Multidimensional Morphometric 3D MRI Analyses for Detecting Brain Abnormalities in Children: Impact of Control Population

Marko Wilke,^{1,2,3,4*} Douglas F. Rose,³ Scott K. Holland,⁵ and James L. Leach⁴

¹Department of Pediatric Neurology and Developmental Medicine, Children's Hospital, University of Tübingen, Germany

²Experimental Pediatric Neuroimaging, Children's Hospital and Department of Neuroradiology, University of Tübingen, Germany

³Department of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

⁴Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

⁵Pediatric Neuroimaging Research Consortium, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

Abstract: Automated morphometric approaches are used to detect epileptogenic structural abnormalities in 3D MR images in adults, using the variance of a control population to obtain z-score maps in an individual patient. Due to the substantial changes the developing human brain undergoes, performing such analyses in children is challenging. This study investigated six features derived from high-resolution T1 datasets in four groups: normal children (1.5T or 3T data), normal clinical scans (3T data), and patients with structural brain lesions (3T data), with each $n = 10$. Normative control data were obtained from the NIH study on normal brain development ($n = 401$). We show that control group size substantially influences the captured variance, directly impacting the patient's z-scores. Interestingly, matching on gender does not seem to be beneficial, which was unexpected. Using data obtained at higher field scanners produces slightly different base rates of suprathreshold voxels, as does using clinically derived normal studies, suggesting a subtle but systematic effect of both factors. Two approaches for controlling suprathreshold voxels in a multidimensional approach (combining features and requiring a minimum cluster size) were shown to be substantial and effective in reducing this number. Finally, specific strengths and limitations of such an approach could be demonstrated in individual cases. *Hum Brain Mapp* 35:3199–3215, 2014. © 2013 Wiley Periodicals, Inc.

Key words: high-resolution structural MRI; automated morphometry; control group size; brain lesion detection; multidimensional analyses

Contract grant sponsor: Deutsche Forschungsgemeinschaft; Contract grant numbers: (WI3630/1-1, M.W.); Contract grant sponsor: NIH; Contract grant numbers: N01-HD02-3343, N01-MH9-0002, and N01-NS-9-2314, -2315, -2316, -2317, -2319 and -2320; Contract grant sponsor: NIDCD; Contract grant number: R01-DC07186; Contract grant sponsor: NICHD (Pediatric Functional Neuroimaging Research Network), Contract grant number: HHSN275200900018C.

*Correspondence to: Marko Wilke, Department of Pediatric Neurology and Developmental Medicine, Children's Hospital, University of Tübingen, Hoppe-Seyler-Str. 1, 72076 Tübingen, Germany. E-mail: Marko.Wilke@med.uni-tuebingen.de

Received for publication 3 May 2013; Revised 24 July 2013; Accepted 2 August 2013

DOI 10.1002/hbm.22395

Published online 4 November 2013 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Magnetic resonance imaging (MRI) is the imaging modality of choice when imaging the developing brain, in the healthy as well as in the diseased state (Raschle et al., 2012; Schaer and Eliez, 2007; Wilke and Holland, 2008). Large-scale MRI studies have identified substantial structural brain changes as a function of both normal and abnormal development (Castellanos et al., 2002; Gogtay et al., 2004; Gothelf et al., 2007; Lenroot et al., 2007; Peterson et al., 2003; Reiss et al., 1996; Schmithorst et al., 2005). This has been shown to have important ramifications for crucial data processing steps, such as spatial normalization and tissue segmentation (Altaie et al., 2008; Hoeksma et al., 2005; Machilsen et al., 2007; Muzik et al., 2000; Wilke et al., 2002, 2003a, 2008; Yoon et al., 2009), exemplifying the importance of selecting an appropriate reference population when assessing image data in children. Among a number of demographic variables, age and gender have been shown to be most important in this respect (Castellanos et al., 2002; Giedd et al., 1996; Gogtay et al., 2004; Good et al., 2001; Lenroot et al., 2007; Wilke et al., 2002, 2007, 2008).

Automated image analysis approaches of high-resolution MRI data have already been shown to be of substantial benefit when assessing individual patients in the context of epileptogenic brain lesions, such as subcortical band heterotopia or malformations of cortical development (Barkovich et al., 2012; Bernasconi et al., 2001, 2011;

Blumcke et al., 2011; Fischl and Dale, 2000; Huppertz et al., 2008; Kassubek et al., 2002; Wilke et al., 2003b; Woermann et al., 1999). The applicability of standard statistical approaches such as voxel-based morphometry (Ashburner and Friston, 2000) when comparing a single patient with a group have been questioned (Mehta et al., 2003; Scarpazza et al., 2013), and an alternative approach is commonly employed in this scenario. Here, data from the patient under study is compared with data from a control population by spatially normalizing it, partitioning it into tissue classes, and converting it into a z-score volume, using the voxelwise mean and standard deviation of the control population (previously processed in the same way; Kassubek et al., 2002; Huppertz et al., 2005). These z-score images may also be combined and will have their highest value where the patient deviates most from “normality” (as defined by the current control population), highlighting suspicious brain areas which can then be more closely inspected visually (Wilke et al., 2003b). An overview of this approach is given in Figure 1. Careful selection of control population with as large a sample as possible, is crucial for such comparisons (Good et al., 2001; Huppertz et al., 2008).

