ORIGINAL RESEARCH

Measuring Brain Volume by MR Imaging: Impact of Measurement Precision and Natural Variation on Sample Size Requirements

R.G. Steen
R.M. Hamer
J.A. Lieberman

BACKGROUND AND PURPOSE: To determine the sample size needed to provide adequate statistical power in studies of brain volume by MR imaging, we examined the precision and variability of measurements in healthy controls.

MATERIALS AND METHODS: A cohort of 52 people (mean age, 25.1 years) was examined at weeks 0 and 12 at 1.5T. We used an axial multisection T1-weighted sequence and a contiguous proton-attenuation/T2-weighted sequence. Data were registered to a probabilistic brain atlas, and an automated atlas-based program was used to segment brain tissue by type and by lobe. We assumed that there were no changes in volume because there were no intervening neurologic events. Sample sizes required to yield 80% statistical power in detecting a significant difference in volume were calculated for various experimental designs, assuming a patient-control volume difference of 5% or 2%.

RESULTS: The precision of most measurements was excellent, but required sample sizes were larger than anticipated. If the goal was to detect a 5% difference in whole brain volume in a 2-sample cross-sectional study, the required sample was 73 patients and 73 controls because brain volume varies between individuals in a way that is not informative about disease effects. For a similar 2-sample longitudinal study, the required sample size was just 5 patients and 5 controls.

CONCLUSIONS: Our results argue strongly for longitudinal studies in preference to cross-sectional studies, especially as research budgets decline. Our findings also suggest that there may be more uncertainty than expected in published MR imaging brain volume studies.

MR imaging makes it possible to visualize the human brain in vivo with exquisite detail and has been used extensively to examine patients with various brain illnesses, including schizophrenia (SZ). There is a large volume of literature to support the notion that there are characteristic brain structural abnormalities in patients with chronic SZ,[1] and a growing amount of literature to support that notion in patients with first-episode SZ.[2] However, experimental variance can be introduced into MR imaging studies[3] during data acquisition (eg, subject position, scanner field variation, image artifacts, scanner-to-scanner variation) or data analysis (eg, image registration, interpolation, bias field correction, manual interaction). These considerations call into question some of the conclusions that have been made about brain volume abnormalities in patients with SZ.[2]

We undertook a study of the precision of brain volume measurement by MR imaging in healthy controls to determine the sample size needed to provide adequate statistical power in future MR imaging studies of brain volume. We evaluated sample sizes required for studies of whole brain volume or volume of various smaller structures in the brain. The basic question is, "If we take 2 sets of measurements from a single subject (or from a single brain structure), do we obtain values that are similar enough that they can be used interchangeably?"

Received June 19, 2006; accepted after revision November 5.

From the Departments of Psychiatry (R.G.S., R.M.H., J.A.L.) and Biostatistics (R.M.H.), University of North Carolina at Chapel Hill, Chapel Hill, NC; and the Department of Psychiatry (J.A.L.), Columbia University, New York.

R.G. Steen was supported by the National Alliance for Research on Schizophrenia and Depression as a Hofmann Trust Investigator. Research was also supported by MH61603 (J.A. Lieberman), the University of North Carolina at Chapel Hill Schizophrenia Research Center, a National Institute of Mental Health Silvio Conte Center for the Neuroscience of Mental Disorders (MH64065), and the Foundation of Hope.

Please address correspondence to R. Grant Steen, PhD, Department of Psychiatry, Campus Box #7160, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7160; e-mail: Grant_Steen@med.unc.edu

DOI 10.3174/ajnr.A0537

Materials and Methods

Subjects

Data for this study were collected as part of a 2-year randomized double-blind clinical trial that compared the efficacy and safety of olanzapine with that of haloperidol in patients experiencing first-episode SZ.[4-6] Patient data from that trial were not used here; instead, we focused on brain volume data from 52 healthy controls, which were not reported in detail before.[6]

Healthy individuals, most of whom were university students, were recruited as patient controls by advertisement, and each person was seen in a face-to-face interview to screen for medical or psychiatric history findings of any kind, which were exclusionary. Controls were imaged at enrollment, then again 12 weeks later on the same scanners, by using the same imaging protocol described for patients.[6] We used data only from subjects who were imaged at both time points and who had no health complaints at either time point. Controls were a mean age of 25.1 $\pm$ 4.0 years at first scanning, with 67.3% being male, and the ethnic composition was 59.6% white, 28.9% African-American, and 11.5% other ethnicity.

For the purposes of this study, we assumed that there should be no changes in adult brain volume over a 12-week period in the absence of an intervening neurologic event, and we assumed that any such events would have been reported by controls or detected by clinicians.

