r_y, r_z, \eta_y, \eta_z) \in \{(1.1, 1.0, 0, 0), (1.1, 1.2, 0, 0), (1.0, 1.0, 10, 0), (1.0, 1.0, 10, 10), (1.0, 1.0, 10, 30)\}. \quad (10)$$

See Fig. 4 for the kernel density estimates of sample distances under the null case and various alternatives.

We repeat the sample generation $N_{mc} = 10000$ times under each alternative case. We count the number of times the null hypothesis is rejected at $\alpha = 0.05$ for K-W test of distributional equality, and ANOVA F-tests (with and without HOV) of equality of mean distances, and find the ratio of number of significant results by each test to N_{mc}. Thus we obtain the empirical power estimates under H_a which are provided in Table 6, where $\hat{\beta}_{KW}$ is the empirical power estimate for K-W test, $\hat{\beta}_{F_1}$ is for ANOVA F-test with HOV, and $\hat{\beta}_{F_2}$ is for ANOVA F-test without HOV. Using the asymptotic normality of the empirical power estimates, we observe that under each of H_a cases with $(r_y, r_z, \eta_y, \eta_z) \in \{(1.1, 1.0, 0, 0), (1.1, 1.2, 0, 0)\}$ the distributions are different, so the larger the r_y and r_z from 1.0, the higher the power estimates for K-W and ANOVA F-tests. Furthermore, as the sample size n increases, the power estimates for K-W

Table 7 Estimated significance levels based on Monte Carlo simulation of distances with two groups $\mathcal{X}$ and $\mathcal{Y}$ with sizes n_x and n_z, respectively, with $N_{mc} = 10000$ Monte Carlo replicates. $\hat{\alpha}_W$ is the empirical size estimate for Wilcoxon rank sum test, $\hat{\alpha}_t$ is for t-test, $\hat{\alpha}_{KS}$ is for K-S test; $\hat{\alpha}_{W,t}$, $\hat{\alpha}_{W,KS}$, and $\hat{\alpha}_{t,KS}$ are the values of proportion of agreement between the indicated tests in the subscripts. The superscript labeling for conservativeness and liberalness of empirical sizes and for proportions of agreement values are as in Table 5 and for ordering of the power estimates for each row is as in Table 6

(n_x, n_y)	Empirical size			Prop. of agreement		
	$\hat{\alpha}_W$	$\hat{\alpha}_t$	$\hat{\alpha}_{KS}$	$\hat{\alpha}_{W,t}$	$\hat{\alpha}_{W,KS}$	$\hat{\alpha}_{t,KS}$
Two-sided tests						
(1000, 1000)	0.0517^{a}	0.0505^{a}	0.0486^{a}	0.0403^{l}	0.0305^{l}	0.0273^{l}
(5000, 10000)	$0.0457^{b,<}$	$0.0463^{b,<}$	0.0465^{b}	0.0356^{l}	0.0273^{l}	0.0244^{l}
(7500, 10000)	0.0493^{a}	$0.0463^{a,<}$	0.0464^{a}	0.0385^{l}	0.0282^{l}	0.0246^{l}
(10000, 10000)	0.0518^{a}	0.0525^{a}	0.0501^{a}	0.0421^{l}	0.0320^{l}	0.0281^{l}
Left-sided tests (i.e., $\mathcal{X}$ values tend to be smaller than $\mathcal{Y}$ values)						
(1000, 1000)	0.0517^{a}	0.0527^{a}	0.0486^{a}	0.0440^{l}	0.0329^{l}	0.0305^{l}
(5000, 10000)	0.0470^{a}	0.0489^{a}	0.0492^{a}	0.0382^{l}	0.0311^{l}	0.0282^{l}
(7500, 10000)	0.0490^{a}	0.0493^{a}	0.0478^{a}	0.0399^{l}	0.0322^{l}	0.0284^{l}
(10000, 10000)	0.0517^{a}	0.0514^{a}	0.0494^{a}	0.0426^{l}	0.0330^{l}	0.0301^{l}
Right-sided tests (i.e., $\mathcal{X}$ values tend to be larger than $\mathcal{Y}$ values)						
(1000, 1000)	0.0521^{a}	0.0502^{a}	0.0491^{a}	0.0409^{l}	0.0337^{l}	0.0294^{l}
(5000, 10000)	0.0486^{a}	0.0502^{a}	0.0478^{a}	0.0405^{l}	0.0308^{l}	0.0285^{l}
(7500, 10000)	0.0479^{a}	0.0469^{a}	0.0495^{a}	0.0391^{l}	0.0325^{l}	0.0287^{l}
(10000, 10000)	0.0532^{a}	0.0517^{ab}	0.0469^{b}	0.0435^{l}	0.0354^{l}	0.0311^{l}