Automated analysis approaches aimed at finding abnormal brain tissue in an individual child based on comparisons with a group norm are faced with the dilemma of closely matching on age and gender and the simultaneous requirement of including a large control population in order to ensure that structural variability (which also

Figure 1.

Overview of the general approach: data from a single individual (1) is preprocessed and partitioned into tissue classes, yielded GM, WM, CSF, a junction image, and a partial volume estimate map (2). Using age and gender as criteria, control subjects are selected from a database (3). The mean and standard deviation of these are

used to transform the individual's data into z-score maps, also yielding an additional feature of tissue composition (4). These maps can then be combined (5) to yield a final estimate of abnormality. See text for more details. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE I. Age and gender distribution of the control population from the NIH study

	Age 4–5.9 years	Age 6–7.9 years	Age 8–9.9 years	Age 10–11.9 years	Age 12–13.9 year	Age 14–15.9 years	Age 16–18.9 years	Total
Boys	19	41	31	21	30	21	28	191
Girls	18	51	31	40	23	20	27	210

changes with age; Wilke and Holland, 2003) is appropriately captured. Further, only a finite number of MRI datasets from normal children are available at any given site, which may require using data from a data repository, such as the NIH study on normal brain development (Evans et al., 2006), adding additional variance (as these children may have been scanned on different scanners, or different field strengths). Finally, different features are usually extracted from the segmented tissue maps (Hupertz et al., 2005, 2008; Wilke et al., 2011) which contributes to enhancing the detection rate of abnormalities in an individual (Bruggeman et al., 2007, 2009; Wilke et al., 2011). However, this multiple feature extraction also increases the chance of false positive findings (or, more generally, suprathreshold voxels), which is of particular relevance as multiple comparisons are not usually rigidly controlled in this setting (as is otherwise necessary when performing mass-univariate testing; Nichols, 2012).

This study is aimed at assessing which factors are most important when selecting a matched control group for use in automated data processing approach designed to detect focal brain abnormalities in an individual child, and to determine the base rate of suprathreshold voxels that can be expected when performing such analyses on multiple features.

supported by the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute of Neurological Disorders and Stroke (Evans et al., 2006). We used data from objective 1, which included children from age about 5–18 years (age was defined in months at date of scan). These scans were all obtained on 1.5T MRI systems. Following data quality assessment, a final 401 subjects were included (the same subjects that were included in Wilke et al., 2008, minus three that were removed following inadequate segmentation with the current settings [see below]), with a mean age of 128.2 ± 46.46 months at date of scan (range, 57–223 months [4.75–18.58 years]), 191 boys and 212 girls (see Table I for an overview of the age and gender distribution). These subjects constituted the normal control database. In order to address the different questions, a number of scenarios were defined for different subgroups (see also Table II). The different subgroups constitute different cases of subject data acquisition or status (step 1 in Fig. 1), while the different scenarios constitute different cases of subject selection from the control database (step 3 in Fig. 1). The impact of different pre-processing approaches (step 2 in Fig. 1) was extensively analyzed previously (Wilke et al., 2003b) and is therefore not investigated here.

METHODS AND SUBJECTS

Subjects: Origin of the Data

Data used in the preparation of this article were obtained from the NIH Pediatric MRI Data Repository created by the NIH MRI Study of Normal Brain Development. This is a multisite, longitudinal study of typically developing children from ages newborn through young adulthood, conducted by the Brain Development Cooperative Group and

Subgroup I: Healthy Controls (1.5T)

In order to assess the effect of matching on gender and the size of the control population, we identified 10 children from the NIH dataset that served as typical controls, two children each (one boy, one girl) of 8, 10, 12, 14, and 16 years of age (± 2 months); see Table II. Of course, all voxels will be found at a threshold of $\text{abs}(z) > 0$, so it is rather their presence at increasing thresholds that is of interest. As no attempt was

TABLE II. Demographic details of the four subgroups

	Subgroup I	Subgroup II	Subgroup III	Subgroup IV
Data origin	NIH study	CCHMC	CCHMC	CCHMC
Field strength	1.5T	3T	3T	3T
Normal controls	Yes	Yes	No	No
MRI read as normal	Yes	Yes	Yes	No
Number of subjects	10	10	10	10
Boys/girls	5/5	5/5	5/5	7/3
Age [mean $\pm$ SD]	12 ± 3	12 ± 3	12 ± 3	11 ± 4

NIH, National Institutes of Health; CCHMC, Cincinnati Children's Hospital Medical Center.

made to define a z -threshold that distinguishes “expected” from “unexpected” suprathreshold voxels, we used the convention of “suprathreshold voxels” (instead of true or false positives). This subgroup allowed us to identify the effect of matching on age and gender as well as yielding the base rate of suprathreshold voxels at different z -scores in healthy children scanned at 1.5T.