Image Acquisition

Rigorous quality-control procedures were used to ensure that all images were acquired and analyzed by identical methods.[3] All MR im-

aging data were collected at 1.5T and analyzed blind as to group membership.[6] A scout sequence was run on each subject to help in section positioning; then T1-weighted and T2-weighted image sets were acquired from each subject in the axial plane. A 3D T1-weighted inversion-recovery prepared spoiled gradient-recalled acquisition in steady state was acquired (TR = 12.3 msec, TE = 5.4 msec, flip angle = 20°, section thickness = 1.5 mm, FOV = 24 cm, matrix = 256 × 256, 124 sections). Then a contiguous proton-attenuation/T2-weighted fast spin-echo sequence was acquired (TR = 4000 msec, TE = 15 and 105 msec, flip angle = 90°, section thickness = 3.0 mm, FOV = 24 cm; matrix = 256 × 256, 60 sections). Parameters were optimized to show gray matter (GM) and white matter (WM) with good contrast and to yield reproducible segmentation with a fully automated program.

Image Processing

All patient and control images were centrally analyzed in a multistep process designed to minimize operator interaction.[6] Processing included a bias-field correction step to adjust for intensity inhomogeneities in the images. Baseline T1-weighted data were registered to a probabilistic brain atlas, so that all brains could be analyzed and displayed in a standard coordinate system. Then T2-weighted data were registered to the T1-weighted data within the segmentation algorithm. These images formed the basis for a 3-channel segmentation, which used an automatic atlas-based segmentation program (expectation maximization segmentation) to separate brain tissue into GM, WM, and CSF.[7]

The probabilistic brain atlas driving the tissue segmentation also provided a Talairach-based parcellation, dividing the left and right hemispheres, which coarsely represented the frontal, temporal, parietal, and occipital lobes.[6] Atlas registration overlaid these representations onto each scan, thereby creating a fully automatic parcellation for each dataset. Caudate volume was obtained by manual outlining of the caudate head, after rigorous operator training and standardization of methods.[6] Most of the tools used were fully automatic (atlas registration, interscan registration, tissue segmentation, parcellation), which made these procedures robust against rater drift.

Data Analysis

All data from healthy control subjects imaged at both week 0 (baseline) and week 12 were analyzed. We did not attempt to evaluate longer follow-up data because the healthy brain can potentially change over long follow-up intervals[8,9] and because we wanted to characterize measurement precision in the absence of biologic change. Statistical analysis was done by using SAS System software (Version 9.1 TS1M2; SAS Institute, Cary, NC) to compare each subject at week 12 with the same subject at week 0, by using no covariates in the analysis.

To determine the sample size required for studies of brain volume, we made several key assumptions. First, we assumed that the minimal acceptable level of power was 80%. Second, we assumed that there were no biologic differences in controls between week 0 and week 12, so that volumetric changes over this time period must have been due only to error or random variation. Finally, we assumed that patients would differ in brain volume from controls by a small amount, either 5% or 2% in different simulations. Then we used the measured variance at week 0 and week 12 and the variance in the change scores between week 0 and week 12 to calculate the sample sizes necessary to detect both a 5% and a 2% change in the week 12 values for several different study designs.

Power and sample size calculations were performed by using PROC POWER in SAS, Version 9.1. For the cross-sectional 2-group study design, calculations were based on a 2-sample *t* test on means (cross-sectional). For the longitudinal 1-group study, calculations were based on a 1-sample *t* test. For the longitudinal 2-group (change) study, calculations were based on a 2-sample *t* test on mean change scores. All tests were 2-tailed, and the 2-sample tests assumed equal variance in both samples. Mean brain volume, mean change, mean difference, and difference in mean change were all calculated as both 5% and 2% of the baseline values or as 5% and 2% differences between groups for the cross-sectional comparisons. The computer program used noncentral *t* distributions based on hypothesized effect sizes to estimate power and sample size required for 80% statistical power.

Results

Comparison of Baseline to Follow-Up Data

The mean difference between week 0 and week 12 was generally quite small. For whole brain volume, there was only a 2.2 cm^3 (mL) discrepancy in mean volume between week 0 and week 12 (Table 1). This represented a 0.22% difference, and both the Pearson correlation coefficient and the concordance value were 0.98. We note that because the image parcellation method was fully automated, analyzing the same dataset twice would have produced identical results each time (concordance value = 1.00).