and ANOVA F-tests also increase. Notice that under these alternatives, the K-W test tends to be more powerful than ANOVA F-tests, since such alternatives influence the ranking (hence the distribution) of the distances, more than the mean of the distances. Furthermore, under these alternatives, it is not the size or scale that is really different; it is the difference in shape that is more emphasized. The size component is distance with respect to the GM/WM surface; i.e., if the GM voxels from the GM/WM surface are at about the same distance, the K-W test is more sensitive to the differences in the distributions of the LCDM distances. We also note that ANOVA F-tests with and without HOV have about the same power estimates.

Under each of alternative cases with

$$(r_y, r_z, \eta_y, \eta_z) \in \{(1.0, 1.0, 10, 0), (1.0, 1.0, 10, 10), (1.0, 1.0, 10, 30)\} \tag{11}$$

as η_y and η_z deviate more from 0, the power estimates for K-W and ANOVA F-tests increase. Note that as n increases, the power estimates also increase under these alternative cases. Under these second type of alternatives, ANOVA F-tests tend to be more powerful, since the right skewness (tail) of distances are more emphasized, which in turn implies that the differences in the mean distances are emphasized more. Under these alternatives, both the size or scale and shape are different. If the GM voxels from the GM/WM surface are at different distances, ANOVA F-tests are more sensitive to the differences in LCDM distances. We also note that both ANOVA F-tests (with and without HOV) have about the same power estimates.

4.4 Empirical Size Estimates for the Two-Sample Case

For the null hypothesis for the two-sample case, we generate two samples $\mathcal{X}$ and $\mathcal{Y}$ each of size n_x and n_y, respectively. Each sample is generated as described in Sect. 4.2. We repeat the sample generation $N_{mc} = 10000$ times.

We count the number of times the null hypothesis is rejected at $\alpha = 0.05$ for Lilliefor's test of normality, Wilcoxon rank sum test of distributional equality, t-test of equality of mean distances, and K-S test of equality of cdfs, and find the ratio of the number of significant results by each test to N_{mc}, thereby obtain the estimated significance levels. Unlike the multi-sample case, for the two-sample case, except for Lilliefor's test there are three types of alternative hypotheses possible: two-sided, left, and right-sided alternatives. The estimated significance levels are provided in Table 7, where $\hat{\alpha}_W$ is the empirical size estimate for Wilcoxon rank sum test, $\hat{\alpha}_t$ is for t-test, $\hat{\alpha}_{KS}$ is for K-S test. Furthermore, $\hat{\alpha}_{W,t}$ is the proportion of agreement between Wilcoxon rank sum and t-tests, $\hat{\alpha}_{W,KS}$ is the proportion of agreement between Wilcoxon rank sum and K-S tests, and $\hat{\alpha}_{t,KS}$ is the proportion of agreement between t-test and K-S test. We omit

the Lilliefor's test, since by construction, our samples are severely non-normal, so normality is rejected for virtually all samples generated. Observe that under H_o, the empirical significance levels are about the desired level for all three types of alternatives, although Wilcoxon tests are slightly liberal, while K-S test is slightly conservative. Hence, if the distances are not that different; i.e., the frequency of distances for each bin and the distances for each bin are identically distributed for each group, the inherent spatial correlation does not influence the significance levels. However, Wilcoxon rank sum, t-test, and K-S methods test different hypotheses, so their acceptance and rejection regions are significantly different for LCDM distances, since the proportion of agreement for each pair is significantly smaller than the minimum of the empirical size estimates for each pair of tests.