Subgroup II: Healthy Controls (3T)

In order to assess the effect of field strength, we searched the local database of healthy children participating in research studies at Cincinnati Children’s Hospital Medical Center for subjects that were scanned at 3T and satisfied the following criteria: no history of a neurological disorder (including epilepsy in particular), normal neurological exam, no acute or chronic illness, and no major head trauma. From these data, 10 children were identified that were matched on age and gender to the healthy controls in subgroup I, allowing for the same controls to be used in these two scenarios (see Tables II and III for details). The scans were read as normal by a board-certified pediatric neuroradiologist (JLL). This subgroup allowed us to identify the effect of field strength (when compared with subgroup I).

Subgroup III: Clinical Controls (3T)

Inclusion of this cohort was motivated by the fact that a lack of truly healthy control data available at a site may require researchers to use imaging data from children who were scanned for a clinical indication but whose scan was read as normal. This group might also be referred to as a “pseudonormal control population.” This is recognized to be suboptimal but may be necessary in specific research circumstances (Courchesne and Plante, 1996; Evans et al., 2006; Rivkin, 2000; Wilke and Holland, 2008). In order to assess the effect of using such patients as controls, we identified children from Cincinnati Children’s Hospital Medical Center radiology patient database that were scanned on a 3T MRI scanner, were read as normal by a board-certified pediatric neuroradiologist (JLL), and satisfied the following criteria: scanned for either headache, syncope, or dizziness, no history of a neurological disorder (including epilepsy in particular) or major head trauma, and no chronic illness. Ten children were identified that were matched on age and gender to the healthy controls in subgroup I and II, again allowing for the same controls to be used in these two scenarios (see Tables II and III for details). This subgroup allowed us to identify potential differences between truly versus pseudonormal children (when compared with subgroup II).

Subgroup IV: Patients (3T)

In order to qualitatively assess global effects when processing data with structural abnormalities, we identified

children from the local radiology patient database that were scanned on a 3T MRI scanner and that satisfied the following criteria: clearly identifiable brain abnormalities in cortical or subcortical regions (including cortical dysplasia, closed lip schizencephaly, heterotopic gray matter), but no other gross alteration in brain structure. Again, 10 children were identified in the age range covered by the normal control database: as these were not completely matched in age and gender to the other subgroups, their control group was chosen to have an age range of 48 months, without gender matching. This also means that they cannot statistically be compared with the other groups as their control group compositions differ. Their diagnoses were periventricular nodular heterotopia ($n = 6$), bilateral polymicrogyria and schizencephaly ($n = 1$), focal cortical dysplasia ($n = 1$), polymicrogyria ($n = 1$), and subcortical band heterotopia ($n = 1$). For further details, see Table II. This subgroup allowed us to identify the effect of pathology (when compared with subgroup II).

Image Data Acquisition and Preprocessing

For all subjects, a single high-resolution T_1 -weighted three-dimensional (3D) dataset was used. For subgroup I (NIH cohort), data were acquired on a total of seven MRI scanners of 1.5T field strength with the following parameters: TR = 90 ± 160 ms, TE = 10 ± 0.9 ms, pulse angle = $37.9 \pm 20.36^\circ$, matrix = $[255.68 \pm 3.19] \times [229.59 \pm 67.54] \times [148.96 \pm 41.35]$, voxel volume 1.44 ± 0.58 mm³. For subgroup II (healthy controls scanned at 3T), data were acquired on a Philips Achieva 3T System (Philips, Best, The Netherlands, $n = 10$) with the following parameters: TR = 8 ms, TE = 3.7 ms, pulse angle = 8° , matrix = $256 \times 256 \times 192$, voxel volume 1 mm³. For subgroups III and IV (clinical controls and patients scanned at 3T), data were acquired on either a Philips Achieva 3T System (P; Philips, Best, The Netherlands) or on a Siemens Trio 3T System (S; Siemens Medizintechnik, Erlangen, Germany) with the following parameters: TR (P/S) = 9.9/2,000 ms, TE (P/S) = 4.6/2.93 ms, flip angle (P/S) = $8/12^\circ$, matrix (P/S) = $[320/512] \times [320/512] \times [160-170/160-170]$, voxel volume 0.78 ± 0.05 mm³. Locally acquired images were anonymized prior to processing, and all procedures were in accordance with local institutional review board requirements.

All processing and analyses steps were performed using functionality available within the SPM8 software package (Wellcome Department of Imaging Neuroscience, University College London, UK) or using custom scripts and functions, running within Matlab (The Mathworks, Natick, MA). Segmentation and spatial normalization was achieved using functionality available within the vbm8 toolbox (Gaser, 2012), which offers the major advantage of not using prior information on tissue localization when performing the segmentation (Gaser et al., 2007). For the segmentation of infant data, this has already been shown