For left frontal WM, there was a 0.6-mL mean difference in volume between baseline and week 12, which represented a 0.58% difference, and both the Pearson and the concordance values were approximately 0.93. Even for the caudate, which shows the largest proportional difference between week 0 and week 12, there was only a 0.8% difference, and both the Pearson and concordance values showed a strong correlation. Data in Table 1 suggest that precision in this study was excellent overall and comparable with other published brain volumetric studies.[2]

Analysis of Error

To characterize error that might corrupt MR imaging findings, we did an analysis of the absolute differences between week 0 and week 12 (Table 2). The mean volume difference between week 0 and week 12 was generally quite small (Table 1), but this could have been an artifact. If all differences are random, then one would expect some volumes to increase and others to decrease, so that the net result could be zero because volume increases are offset by volume decreases. However, even if random variations in individual measurements average to a small value, there could still be a substantial reduction in the interchangeability of data between week 0 and week 12. To evaluate this possibility, we calculated the absolute magnitude of the difference between week 0 and week 12. In whole brain, the absolute magnitude of change was roughly 8-fold larger than the mean change, or approximately 2%. The greatest single brain volume decrease was 68.3 mL or −5.7%, whereas the greatest single volume increase was 77.3 mL or +6.5%. Such changes are clearly not consistent with the small changes expected in the volume of an adult brain over 12 weeks.

Table 1: Volume (milliliters) of whole brain and brain lobes in 52 control subjects, measured at baseline and again at 12 weeks postbaseline*

	Baseline		Week 12		Mean Volume Difference		Percent Difference	Pearson	
Structure	Mean	SD	Mean	SD	Mean	SD	Mean	Correlation	Concord
Whole brain	1190.4	127.0	1192.6	126.9	2.2	25.2	0.22	0.98	0.98
L frontal GM	157.2	16.8	158.6	14.8	1.4	4.7	1.08	0.96	0.95
R frontal GM	160.5	17.7	161.3	16.5	0.8	4.1	0.65	0.97	0.97
L occipital GM	59.3	6.9	59.1	6.4	−0.2	2.1	−0.20	0.95	0.95
R occipital GM	59.4	6.8	59.4	6.3	0.0	2.2	0.11	0.95	0.95
L parietal GM	62.7	7.1	62.4	6.8	−0.3	2.0	−0.34	0.96	0.96
R parietal GM	63.7	7.4	63.6	7.0	−0.1	2.5	−0.01	0.94	0.94
L temporal GM	67.8	8.1	68.0	7.4	0.2	3.0	0.59	0.93	0.92
R temporal GM	68.6	8.7	68.8	7.5	0.2	3.3	0.72	0.93	0.91
L frontal WM	141.9	14.5	142.4	13.4	0.6	5.3	0.58	0.93	0.93
R frontal WM	142.1	16.0	142.3	14.6	0.2	5.3	0.37	0.94	0.94
L occipital WM	17.1	2.9	16.9	2.9	−0.2	1.4	−0.64	0.89	0.88
R occipital WM	17.5	2.6	17.3	2.5	−0.2	1.3	−0.77	0.88	0.88
L parietal WM	46.1	5.7	45.6	5.6	−0.4	1.7	−0.87	0.96	0.95
R parietal WM	46.0	5.0	45.6	4.8	−0.3	1.8	−0.62	0.93	0.93
L temporal WM	36.0	4.9	36.0	4.1	0.0	2.1	0.55	0.90	0.89
R temporal WM	36.1	5.1	36.1	4.3	0.1	2.1	0.71	0.91	0.90
L caudate	4.3	0.6	4.3	0.6	0.0	0.5	1.46	0.58	0.58
R caudate	4.8	0.7	4.7	0.7	0.0	0.6	−0.21	0.57	0.56

Note:—Concord indicates concordance; L, left; R, right; GM, gray matter; WM, white matter.
* None of the mean volume differences were significant by paired-sample *t* test, and the Pearson correlation between week 0 and week 12 was generally quite high. Concordance is also a measure of the degree to which values at week 0 and week 12 are correlated, unlike the Pearson, which does not take account of the sample mean, concordance is sensitive to changes in sample mean.

Table 2: Brain volume (milliliters) differences between values at week 0 and week 12, expressed in several ways*

	Absolute Difference		Max Volume Change		Max Percent Change	
Structure	Mean	SD	Decrease	Increase	Decrease	Increase
Whole brain	17.9	17.7	−68.3	77.3	−5.7	6.5
L frontal GM	3.3	3.6	−6.7	19.8	−4.3	12.6
R frontal GM	3.0	2.9	−6.3	17.8	−3.9	11.1
L occipital GM	1.6	1.4	−5.2	7.4	−8.8	12.5
R occipital GM	1.4	1.6	−5.1	6.8	−8.6	11.4
L parietal GM	1.6	1.2	−5.9	4.8	−9.4	7.7
R parietal GM	1.7	1.8	−7.1	8.5	−11.1	13.3
L temporal GM	2.2	2.1	−6.0	9.4	−8.9	13.9
R temporal GM	2.4	2.3	−7.7	10.7	−11.2	15.6
L frontal WM	3.8	3.7	−9.7	21.3	−6.8	15.0
R frontal WM	3.9	3.6	−11.1	21.1	−7.8	14.8
L occipital WM	1.0	1.0	−3.4	3.6	−19.9	21.1
R occipital WM	1.0	0.8	−2.9	3.2	−16.6	18.3
L parietal WM	1.3	1.2	−7.1	3.6	−15.4	7.8
R parietal WM	1.5	1.1	−5.1	3.5	−11.1	7.6
L temporal WM	1.5	1.5	−6.5	5.7	−18.1	15.8
R temporal WM	1.6	1.4	−5.0	6.5	−13.9	18.0
L caudate	0.4	0.3	−1.2	2.1	−27.9	48.8
R caudate	0.4	0.4	−2.1	1.8	−43.8	37.5