4.5 Empirical Power Estimates for the Two-Sample Case

For the alternative hypotheses, we generate samples $\mathcal{X}$ and $\mathcal{Y}$ as in Sect. 4.3 also. Note that when $r_y = 1$ and $\eta_y = 0$, we obtain the null case. The five alternative cases we consider are $(r_y, \eta_y) \in \{(1.1, 0), (1.2, 0), (1.0, 10), (1.0, 30), (1.0, 50)\}$. We count the number of times the null hypothesis is rejected for Lilliefor's test of normality, Wilcoxon rank sum test of distributional equality, t-test of equality of mean distances, and K-S test of equality of cdfs, thereby obtain the estimated significance levels as before. The power estimates are provided in Table 8, where $\hat{\beta}_W$ is the power estimate for Wilcoxon rank sum test, $\hat{\beta}_t$ is for t-test, $\hat{\beta}_{KS}$ is for K-S test.

Under the alternative cases with $(r_y, \eta_y) \in \{(1.1, 0), (1.2, 0)\}$, we see that the distributions start to differ. As r_y deviates further away from 1.0, then the power estimates for Wilcoxon rank sum, t-test, and K-S tests increase. Furthermore, as the sample size n increases, the power estimates for Wilcoxon test, t-test, and K-S test also increase. Observe that as in the multi-sample case, under these alternatives, Wilcoxon test is more powerful than t-test, since the ranking of the distances are affected more than the mean distances under these alternatives. But K-S test has the highest power estimates for sample sizes larger than 1000. Thus, for differences in shape rather than the distance from the GM/WM surface, K-S test and Wilcoxon rank sum test are more sensitive (i.e., powerful) than t-test. Furthermore, as the sample sizes increase, the left-sided tests become more powerful than their two-sided counterparts. Notice that we omit the power estimates for the right-sided alternatives, since by construction (i.e., due to our parameter choices in our simulations) $\mathcal{X}$ values tend to be smaller than $\mathcal{Y}$ values for these alternatives; hence the tests virtually have no power for the right-sided alternatives.

Under the H_a cases with $(r_y, \eta_y) \in \{(1.0, 10), (1.0, 30), (1.0, 50)\}$, as η_y deviates further away from 0, the power estimates for Wilcoxon rank sum, t-test, and K-S tests increase. Note that as n increases, the power estimates also increase under each alternative case. Under these alternatives, t-test is more powerful than Wilcoxon test, since mean distances are more affected than the rankings under such alternatives. However, K-S test has higher power estimates for larger deviations from the null case. These alternatives imply that the distances of the GM voxels are at different scales, t-test has the best performance for small differences, while for large differences, K-S has the best performance. Furthermore, as the sample sizes increase, the left-sided tests become more powerful than their two-sided counterparts. Again, we omit the power estimates for the right-sided alternatives, because, by construction, $\mathcal{X}$ values tend to be smaller than $\mathcal{Y}$ values for these alternatives.

We do not report the power estimates for Lilliefor's test of normality, since by construction our data is severely non-normal, and we get power estimates of 1.000 under both null and alternative cases.

5 Discussion and Conclusions

Pooled LCDM distances, when used as a single variable, provide a method to analyze heterogeneous forms of morphometric differences. When the LCDM distances of the subjects in the same diagnostic group are pooled, common morphometric traits of the ROI for that group are accentuated. Conversely, the morphometric traits not common for all the subjects in a group but specific to a particular subject are downplayed. The most common morphometric traits in a relevant ROI in a particular group may be associated with the diagnosis of the group and pooled LCDM distances carry on the most common characteristics, so they have the potential as demonstrated here to be very sensitive in detecting the diagnosis-specific traits of the ROI. As a result, they can indicate changes in the ROI highly associated with disease (major depression in the VMPFC in this article) or associated with being at genetic risk for the development of a specific condition. When pooled distances yield significant results, it implies that ROI significantly differ in morphometry (shape or size). However, it does not indicate the specific location within a ROI where such differences occur which might be important for understanding the underlying neurobiology. This may require the use of censoring which is the topic of another paper.