to be advantageous (Altaie et al., 2008), and for segmenting pathological brain imaging data, segmentation without relying on a priori tissue probability maps avoids the conflict of the actual (abnormal) brain and the information encoded in the (normal) tissue priors (Seghier et al., 2008, Wilke et al., 2011). Otherwise, the procedure is similar to the unified segmentation algorithm available within SPM8 (Ashburner and Friston, 2005) in that image inhomogeneities are removed in an iterative process. Spatial normalization was based on custom-generated pediatric priors based on an earlier analysis of the control population (Wilke et al., 2008). In order to enforce a harder segmentation of gray and white matter, the number of Gaussians modeling these tissue classes was set to 1 (Huppertz et al., 2008). Further, the anisotropic filtering available in vbm8 (Manjon et al., 2010) was disabled as the blurring of the gray/white-matter boundary zone is an important imaging feature in some brain malformations (Bernasconi et al., 2011; Colombo et al., 2009) that could be ameliorated by such a procedure. Data sampling was increased to be every 2 (instead of 3) mm for the same consideration. In order to minimize interpolation artifacts, 7th degree B-spline interpolation (Unser, 1999) was used throughout whenever possible. Settings were otherwise left at their default values, including the use of a Hidden Markov Random Field (HMRF; Cuadra et al., 2005) approach with a small prior probability weighting of 0.15. Images were written out to $121 \times 145 \times 121$ voxels with the standard $1.5 \times 1.5 \times 1.5$ mm resolution. The exact same settings, including the same priors, were used for all subjects.

Features

As it has been shown that combining information from different approaches allows for a more sensitive detection of abnormalities (Huppertz et al., 2005; Bruggemann et al., 2007, 2009, Wilke et al., 2011), we assessed a number of different features extracted from the images (step 4 in Fig. 1). First, the main intrabrain tissue classes of gray matter, white matter, and CSF (Ashburner and Friston, 2005) were compared. Additionally, a new approach toward assessing tissue composition (Wilke et al., 2011) was implemented. This feature is calculated by summing the absolute differences in tissue probabilities in a voxelwise fashion over the whole 3D volume between an individual patient (p) and the mean of a control population (c), such that

$$TC = |(GM_c - GM_p)| + |(WM_c - WM_p)| + |(CSF_c - CSF_p)|$$

A further feature previously suggested to be of value when assessing subtle structural brain anomalies is the so-called “junction image” (Huppertz et al., 2005, 2008). Here, the signal intensity of “average” GM and WM voxels is determined first. This is achieved by using the GM, WM tissue maps to identify voxels with intermediate signal intensity in the original T_1 , which can be expected to be more prominent in regions with a pathological tissue inter-

face, such as near brain lesions and in the vicinity of malformations of cortical development (Bernasconi et al., 2011; Colombo et al., 2009; Kassubek et al., 2002). This map is then binarized and convolved with a 5-voxel cubic binary 3D convolution filter (Huppertz et al., 2008). In order to be able to compare the results from this map with the other features, we did not mask out any brain region. A final feature aimed at exploiting subtle tissue intensity fluctuations was derived from the “partial volume estimate” available within vbm8 (Tohka et al., 2004). This parameter is also given in the form of a 3D map, containing voxel-wise information about the tissue class assignment, describing the partial volume effects as estimated during segmentation. Similar to the junction image, the partial volume estimate map should therefore be sensitive to tissue border regions in the brain where more than one tissue type is estimated to be present in a voxel.

For all features, a z-score image was calculated from the patient's input data by subtracting the mean and dividing by the standard deviation of the selected control subjects, yielding a total of six z-score maps. For the first three features, this was achieved by computing a simple mean and standard deviation over the current set of control subjects on a voxel-wise basis. For the last three features (tissue composition map, junction map, and partial volume estimate), every individual from a group of n control subjects was compared to the mean of these n control subjects in the same way as the patient, in order to determine the range of normal. From these n maps, a mean and standard deviation was then computed. In order to achieve a local pooling of the variance, the standard deviation images were smoothed by a 6-mm full-width-at-half-maximum (FWHM) Gaussian filter prior to this, as done before (Huppertz et al., 2008). As pathology may manifest itself with both abnormally high or abnormally low z-scores (a score of $z = -4$ is equally abnormal as $z = 4$), we investigated absolute z-scores only.

Scenarios

One of the a priori assumptions of the current approach was that there can be no single “normal pediatric reference population” as the variability induced by including children of all ages would preclude the sensitive detection of abnormality at any given age. Consequently, appropriate data had to be selected from the whole database according to the main factors explaining variance (age and gender; Castellanos et al., 2002; Giedd et al., 1996; Gogtay et al., 2004; Good et al., 2001; Lenroot et al., 2007; Wilke et al., 2002, 2007, 2008). Therefore, these two factors were considered when matching a control population to a subject. We defined three age ranges (12, 24, and 48 months, i.e., subject age ± 6 , 12, and 24 months, respectively), each with or without matching on gender, resulting in six scenarios that were applied to subgroup I–III (see Table III). These scenarios were coded according to the age range (ar_months)

TABLE III. Overview of the six scenarios applied to subgroups I–III and subgroup IV

Scenario age Range	Gender matched	Scenario code	Number of subjects, min (max), subgroup I–III	Number of subjects, min (max), subgroup IV
12 months ($\pm$ 6 months)	Yes	ar_12_g_1	8 (18)	NA
	No	ar_12_g_0	19 (37)	NA
24 months ($\pm$ 12 months)	Yes	ar_24_g_1	21 (39)	NA
	No	ar_24_g_0	40 (72)	NA
48 months ($\pm$ 24 months)	Yes	ar_48_g_1	43 (85)	NA
	No	ar_48_g_0	87 (156)	75 (156)

Note that while more subjects were available for some scenarios, the minimum number available for all subjects was used in subgroups I–III in order to allow for direct comparisons.