Note:—L, left; R, right; Max, maximum; GM, gray matter; WM, white matter.
* The absolute difference is unaffected by the sign of the change from week 0 to week 12. The maximal decrease and increase from week 0 to week 12 is expressed as volume (with units of milliliters) and as percent change (with respect to baseline).

Table 3: Minimal sample size required to detect a 5% change in volume with at least 80% statistical power, under a variety of assumptions about study design*

	Cross-Sectional		Longitudinal		Longitudinal	
	2 Groups		1 Group		2 Groups (change)	
Structure	No.	Power	No.	Power	No.	Power
Whole brain	146	80.3	4	86.7	10	90.2
L frontal GM	146	80.2	5	80.0	14	82.3
R frontal GM	156	80.4	5	90.0	12	86.5
L occipital GM	170	80.1	7	86.0	20	83.3
R occipital GM	166	80.4	7	85.6	20	82.9
L parietal GM	166	80.3	6	87.1	16	84.0
R parietal GM	172	80.4	7	80.3	22	81.3
L temporal GM	184	80.4	9	83.2	28	81.0
R temporal GM	204	80.1	10	82.2	32	80.1
L frontal WM	134	80.2	7	83.4	20	80.4
R frontal WM	162	80.2	7	83.8	20	80.9
L occipital WM	362	80.1	23	81.1	84	80.2
R occipital WM	280	80.3	19	81.4	68	80.3
L parietal WM	192	80.0	7	85.0	20	82.2
R parietal WM	150	80.1	8	86.2	22	80.3
L temporal WM	234	80.1	13	80.7	46	80.9
R temporal WM	256	80.3	13	80.4	46	80.5
L caudate	210	80.1	47	80.5	180	80.0
R caudate	244	80.3	54	80.4	208	80.0

Note:—L, left; R, right; GM, gray matter; WM, white matter.
* Estimated statistical power is shown for a sample size that was calculated to yield at least 80% power.

In the caudate, the largest volume decrease was 44%, whereas the largest volume increase was 49%. Because the caudate is rather small and its margins can be hard to define, it should not be surprising that it is measured with less reliability than whole brain.

Sample Size Estimates

The sample size required to detect a 5% difference in the week 12 values is shown in Table 3 for several different study designs. The most common study design is to compare patients with controls at a single time point,[2] to measure volume differences at baseline. Even in the whole brain, where volume measurements are made with the greatest precision and where artifacts arising from parcellation cannot be a factor, the required sample size for a cross-sectional study design is 146 subjects, equally apportioned between patients and controls. If

Table 4: Minimal sample size required to detect a 2% change in volume with at least 80% statistical power, under a variety of assumptions about study design*

	Cross-sectional		Longitudinal		Longitudinal	
	2 groups		1 group		2 Groups (change)	
Structure	No.	Power	No.	Power	No.	Power
Whole brain	896	80.0	11	80.5	38	80.8
L frontal GM	898	80.0	20	81.5	72	80.4
R frontal GM	956	80.0	15	80.8	54	80.9
L occipital GM	1050	80.0	28	80.6	106	80.6
R occipital GM	1016	80.0	28	80.2	106	80.2
L parietal GM	1022	80.1	22	81.2	80	80.2
R parietal GM	1056	80.1	33	81.2	124	80.6
L temporal GM	1130	80.1	42	80.5	162	80.5
R temporal GM	1262	80.0	49	80.3	190	80.3
L frontal WM	824	80.1	30	80.5	114	80.5
R frontal WM	998	80.0	30	81.0	112	80.3
L occipital WM	2246	80.0	130	80.1	514	80.1
R occipital WM	1730	80.0	105	80.3	412	80.1
L parietal WM	1190	80.1	29	80.9	108	80.2
R parietal WM	926	80.0	33	80.2	126	80.2
L temporal WM	1448	80.0	70	80.5	272	80.2
R temporal WM	1580	80.0	70	80.2	274	80.1
L caudate	1298	80.0	280	80.0	1116	80.1
R caudate	1506	80.0	324	80.1	1290	80.1

Note:—L, left; R, right; GM, gray matter; WM, white matter.
* Estimated statistical power is shown for a sample size that was calculated to yield at least 80% power.

the study aim is to detect a 5% change in a single group with time so that each subject acts as his or her own control, the required sample size is only 4 subjects. If the study aim is to detect a 5% change in 1 group, in contrast to no change in a second group, the required sample size is 10 subjects, equally apportioned between the groups.