We use Kruskal-Wallis (K-W) and ANOVA F-tests (with or without HOV) for multi-group comparisons, Wilcoxon rank sum, Kolmogorov-Smirnov (K-S), and t-tests for two-group comparisons (the first two of these tests used to test distributional differences and the third is used to test mean differences due to a location parameter). But these tests require within sample independence which is violated due to

Table 8 The power estimates based on Monte Carlo simulation of distances with two groups, $\mathcal{X}$, and $\mathcal{Y}$, with sizes n_x, and n_y, respectively, with $N_{mc} = 10000$ Monte Carlo replicates. $\hat{\beta}_W$ is the power estimate for Wilcoxon rank sum test, $\hat{\beta}_t$ is for t-test, $\hat{\beta}_{KS}$ is for K-S test. The superscript labeling for ordering of the power estimates in each row is as in Table 6

(n_x, n_y)	Two-sided			Left-sided		
	$\hat{\beta}_W$	$\hat{\beta}_t$	$\hat{\beta}_{KS}$	$\hat{\beta}_W$	$\hat{\beta}_t$	$\hat{\beta}_{KS}$
$r_y = 10.1;\ \eta_y = 0$						
(1000, 1000)	0.1317^{a}	0.1264^{a}	0.0788^{b}	0.0742^{a}	0.0712^{a}	0.0750^{a}
(5000, 10000)	0.2723^{b}	0.2520^{c}	0.3734	0.3816^{b}	0.3600^{c}	0.5122^{a}
(10000, 5000)	0.2720^{b}	0.2507^{c}	0.3753	0.3838^{b}	0.3572^{c}	0.5157^{a}
(7500, 10000)	0.3242^{b}	0.3046^{c}	0.4731	0.4425^{b}	0.4178^{c}	0.6139^{a}
(10000, 7500)	0.3305^{b}	0.3100^{c}	0.4850	0.4455^{b}	0.4204^{c}	0.6253^{a}
(10000, 10000)	0.3662^{b}	0.3362^{c}	0.5504	0.4924^{b}	0.4588^{c}	0.6861^{a}
$r_y = 1.2;\ \eta_y = 0$						
(1000, 1000)	0.2635^{a}	0.2533^{a}	0.1838^{b}	0.1695^{b}	0.1630^{b}	0.1813^{a}
(5000, 10000)	0.7606^{b}	0.7331^{c}	0.9401^{a}	0.8463^{b}	0.8250^{c}	0.9755^{a}
(10000, 5000)	0.7588^{b}	0.7269^{c}	0.9421^{a}	0.8437^{b}	0.8178^{c}	0.9765^{a}
(7500, 10000)	0.8514^{b}	0.8282^{c}	0.9839^{a}	0.9121^{b}	0.8950^{c}	0.9945^{a}
(10000, 7500)	0.8561^{b}	0.8300^{c}	0.9845^{a}	0.9133^{a}	0.8969^{b}	0.8882^{c}
(10000, 10000)	0.8976^{b}	0.8750^{c}	0.9935^{a}	0.9468^{b}	0.9312^{c}	0.9982^{a}
$r_y = 1.0;\ \eta_y = 10$						
(1000, 1000)	0.0772^{c}	0.1173^{a}	0.0514^{d}	0.0506^{c}	0.0677^{b}	0.0477^{c}
(5000, 10000)	0.0871^{c}	0.2222^{b}	0.0673^{d}	0.1361^{c}	0.3297^{b}	0.1089^{d}
(10000, 5000)	0.0841^{c}	0.2186^{b}	0.0670^{d}	0.1390^{c}	0.3232^{b}	0.1076^{d}
(7500, 10000)	0.0951^{c}	0.2638^{b}	0.0737^{d}	0.1497^{c}	0.3786^{b}	0.1159^{d}
(10000, 7500)	0.0995^{c}	0.2630^{b}	0.0748^{d}	0.1560^{c}	0.3725^{b}	0.1161^{d}
(10000, 10000)	0.1018^{c}	0.2978^{b}	0.0743^{d}	0.1628^{c}	0.4132^{b}	0.1200^{d}
$r_y = 1.0;\ \eta_y = 30$						
(1000, 1000)	0.1760^{b}	0.2887^{a}	0.0878^{c}	0.1028^{c}	0.1885^{b}	0.0793^{d}
(5000, 10000)	0.4677^{d}	0.8254^{b}	0.7080^{c}	0.5927^{c}	0.8881^{b}	0.8911^{b}
(10000, 5000)	0.4668^{d}	0.8094^{b}	0.6901^{c}	0.5918^{d}	0.8807^{b}	0.8659^{c}
(7500, 10000)	0.5578^{d}	0.8987^{c}	0.9078^{b}	0.6773^{d}	0.9435^{c}	0.9792^{b}
(10000, 7500)	0.5509^{c}	0.8976^{b}	0.8983^{b}	0.6750^{d}	0.9438^{c}	0.9713^{b}
(10000, 10000)	0.6188^{d}	0.9369^{c}	0.9691^{b}	0.7339^{d}	0.9679^{c}	0.9942^{b}
$r_y = 1.0;\ \eta_y = 50$						
(1000, 1000)	0.3361^{c}	0.4865^{a}	0.2041^{d}	0.2266^{c}	0.3521^{b}	0.2048^{d}
(5000, 10000)	0.8876^{c}	0.9842^{b}	0.9980^{b}	0.9363^{c}	0.9936^{b}	0.9998^{a}
(10000, 5000)	0.8830^{c}	0.9844^{b}	0.9980^{a}	0.9325^{d}	0.9931^{c}	1.000^{a}
(7500, 10000)	0.9478^{c}	0.9964^{b}	1.000^{a}	0.9932^{c}	0.9986^{b}	1.000^{a}
(10000, 7500)	0.9473^{c}	0.9961^{b}	1.000^{a}	0.9741^{c}	0.9984^{b}	1.000^{a}
(10000, 10000)	0.9716^{c}	0.9984^{b}	1.000^{a}	0.9847^{c}	0.9995^{b}	1.000^{a}