NA, not applicable; see text for details.

and whether gender was matched (g_1) or not (g_0): ar_12_g_1, ar_12_g_0, ar_24_g_1, ar_24_g_0, ar_48_g_1, and ar_48_g_0.

As each database will only have a finite number of subjects, stricter criteria will by default result in a smaller size of the control population (see Table III). We therefore first identified the scenario for each subject in subgroup I (healthy controls from the NIH study) with the largest number of control subjects, with and without matching on gender, resulting in 10 (partially overlapping) sets of $n = 43$ and $n = 87$ control subjects, respectively. We then iteratively removed two subjects at random from these datasets. For this simulation, GM variability was assessed using the standard deviation of the (decreasing) population in nine automatically defined regions of interest, as done before (Wilke, 2012a). Briefly, this approach identifies the eight cortical regions that are closest to the eight corners of the volumes, as well as the center of the volume. The regions are coded by their location in the x (left/right, L/R), y (anterior/posterior, A/P), and z -dimension (superior/inferior, S/I), resulting in RPI, RPS, RAS, LAS, LPS, LAI, RAI, LPI, and CTR (center). Within these nine regions, the standard deviation in a 20^3 voxel cube is calculated from those voxels that show an above-average variance (this keeps the resulting mean value from being dominated by noncontributing voxels). This approach allowed us to investigate the effects of decreasing sample size on variance, for both gender-matched and unmatched settings.

For each subject in each subgroup, control subjects were drawn from the control subject database for each of the six scenarios (see Table III). The minimum number of control subjects for one subject in a group determined the number of subjects for the other subjects (e.g., the smallest number of control subjects for the gender-matched controls in an age range of 12 months was 8; this was set to be the number of control subjects for all other subjects for this age range in order to allow result comparison between groups). For this simulation, the rate of voxels exceeding a given z -score was calculated for each feature, allowing us to assess the base rate of suprathreshold voxels for every z -threshold of 0–10, for every scenario and every subgroup.

Finally, we wanted to assess the effect of combining features (step 5 in Fig. 1), based on the hypothesis that, although the different features are correlated, the combination of different features should increase sensitivity and/or specificity by removing spurious suprathreshold voxels. It is expected that suprathreshold voxels in one map, if corroborated in another map, may be more indicative of a true abnormality (Bruggeman et al., 2007, 2009; Wilke et al., 2011). Conversely, deviating voxels in only one feature might be expected to have lower credibility if all other features detect no abnormality. It should be noted that the aim of the current investigation was not to establish fixed thresholds to identify pathology in a given patient (for this, the range of possible lesions was considered too wide, and the imaging characteristics too variable). Instead, our primary interest was to evaluate whether combining features would impact the rate of positive findings in healthy children (subgroup I) to establish baseline data and inform further investigations into pathologic states. We examined two approaches: (1) requiring results to be above a given threshold in two, three, and four features, and (2) requiring the suprathreshold results to exceed a given volume (i.e., number of contiguous voxels). To this effect, we examined cluster sizes of 0 (i.e., no extent threshold) and 148 as well as 296 voxels (corresponding to 0, 500, and 1,000 μ l). Again, all results were assessed using all possible combinations of these constraints at z -score thresholds of 0–10.

Implementation of the Algorithm

The algorithm was implemented in Matlab. An option for generating further customized reference databases (e.g., for younger children, or locally acquired reference data) was included. We implemented a dynamic approach in which the processed data from the whole reference population is available in a database, but the subsample to which a given patient is ultimately compared is determined by the patient's age and gender (see above). Following processing, the user is presented with the results

in an interactive graphical display, allowing exploration of the dataset by changing the z -score, the number of features in which a finding is required to be suprathreshold, and the cluster size a suprathreshold effect has to meet.

Statistical Testing

Because of small numbers, nonparametric statistical tests were used throughout, using functionality available in Matlab (the Mathworks, Natick, MA). Correlations were computed using Spearman's rank correlation, while group differences were assessed using the Mann-Whitney U -test, with significance being assumed at $P \leq 0.05$, corrected for multiple comparisons.

RESULTS

Influence of Control Group Size and Composition

When assessing the impact of increasing the group size for all subjects in nine automatically selected brain regions, either with (minimum $n = 43$) or without (minimum $n = 87$) matching for gender, the resulting patterns are remarkably similar (see Fig. 2). There is large overlap between the variance of same-gender versus same-and-different-gender control groups (up to the joint minimum $n = 43$), with no difference achieving significance, indicating that similar variance is captured irrespective of matching on gender. When assessing the correlation with age, no correlation achieved significance, indicating that the variance captured in the control population was not significantly wider in older or younger children. Further, the ongoing increase in observable variance in several regions with increasing subject number (most pronounced in CTR, see Fig. 2) suggests that, for these regions, increasing control group size continues to add variance even at $n = 87$, indicating that the full range of variability is not yet explained with smaller groups.

Base Rate of Suprathreshold Voxels

As expected, there is a strong inverse relation of the z -score and the voxels surviving this threshold in the z -score maps. This effect is not equally pronounced for all features, with the tissue composition and CSF showing the highest rate of suprathreshold voxels (see Fig. 3).