For frontal WM, if the study aim is to detect a 5% change in a single group with time, the required sample size is only 7 subjects, all in the same group (Table 3). If the study aim is to detect a 5% change in 1 group, in contrast to no significant change in a second group, the required sample size is 20 subjects, equally apportioned between the 2 groups. Yet, if the goal is to detect a 5% difference between 2 groups, a total sample size of 134–162 subjects is required, equally apportioned between patients and controls.

For the caudate, which is measured with much less precision, the required sample sizes are accordingly larger (Table 3). To detect a 5% change with time in caudate volume in a single group requires a sample size of 47–54 subjects, but to detect a 5% difference in change rate between 2 groups requires a sample size of 180–208 subjects. Finally, to have 80% power to detect a 5% difference in caudate volume between 2 groups at baseline requires a sample of 210–244 subjects.

We also calculated the sample size required to detect a 2% difference between patients and controls for several study designs (Table 4). In the whole brain, if the study aim is to detect a 2% change in a single group with time so that each subject can act as his or her own control, the required sample size is 11 subjects. If the study aim is to detect a 2% change in 1 group, in contrast to no significant change in a second group, the required sample size is 38 subjects, apportioned equally between the 2 groups. However, the required sample size for a cross-sectional study with 80% power to detect a 2% difference between patients and controls is 896 subjects overall.

Discussion

Our results suggest that the sample sizes necessary to obtain 80% statistical power to detect a 5% difference in brain volume are considerably larger than anticipated (Table 3), even though the precision of most measurements was quite good (Table 1). The sample sizes required to detect a significant difference are correspondingly larger if an assumption is made that there is actually a 2% difference in brain volume between patients and controls (Table 4). Our findings suggest that there may be more uncertainty than expected in brain volumetric findings.

In considering whether 2 sets of measurements are equivalent, a critical consideration is the intended purpose of the comparison. If researchers are probing for large differences in a cross-sectional study or for large changes in a longitudinal study, then the strict equivalence of 2 measurements is a less important issue. If the effect size sought is large, then measurement precision need not be great to detect such a difference. Conversely, if effect sizes are small, as they are likely to be in most brain imaging studies, then one needs very precise measurements to detect a difference.

There are several ways to determine the interchangeability of 2 sets of measurements that have a continuous distribution. Traditional psychometric testing uses interclass correlations (eg, Pearson correlations), which arise from test theory. High correlation values are required for a measurement to have adequate validity; in test theory, reliability is an upper bound on validity because a test cannot be more valid than it is reliable. However, concordance is more appropriate than the common Pearson product moment correlation for assessing the interchangeability of scores because concordance is sensitive to differences in sample means as well as to the linear relationship between 2 sets of scores. For example, if 2 sets of brain volume measurements were available and all volumes in 1 set were exactly 5-fold larger than in the other set, the Pearson correlation would be 1.00, whereas the concordance correlation would be much less than 1.00, showing that the 2 datasets are not interchangeable. In our analysis, we did not test for statistical significance of the correlation between week 0 and week 12; although this value would have been significant, it would not have been meaningful because the usual significance test for a correlation tests a null hypothesis that there is zero correlation between 2 measurements. Such a null hypothesis is not appropriate in this study; if 2 datasets are to be used interchangeably, they must have a correlation close to unity.

The concordance correlations reported here (Table 1) are generally quite high. Concordance for the whole brain is 0.98, whereas the concordance for GM averages 0.94 ± 0.02 and for WM, averages 0.91 ± 0.03. Certain structures have lower concordance correlations (eg, caudate = 0.56), showing that there is more imprecision in the measurement of small structures or structures with indefinite boundaries (such as the head of the caudate).

Spatial resolution of the imaging method was rather low (individual voxels were $0.9 \times 0.9 \times 1.5$ mm in the T1-weighted sequence and $0.9 \times 0.9 \times 3.0$ mm in the T2-weighted sequence), so it is possible that some voxels con-

tained a mixture of GM and WM. A "mixed" voxel would be classified as either GM or WM, depending on the exact proportion of each tissue in the voxel, as well as on a host of other factors, [2] so problems in segmentation could account for some of the variation that we report. However, given the nature of our dataset, we cannot calculate exactly how much experimental variance was due to errors in data acquisition (eg, subject position, scanner-field variation, image artifacts, scanner-to-scanner variation) and how much was due to errors in data analysis (eg, image registration, interpolation, bias field correction, manual interaction).