the spatial correlation between LCDM distances of nearby voxels. Furthermore, parametric tests require normality of the samples also, which is again violated due to the heavy right skewness of the LCDM distances. However, our Monte Carlo analysis indicates that the influence of these violations is mild or negligible. Furthermore, the tests are more sensitive against different alternatives. In particular, K-W and Wilcoxon tests (i.e., the nonparametric tests) are more sensitive to distributional differences in a ROI with similar laminar thickness, while ANOVA F-tests and t-test (i.e., parametric tests) are more sensitive against the differences in the means, that is, differences in average GM thickness (i.e., laminar thickness values). On the other hand, K-S test is more sensitive to the largest difference in the cdfs of the LCDM distances.

Although the focus of this paper is the description of new morphologic image processing methods, as an illustrative example, we use GM tissue in the Ventral Medial Prefrontal Cortex (VMPFC) as the ROI for three groups of subjects; namely, subjects with major depressive disorder (MDD), subjects at high risk (HR) for MDD, and unrelated healthy control subjects (Ctrl). Based on previous re-

sults from other groups with older adult populations, we expected to find cortical differences associated with affective disorders in this region, however the nature of the changes or if they are present in younger populations has not been well characterized. Our study comprises of adolescent and young adult (MDD, HR) and (Ctrl, Ctrl) co-twin pairs. We found that gray matter distances in left and right VMPFC tend to decrease associated with MDD or being at HR for MDD, which is a characteristic that would be associated with cortical thinning. We thus observe a significant reduction in laminar thickness of VMPFC and perhaps shrinkage in MDD when compared to Ctrl subjects. However the same trend is also seen in the HR subjects, who are typical healthy individuals except for their genetic relation to the depressed cotwins. Thus this study does not support that all of the changes in morphometry of VMPFCs is related directly to major depression. It could be possible that VMPFC tend to shrink due to depression, but as similar shrinkage is seen in HR subjects, it could also be the case that specific genetic factors might predispose to this morphometric difference in VMPFC which in turn leads to vulnerability for developing depression in young individuals. Furthermore, in the pooled LCDM distance analysis, we find that the central values (i.e., means and medians) of the pooled distances in left VMPFCs of MDD and HR subjects are not significantly different, but the orderings of the central values of LCDM distances are MDD $<$ Ctrl and HR $<$ Ctrl; in right VMPFCs the ordering is as HR $<$ MDD $<$ Ctrl. Our findings here support that there are significant lateralization differences in the contribution of this region to affective disorders; similar asymmetry or lateralization findings have been previously reported in functional and structural studies [13, 18] and functional lateralization in this region has also been reported in animal models [9]. The cdf comparisons indicate that it is more likely for left VMPFCs of MDD or HR subjects to be thinner than those of Ctrl subjects which confirm the above findings about cortical thinning. However no such stochastic ordering occurs for the right VMPFCs, which only indicates the cdf orderings depend on the distance values in the right VMPFCs.