When assessing the rate of suprathreshold voxels as a function of gender, no difference remained significant after correcting for multiple testing, indicating that there were not significantly more suprathreshold voxels in boys or girls, at any threshold.

When assessing the correlation of the rate of suprathreshold voxels with age, only one correlation (at the maximum threshold of $z = 10$) remained significant after correcting for multiple testing (with Spearman's

Figure 2.

Results of simulations assessing group size (number of subjects) versus variance (standard deviation) for gender-matched (black plots, maximum $n = 43$) versus nongender-matched (grey plots, maximum $n = 87$) control groups, in nine different brain regions (right/left, anterior/posterior, superior/inferior + center). Note large overlap in variance between both (gender-matched and nonmatched) groups. See text for more details.

$\rho = 0.8909$ at $P = 0.000542$), indicating that at this threshold only, older children had more suprathreshold voxels.

When assessing the effect of increasing group size on the rate of suprathreshold voxels, a clear reduction in the rate of suprathreshold voxels can be seen at practically all thresholds in all scenarios. This effect is enumerated in Table IV. The effect is most pronounced for the tissue composition and the CSF tissue class, indicating that a larger group size will substantially reduce the rate of suprathreshold voxels.

Effects of Field Strength and “Truly Versus Pseudonormal”

When assessing the effect of field strength by comparing subgroup II with subgroup I, there are small but consistent differences (Fig. 4). Overall, the differences are marginal

Figure 3.

Ratio of suprathreshold voxels as a function of z-score for all six features for scenario I (ar_12_g-I). See also Table IV. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

on the global level, with the most pronounced changes occurring in the CSF class. None of the differences is significant.

When assessing the effect of using a pseudonormal population (subgroup III), there are again only small differences when compared with the truly normal population (subgroup II; Fig. 5). Again, the most pronounced changes occur in the CSF class. None of the differences is significant.

Using Multiple Features in Automated Morphometric Analysis Approaches

The two approaches investigated here to control the rate of suprathreshold voxels (requiring two, three, or four features to be suprathreshold and introducing a cluster size) are both effective in reducing the rate of suprathreshold voxels (Fig. 6). Forcing results to be suprathreshold in more than one feature clearly has a stronger effect than increasing the cluster size from 148 to 296 voxels, but as expected, both approaches are substantially more powerful at higher thresholds.

Performance in Patients

When assessing the global effects of investigating patients (subgroup IV) versus truly normal children (subgroup II), it is remarkable that the ratio of suprathreshold

TABLE IV. Ratio of suprathreshold voxels for all features and all scenarios

z-Score	GM			WM			CSF			TC			JI			PVE		
	12M	24M	48M	12M	24M	48M	12M	24M	48M	12M	24M	48M	12M	24M	48M	12M	24M	48M
0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
1	48.26	38.66	36.40	48.64	35.55	31.24	66.91	37.97	27.75	32.90	18.80	17.66	82.99	81.78	81.68	32.85	31.35	31.44
	39.43	36.75	35.11	36.45	31.94	28.80	40.79	28.57	22.68	19.88	18.18	17.06	82.28	81.71	81.62	31.76	31.66	31.58
2	12.28	8.22	7.43	12.92	8.60	7.40	32.19	16.29	12.29	14.49	8.59	8.13	61.67	60.20	59.91	8.05	6.95	6.97
	8.52	7.62	7.12	8.66	7.61	6.80	17.67	12.62	10.17	9.08	8.44	7.84	60.72	59.98	59.85	7.20	7.06	7.01
3	3.24	1.75	1.48	4.51	2.75	2.25	16.94	8.73	6.74	8.01	4.06	3.68	34.73	31.79	30.45	2.18	1.77	1.76
	1.85	1.54	1.39	2.78	2.33	2.01	9.45	6.86	5.54	4.51	3.90	3.48	31.91	30.61	30.00	1.87	1.82	1.80
4	1.01	0.51	0.44	1.89	1.03	0.80	10.04	5.12	3.88	5.30	1.63	1.36	11.61	8.83	7.66	.75	.56	.55
	0.53	0.45	0.41	1.05	0.84	0.69	5.65	3.96	3.11	1.96	1.47	1.22	8.74	7.70	7.32	0.62	0.60	0.58
5	0.37	0.14	0.13	0.90	0.43	0.32	6.49	3.17	2.31	2.68	0.63	0.50	2.63	1.52	1.12	0.28	0.17	0.17
	0.17	0.13	0.12	0.45	0.34	0.26	3.56	2.38	1.77	0.81	0.56	0.44	1.38	1.10	1.03	0.21	0.19	0.18
6	0.15	0.05	0.04	0.48	0.20	0.14	4.48	2.07	1.42	1.59	0.25	0.20	0.56	0.24	0.15	0.11	0.05	0.05
	0.06	0.05	0.04	0.21	0.16	0.12	2.38	1.48	1.04	0.35	0.23	0.18	0.20	0.14	0.16	0.07	0.06	0.05
7	0.08	0.03	0.02	0.27	0.11	0.07	3.28	1.42	0.93	0.88	0.12	0.09	0.17	0.06	0.04	0.04	0.01	0.01
	0.03	0.03	0.02	0.12	0.08	0.06	1.66	0.96	0.64	0.17	0.11	0.08	0.03	0.02	0.02	0.02	0.02	0.01
8	0.05	0.02	0.01	0.17	0.07	0.04	2.50	1.02	0.64	0.47	0.06	0.04	0.07	0.02	0.01	0.02	0.01	<0.01
	0.02	0.02	0.01	0.07	0.05	0.03	1.22	0.66	0.42	0.09	0.05	0.04	<0.01	<0.01	<0.01	0.01	0.01	<0.01
9	0.03	0.01	0.01	0.11	0.04	0.02	1.98	0.75	0.46	0.26	0.03	0.02	0.03	0.01	<0.01	0.01	<0.01	<0.01
	0.02	0.01	0.01	0.05	0.03	0.02	0.92	0.48	0.29	0.05	0.03	0.02	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
10	0.02	0.01	0.01	0.08	0.03	0.01	1.60	0.58	0.34	0.16	0.02	0.01	0.02	<0.01	<0.01	<0.01	<0.01	<0.01
	0.01	0.01	0.01	0.03	0.02	0.01	0.72	0.36	0.21	0.03	0.02	0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Shown are values for age ranges = 12/24/48 months, with gender matching (upper values, $n = 8/21/43$) and without (lower values, $n = 19/40/87$). See also Table III.