An unexpected finding is that the sample size required to evaluate whole brain volume differences between patients and controls is substantial, even though large structures can be measured with a great deal of precision (Tables 3 and 4). This is because total brain volume varies substantially from 1 person to another in a way that is probably not informative about health or disease effects. In our study, the smallest brain volume was 783 mL, whereas the largest brain volume was 1414 mL. Therefore, the largest brain was 81% bigger than the smallest brain, even though all of our subjects were healthy, of normal intelligence, and functioning at a high level.

Perhaps it should not be surprising that brain size varies in ways that are not biologically informative. Men have brains that are, on average, ~9% larger than those of women, after controlling for all known covariates including body size.[10,11] GM volume is significantly correlated with verbal intelligence quotient (IQ), performance IQ, and full-scale IQ, yet only 12%–31% of the total variance in IQ can be explained by GM volume.[12] Patients with SZ have brains that are, on average, only 2% lighter in weight than age- and sex-matched controls ($P < .04$), but the effects of both age and sex are far more significant ($P < .0001$) than the effect of disease.[13]

The sample size required to characterize small structures such as frontal GM is substantially larger than the sample size required to characterize whole brain volume (Tables 3 and 4). This is presumably because of greater imprecision in the measurement of small-volume structures. Yet the volume of frontal GM also has a certain amount of natural person-to-person variation that is unrelated to disease effects, as is true of the whole brain. The difference in sample size required for longitudinal-versus-cross-sectional studies gives an indication of how sample size is influenced both by measurement precision and by natural variation. If a subject is compared with himself, as in a longitudinal study, then measurement precision is the only factor that can affect the sample size. If a subject is compared with another subject, as in a cross-sectional study, then both measurement precision and natural variation affect the sample size. When a 5% difference between groups is anticipated (Table 3), a cross-sectional 2-group comparison of GM volume requires, on average, 170.5 subjects, whereas a longitudinal 2-group comparison requires, on average, 20.5 subjects. When subtle differences are anticipated between patients and controls (Table 4), a cross-sectional 2-group comparison of GM volume requires, on average, 1048.8 subjects, whereas a longitudinal 2-group comparison requires an average of 111.8 subjects. Thus, a cross-sectional study of GM volume requires 8- to 9-fold more subjects than a longitudinal study, largely because of person-to-person variation. Yet such variation in GM volume may have no clinical significance.

One approach to compensating for individual variation in volume of brain structures is to normalize brain structure volume measurements to the total intracranial volume of each subject.[14] This might make it easier to compare hippocampal volume between subject groups, but this approach has several drawbacks. If hippocampal volume is normalized to intracranial volume, a ratio is formed that should not be analyzed with the same statistical tests that are used for raw (uncorrected) values. Furthermore, this approach will tend to minimize volumetric changes that are likely to be small anyway, thereby making it harder to achieve statistical significance. For example, if hippocampal volume is 7 mL total and the brain volume is 1400 mL, then the ratio of hippocampal volume to brain volume is only 0.005 or 0.5%. Such numbers are more difficult to use in statistical tests than the raw value of 7 mL would be. A stronger experimental design is to match patients and controls for total intracranial volume, but this is not always possible.

What implications do our findings have for the earlier study[6] of brain volume change in patients with SZ receiving olanzapine or haloperidol? That study was a longitudinal analysis of 2 groups, which used change scores as an end point, so a total of only 10 subjects would be required for an evaluation of changes in whole brain volume (Table 3). That study actually included 164 patients,[6] evenly allocated between 2 treatment groups, so there was more than adequate power to detect a 5% change in total brain volume between groups. In fact, power was even adequate to detect a 2% change in total brain volume (Table 4).

However, most studies of brain volume in SZ are cross-sectional, not longitudinal,[2] and our findings could have implications for any such cross-sectional studies. In a recent review of 180 cross-sectional studies of patients with chronic SZ,[1] only 11 studies apportioned at least 146 subjects between 2 study groups. In a meta-analysis of 47 cross-sectional studies of patients with first-episode SZ,[2] no studies apportioned at least 146 subjects between 2 study groups. Thus, much of what we think we know about how SZ affects brain volume is open to question.

It is especially problematic that some of the brain volume changes that have been described, especially in patients with first-episode SZ, involve volume deficits smaller than the 5% difference that we assumed here. For example, a meta-analysis of whole brain volume deficit in patients with first-episode SZ, which included 524 patients and 650 healthy controls in a cross-sectional design, concluded that the first-episode patient brain is only 2.7% smaller than the control brain.[2] This finding agrees well with the finding that brain weight is 2% less in patients with SZ.[13] Yet, if the difference in brain volume between first-episode patients and controls is actually 2%–3%, then none of the contributing studies in the meta-analysis were adequately powered to detect such a small difference.