We demonstrate that pooled LCDM distances may provide a useful tool in detecting morphometric differences associated with specific disorders which affect the cortex. There is increasing recognition that different cortical features such as surface area or thickness may provide clues to different underlying pathology [6]. For instance increased GM distribution at shorter distances may represent increased surface area or increased curvature which could be further investigated via different methods. Attaining similar maximum long distances with a lower gray matter concentration at nearby long distances could indicate achievement of expected cortical thickness with loss of thickness in certain subregions within the ROI. Additional characterization of cortex may lead to improved sensitivity to detect differences associated with specific disorders. For example, the thickness at each point on the surface can be measured, which means mapping the surfaces to a template and then doing the statistics at each point on the surface. For this purpose, the first step is to apply LLCDM and then apply LDDMM-Surface (see [20, 21]). We also note that the LCDM based methodology used in this article can be applied to many different cortical regions.

Acknowledgements We would like to thank the editors and anonymous referees whose constructive remarks and suggestions greatly improved the presentation and flow of this article. Most of the Monte Carlo simulations presented in this article were executed at Koç University High Performance Computing Laboratory. Research supported by R01-MH62626-01, P41-RR15241.

References

1. Barker, A.R., Priebe, C.E., Miller, M.I., Hosakere, M., Lee, N., Ratnanather, J.T., Wang, L., Gado, M., Morris, J.C., Csernansky, J.C.: Statistical testing on labeled cortical distance maps to identify dementia progression. In: Joint Statistical Meeting, Section on Nonparametric Statistics. American Statistical Association, San Francisco (2003)
2. Botteron, K.N., Raichle, M.E., Drevets, W.C., Heath, A.C., Todd, R.D.: Volumetric reduction in the left subgenual prefrontal cortex in early onset depression. Biol. Psychiatry **51**(4), 342–344 (2002)
3. Bridge, H., Clare, S., Jenkinson, M., Jezzard, P., Parker, A.J., Matthews, P.M.: Independent anatomical and functional measures of the V1/V2 boundary in human visual cortex. J. Vis. **5**(2), 93–102 (2005)
4. Ceyhan, E., Hosakere, M., Nishino, T., Alexopoulos, J., Todd, R.D., Botteron, K.N., Miller, M.I., Ratnanather, J.T.: The use of labeled cortical distance maps for quantization and analysis of anatomical morphometry of brain tissues. Koç University, Istanbul (2008). Technical Report KU-EC-08-2: Available as arXiv:0805.3835v1 [stat.CO]
5. Ceyhan, E., Hosakere, M., Nishino, T., Babb, C., Todd, R.D., Ratnanather, J.T., Botteron, K.N.: Statistical analysis of morphometric measures based on labeled cortical distance maps. In: Fifth International Symposium on Image and Signal Processing and Analysis (ISPA 2007), Istanbul, Turkey (2007)
6. Chenn, A., Walsh, C.A.: Increased neuronal production, enlarged forebrains and cytoarchitectural distortions in beta-catenin overexpressing transgenic mice. Cereb. Cortex **13**, 599–606 (2003)
7. Chung, M.K., Robbins, S.M., Dalton, K.M., Davidson, R.J., Alexander, A.L., Evans, A.C.: Cortical thickness analysis in autism with heat kernel smoothing. NeuroImage **25**(4), 1256–1265 (2005)
8. Conover, W.: Practical Nonparametric Statistics, 3rd edn. Wiley, New York (1999)
9. Czéh, B., Müller-Keuker, J.I., Rygula, R., Abumaria, N., Hiemke, C., Domenici, E., Fuchs, E.: Chronic social stress inhibits cell proliferation in the adult medial prefrontal cortex: hemispheric asymmetry and reversal by fluoxetine treatment. Neuropsychopharmacology **32**(7), 1490–1503 (2007)
10. Drevets, W.C., Price, J.L., Simpson, J.R., Todd, R.D., Reich, T., Vannier, M., Raichle, M.E.: Subgenual prefrontal cortex abnormalities in mood disorders. Nature **386**, 824–827 (1997)
11. Elkis, H., Friedman, L., Buckley, P.F., Lee, H.S., Lys, C., Kaufman, B., Meltzer, H.Y.: Increased prefrontal sulcal prominence in