Figure 4.

Ratio of suprathreshold voxels as a function of z-score for all six features when comparing subgroup II (healthy controls, scanned at 3T) with subgroup I (healthy controls, scanned at 1.5T). Note slightly higher rate of surviving voxels for healthy children scanned at 3T versus 1.5T, when compared with reference data obtained at 1.5T. See also Table II.

voxels as a function of z-score for all six features shifts toward higher values for pediatric patients (Fig. 7). Individual exemplary results of patients with polymicrogyria, schizencephaly, and subcortical band heterotopia are shown in Figure 8 (top, middle, and bottom row, respectively; see Discussion section for more details).

DISCUSSION

In this work, we explore the effects of the size and composition of the control group when automatically assessing pediatric whole-brain, structural MRI data.

Influence of Control Group Size and Composition

The optimal group size for a structural neuroimaging study is surprisingly understudied: while previous simulations have shown a definite increase in the quality of mean images resulting from larger group sizes (Wilke et al., 2008), there is no accepted answer to the question of “how many subjects constitute a study.” For functional MRI, both lower and upper limits have been suggested (Friston, 2012; Friston et al., 1999; Ioannidis, 2005), as well as approaches toward exploring the power of such studies (Thirion et al.,

2007, Wilke, 2012b), but no such recommendations exist for structural imaging studies. As mentioned in the introduction, the issue of matching pertinent features that drive the variance in a given sample becomes critical when investigating the developing human brain. We therefore initially set out to assess the effect of increasing the size of the control population from a minimal 7 subjects to 87 subjects (the minimum number available for all subjects within a 48 months age range) for subgroup I. The results shown in Figure 2 demonstrate that the variance captured by a larger group is substantially higher than that in a smaller group. While this may seem trivial, there are a number of interesting points to these observations. In most automatically determined gray matter areas, there is a steep initial increase which is followed by a plateau where the effect of adding subjects decreases substantially. This plateau does not occur for a sample of less than 25–30 subjects for any of the regions we examined. However, this is not a universal lower limit for minimum control sample size, as is most apparent in the central region, covering parts of both thalami. In this region it is evident that even with a control group size of 87 subjects, variance continues to increase. This is likely explained by the fact that this region of mixed gray and white matter tissue is difficult to segment (Nugent et al., 2013), and shows a high intersubject variability (Rademacher et al., 2002). Further, the variance within regions can be substantially different and depends on the specific control population (example: LPS region in Fig. 2), again highlighting that in most cases, “more is better.” Our results suggest that adding subjects up to a minimum group size of 25–30 will always be beneficial, while group sizes exceeding 87 may still capture additional variance, particularly in brain regions with high degree of variability. However, there is not a single “magic number” of subjects required, as the effect clearly is spatially heterogeneous.

It should be noted that the effect of group size on variance (reported here as the standard deviation, which is its square root) is directly relevant as it is used to calculate the z-score in the individual patient. More variance means that the range of normal is wider, and that consequently, the likelihood for detecting subtle pathology will be lower. For example, approaches such as this one will always be less likely to detect a lesion on the interface between gray and white matter, which is highly variable due to individual differences in sulcation patterns (Rademacher et al., 1993). Conversely, a falsely low variance will increase the likelihood to detect normal tissue as abnormal. Therefore, more or less variance is not automatically good or bad. Our simulations (Fig. 2) seem to suggest that, if the variance of a population continues to change systematically upon adding or detracting a single subject, then it has not yet stabilized, and it may be worthwhile to further increase sample size. It should also be remembered that this is a dataset consisting of children. Simulations in adults, when the most rapid changes during brain development have already occurred, may give a different result.

Figure 5.

Ratio of suprathreshold voxels as a function of z-score for all six features when comparing subgroup III (patients read as normal, scanned at 3T) with subgroup II (healthy controls, scanned at 3T). Note shift toward higher values for pseudonormal children scanned at 3T versus normal children scanned at 3T, when compared with reference data obtained at 1.5T. See also Table II.