We note that our results are directly relevant only to studies that use a pixel-count volumetric method to measure brain volume, whereas an alternative approach to characterizing brain volume is provided by voxel-based morphometry

(VBM). A direct comparison of a volumetric method with VBM showed that VBM could detect significant hippocampal atrophy in a longitudinal study of patients with Alzheimer disease, whereas a volumetric method could identify no significant change.[15] This study was a longitudinal evaluation of patients and controls who were matched for volume of the hippocampus, so the study design was optimized to characterize hippocampal atrophy by VBM, even with a small sample size. In the absence of volumetric matching, natural variability in brain volume will generally force VBM studies to have a large sample size as well, even if VBM is more precise than volumetric methods.

A potential limitation of our study is that we have assumed that volume changes in the adult brain over a period of 12 weeks are due to imprecision of the MR imaging method. However, there are some studies suggesting that human brain volume, measured by MR imaging, can change rather rapidly. The average large-vessel ischemic stroke volume is 54 mL, and this volume of tissue is lost during just 10 hours of stroke evolution.[16] Lack of fluid intake for 16 hours decreases brain volume by 0.55% (or roughly 7 mL), and rehydration can increase total cerebral volume by 0.72%. [17] Acute brain volume changes have also been described in healthy people dehydrated as a result of airplane travel,[18] in young patients having prolonged febrile seizure,[19] in adults recovering from an eating disorder,[20] in patients with bipolar disorder receiving lithium,[21] in patients with multiple sclerosis either left untreated[22] or treated with methylprednisolone,[23] in patients with obsessive-compulsive disorder given paroxetine,[24] in short-stature youth receiving growth hormone therapy,[25] in patients with renal failure who got hemodialysis,[26] and in patients with SZ treated with haloperidol.[6] Among patients with SZ, acute increases in whole brain volume are associated with exacerbation of psychosis, whereas acute decreases in volume are linked to symptom remission.[27] Among alcohol-dependent men, there can be acute changes in brain volume during alcohol withdrawal,[28] and WM volume is correlated with blood hematocrit.[29] In a small study of alcohol-dependent men imaged before and after 1 month of abstinence, total intracranial volume was reported to vary by only 0.4%, but WM volume increased by an average of 10.3%.[30] However, in all of these previous reports, subjects showing an acute brain volume change were demonstrably not healthy before treatment, whereas the subjects in the present study were all well at baseline and well at follow-up. Even if several of our subjects had health issues that were not detected, our sample of subjects was still characterized by remarkably good health overall.

One potential way to compensate for the experimental imprecision that we demonstrate is to model brain volume by using a more sophisticated method. Currently, each individual brain region or tissue type is typically analyzed as if it were changing independently of all other tissue types. Clearly, if a certain voxel is segmented as GM at baseline and WM at follow-up, then there will be intercorrelated changes in volume of both GM and WM. A statistical method should be developed, on the basis of correlated volumetric changes, that would acknowledge that changes in 1 tissue compartment can be offset by changes in another tissue compartment.

Our main finding is that natural variation among very healthy people can swamp all but the largest experimental or disease effects. Our results argue in strong terms for the utility of longitudinal studies in preference to cross-sectional studies, especially as research budgets decline. If one considers only WM tissues in which a 5% change in volume is expected, the average sample size required for a cross-sectional study is 221 subjects, whereas the average sample size required for a longitudinal study is 41 subjects. Even if one allows that a longitudinal study requires that each subject be imaged twice, a cross-sectional study would require approximately 2.7-fold more images to be acquired than a longitudinal study and would be correspondingly more expensive for a comparable level of statistical power.