relatively young patients with unipolar major depression. Psychiatry Res. Neuroimaging **67**(2), 123–134 (1996)
12. Joshi, M., Cui, J., Doolittle, K., Joshi, S., Van Essen, D., Wang, L., Miller, M.I.: Brain segmentation and the generation of cortical surfaces. NeuroImage **9**(5), 461–476 (1999)
13. Killgore, W.D., Gruber, S.A., Yurgelun-Todd, D.A.: Depressed mood and lateralized prefrontal activity during a Stroop task in adolescent children. Neurosci. Lett. **416**(1), 43–48 (2007)
14. Makris, N., Biederman, J., Valera, E.M., Bush, G., Kaiser, J., Kennedy, D.N., Caviness, V.S., Faraone, S.V., Seidman, L.J.: Cortical thinning of the attention and executive function networks in adults with attention-deficit/hyperactivity disorder. Cerebral Cortex (2006)
15. Martinussen, M., Fischl, B., Larsson, H.B., Skranes, J., Kulseng, S., Vangberg, T.R., Vik, T., Brubakk, A.M., Haraldseth, O., Dale, A.M.: Cerebral cortex thickness in 15-year-old adolescents with low birth weight measured by an automated MRI-based method. Brain **128**(11), 2588–2596 (2005)
16. Miller, M.I., Hosakere, M., Barker, A.R., Priebe, C.E., Lee, N., Ratnanather, J.T., Wang, L., Gado, M., Morris, J.C., Csernansky, J.G.: Labeled cortical mantle distance maps of the cingulate quantify differences between dementia of the Alzheimer type and healthy aging. Proc. Natl. Acad. Sci. USA **100**(25), 15172–15177 (2003)
17. Miller, M.I., Massie, A.B., Ratnanather, J.T., Botteron, K.N., Csernansky, J.G.: Bayesian construction of geometrically based cortical thickness metrics. NeuroImage **12**(6), 676–687 (2000)
18. Phillips, M.L., Ladouceur, C.D., Drevets, W.C.: A neural model of voluntary and automatic emotion regulation: implications for understanding the pathophysiology and neurodevelopment of bipolar disorder. Mol. Psychiatry (2007)
19. Preul, C., Lohmann, G., Hund-Georgiadis, M., Guthke, T., von Cramon, D.Y.: Morphometry demonstrates loss of cortical thickness in cerebral microangiopathy. J. Neurol. **252**(4), 441–447 (2005)
20. Qiu, A., Vaillant, M., Barta, P., Ratnanather, J.T., Miller, M.I.: Region-of-interest-based analysis with application of cortical thickness variation of left planum temporale in schizophrenia and psychotic bipolar disorder. Hum. Brain Mapp. **29**(8), 973–985 (2008)
21. Qiu, A., Younes, L., Wang, L., Ratnanather, J.T., Gillepsie, S.K., Kaplan, G., Csernansky, J., Miller, M.I.: Combining anatomical manifold information via diffeomorphic metric mappings for studying cortical thinning of the cingulate gyrus in schizophrenia. NeuroImage **37**(3), 821–833 (2007)
22. Ratnanather, J.T., Barta, P.E., Honeycutt, N.A., Lee, N., Morris, H.M., Dziorny, A.C., Hurdal, M.K., Pearlson, G.D., Miller, M.I.: Dynamic programming generation of boundaries of local coordinatized submanifolds in the neocortex: application to the planum temporale. NeuroImage **20**(1), 359–377 (2003)
23. Ratnanather, J.T., Botteron, K.N., Nishino, T., Massie, A.B., Lal, R.M., Patel, S.G., Peddi, S., Todd, R.D., Miller, M.I.: Validating cortical surface analysis of medial prefrontal cortex. NeuroImage **14**(5), 1058–1069 (2001)
24. Ratnanather, J.T., Wang, L., Nebel, M.B., Hosakere, M., Han, X., Csernansky, J.G., Miller, M.I.: Validation of semiautomated methods for quantifying cingulate cortical metrics in schizophrenia. Psychiatry Res. Neuroimaging **132**(1), 53–68 (2004)
25. Thode Jr., H.: Testing for Normality. Marcel Dekker, New York (2002)
26. Wang, L., Hosakere, M., Trein, J.C., Miller, A., Ratnanather, J.T., Barch, D.M., Thompson, P.A., Qiu, A., Gado, M.H., Miller, M.I., Csernansky, J.G.: Abnormalities of cingulate gyrus neuroanatomy in schizophrenia. Schizophr. Res. **93**(1–3), 66–78 (2007)