References

1. Shenton ME, Dickey CC, Frumin M, et al. **A review of MRI findings in schizophrenia.** *Schizophr Res* 2001;49:1–52
2. Steen, RG, Mull C, McClure R, et al. **Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies.** *Br J Psychiatry* 2006;188:510–18
3. Styner M, Charles HC, Park J, Gerig G. **Multisite validation of image analysis methods: assessing intra- and intersite variability.** In: Sonka M, Fitzpatrick JM, eds. *Medical Imaging 2002: Image Processing.* Proc. SPIE Vol. 4684; 2002:278–86
4. Lieberman JA, Tollefson G, Tohen M. **Comparative efficacy and safety of atypical and conventional antipsychotic drugs in first-episode psychosis: a randomized, double-blind trial of olanzapine versus haloperidol.** *Am J Psychiatry* 2003;160:1396–404
5. Perkins DO, Lieberman JA, Gu H, et al. **Predictors of antipsychotic treatment response in patients with first-episode schizophrenia, schizoaffective, and schizophreniform disorders.** *Br J Psychiatry* 2004;185:18–24
6. Lieberman JA, Tollefson GD, Charles C, et al. **Antipsychotic drug effects on brain morphology in first-episode patients.** *Arch Gen Psychiatry* 2005;62:361–70
7. Van Leemput K, Maes F, Vandermeulen D, et al. **Automated model-based tissue classification of MR images of the brain.** *IEEE Trans Med Imaging* 1999;18:897–908
8. Benedetti B, Charil A, Rovaris M, et al. **Influence of aging on brain gray and white matter changes assessed by conventional, MT, and DT MRI.** *Neurology* 2006;66:535–39
9. Brickman AM, Habeck C, Zarahn E, et al. **Structural MRI covariance patterns associated with normal aging and neuropsychological functioning.** *Neurobiol Aging* 2007;28:284–95. Epub 2006 Feb 15
10. Giedd JN, Rumsey JM, Castellanos FX, et al. **A quantitative MRI study of the corpus callosum in children and adolescents.** *Brain Res Dev Brain Res* 1996;91:274–80
11. Rajapakse JC, Giedd JN, DeCarli C, et al. **A technique for single-channel MR brain tissue segmentation: application to a pediatric sample.** *Mag Reson Imag* 1996;14:1053–65
12. Andreasen NC, Flaum M, Swayze VW 2nd, et al. **Intelligence and brain structure in normal individuals.** *Am J Psychiatry* 1993;150:130–34
13. Harrison PJ, Freemantle N, Geddes JR, et al. **Meta-analysis of brain weight in schizophrenia.** *Schizophr Res* 2003;64:25–34
14. Whitwell JL, Crum WR. **Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging.** *AJNR Am J Neuroradiol* 2001;22:1483–89
15. Testa C, Laakso MP, Sabattoli F, et al. **A comparison between the accuracy of voxel-based morphometry and hippocampal volumetry in Alzheimer's disease.** *J Magn Reson Imaging* 2004;19:274–82
16. Saver JL. **Time is brain-quantified.** *Stroke* 2006;37:263–66
17. Duning T, Kloska S, Steinstrater O, et al. **Dehydration confounds the assessment of brain atrophy.** *Neurology* 2005;64:548–50
18. Cho K. **Chronic "jet lag" produces temporal lobe atrophy and spatial cognitive deficits.** *Nat Neurosci* 2001;4:567–68
19. Sokol DK, Demyer WE, Edwards-Brown M, et al. **From swelling to sclerosis: acute change in mesial hippocampus after prolonged febrile seizures.** *Seizure* 2003;12:237–40
20. Frank GK, Bailer UF, Henry S, et al. **Neuroimaging studies in eating disorders.** *CNS Spectr* 2004;9:539–48
21. Moore GJ, Bebchuk JM, Wilds IB, et al. **Lithium-induced increase in human brain grey matter.** *Lancet* 2000;356:1241–42

22. Hardmeier M, Wagenpfeil S, Freitag P, et al. **Atrophy is detectable within a 3-month period in untreated patients with active relapsing-remitting multiple sclerosis.** *Arch Neurol* 2003;60:1736–39
23. Hoogervorst EL, Polman CH, Barkhof F, et al. **Cerebral volume changes in multiple sclerosis patients treated with high-dose intravenous methylprednisolone.** *Mult Scler* 2002;8:415–19
24. Gilbert AR, Moore GJ, Keshavan MS, et al. **Decrease in thalamic volumes of pediatric patients with obsessive-compulsive disorder who are taking paroxetine.** *Arch Gen Psychiatry* 2000;57:449–56
25. Denton ER, Holden M, Christ E, et al. **The identification of cerebral volume changes in treated growth-hormone deficient adults using serial 3D MR image processing.** *J Comput Assist Tomogr* 2000;24:139–45
26. Walters RJ, Fox NC, Crum WR, et al. **Haemodialysis and cerebral oedema.** *Nephron* 2001;87:143–47
27. Garver DL, Nair TR, Christensen JD, et al. **Brain and ventricle instability during psychotic episodes of the schizophrenias.** *Schizophr Res* 2000;44:11–23
28. Pfefferbaum A, Sullivan EV, Mathalon DH, et al. **Longitudinal changes in magnetic resonance imaging brain volumes in abstinent and relapsed alcoholics.** *Alcohol Clin Exp Res* 1995;19:1177–91
29. Pfefferbaum A, Rosenbloom MJ, Serventi KL, et al. **Brain volumes, RBC status, and hepatic function in alcoholics after 1 and 4 weeks of sobriety: predictors of outcome.** *Am J Psychiatry* 2004;161:1190–96
30. Agartz I, Brag S, et al. **MR volumetry during acute alcohol withdrawal and abstinence: a descriptive study.** *Alcohol Alcohol* 2003;38:71–78