The lack of an effect of gender-matching may seem surprising at first, given the well-documented differences in regional brain volumes between boys and girls (Castellanos et al., 2002; Giedd et al., 1996; Gogtay et al., 2004;

Good et al., 2001; Lenroot et al., 2007; Wilke et al., 2007, 2008). However, it must be borne in mind that the approach presented here uses regional tissue concentration, not volume, estimates (Good et al., 2001), as we could

Figure 6.

Ratio of suprathreshold voxels in a combined z-score image with cluster sizes of 148 voxels (left) and 296 voxels (right), as a function of requiring results to be suprathreshold in one to four features. Note stronger effect of requiring more features (back to front columns) versus requiring larger clusters (left vs. right plot).

Figure 7.

Ratio of suprathreshold voxels as a function of z-score for all six features when comparing subgroup IV (pediatric patients, scanned at 3T) with subgroup II (healthy controls, scanned at 3T). Note shift toward higher values for pediatric patients scanned at 3T versus normal children scanned at 3T, when compared with reference data obtained at 1.5T. See also Table II.

previously show that using tissue concentration improves the detection rate (Wilke et al., 2003b) in this setting. Spatial normalization will remove much of the global as well as local size-differences that are present in native space, thus accounting for much of the gender-associated variance. This is in line with a recent report that confirms that gender-related volumetric differences are largely due to global brain volume differences: if global effects are accounted for, only very few regional volume differences between the genders remain significant (Brain Development Cooperative Group, 2012). Based on these previous and our current findings, we therefore suggest that matching on gender does not seem to be critical in approaches such as this one.

Base Rate of Suprathreshold Voxels

A z-score of $z = [2|3|4]$, identifies voxels outside of 2, 3, or 4 standard deviations from the mean. As we did not distinguish between “below” or “above” the mean (we investigated absolute z-scores), this corresponds to 4.55, 0.27, or 0.0064% of the voxels (assuming a normal distribution). This (given an average number of 665,105 GM voxels in our dataset) corresponds to a total of 30,262, 1,796, and 43 suprathreshold voxels, respectively. This effect is expected and should be factored into account when using z-scores for thresholding. We therefore chose the term “suprathreshold voxels” instead of “false positives” when describing the effects of altering the z-score threshold.

Figure 8.

Illustration of algorithm output in three cases from subgroup IV: upper row: a right-sided polymicrogyria is detected as being strongly abnormal in several features, but note highly noisy results in the CSF class (results not thresholded). Middle row: a bilateral posterior schizencephaly is most strongly detected as abnormal in the CSF class (results thresholded at $z = 4$ and features required

in four maps with an extent threshold of 148 voxels). Bottom row: subcortical band heterotopia detected as being strongly abnormal mainly in central parts, but not in directly subcortical regions due to high local variance (results not thresholded). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Given all the variables influencing these comparisons and the fact that assuming a normal distribution may not be legitimate (Nichols, 2012; Nichols and Holmes, 2002; Manjon et al., 2010; Salmond et al., 2002), it is imperative to know what actual numbers of suprathreshold voxels are to be expected in a given dataset of healthy subjects before applying it to assess patients. We can now provide the answer for three common settings: healthy children scanned at 1.5T (subgroup I, Fig. 3), healthy children scanned at 3T (subgroup II, Fig. 4, cf. Fig. 3 and below), and pediatric patients scanned at 3T that do not exhibit any abnormality to the eye of the experienced observer (subgroup III, Fig. 5 and below).

As expected, the rate of suprathreshold voxels (i.e., voxels surviving a given z -score) goes down substantially when increasing the z -score. The highest rate of suprathreshold voxels is seen in the tissue composition feature (Wilke et al., 2011). The tissue composition is calculated from the voxels of all tissue classes (GM, WM, and CSF), so naturally, the usefulness of this feature to detect abnormality is strongly dependent on the other tissue classes. The main contributor here seems to be CSF, which demonstrates the next highest rate of suprathreshold voxels. CSF volume is very low in children, making spurious segmentation results more likely (Wilke et al., 2003a). Further, the cleaning procedures applied to remove non-brain tissues may further decrease the reliability of this parameter, and automated segmentation approaches such as the one used here (Ashburner and Friston, 2005, with modifications detailed in Gaser et al., 2007) are usually optimized toward achieving a good segmentation of gray and white matter, with CSF being less central. This may account for the CSF class showing the highest rate of all features examined of suprathreshold voxels at higher z -scores. Another factor that might play a role here is that CSF is essentially confined to the subarachnoid space and ventricles; consequently, variance outside of these regions will be very low, which may result in artificially high z -scores. Still, the feature is potentially useful as illustrated in individual patient examples (cf. Fig. 8 and below).

Interestingly and in line with the simulations discussed above, no systematic effect of gender is observable in the base rate of suprathreshold voxels, again suggesting that it is not imperative to match on gender. There is one setting (at $z = 10$) where older children had more suprathreshold voxels; however, this is a very high threshold. Pathology was detected at an average z -score of 2.98 by Kassubek et al., 2002, with a score of >6 considered “likely representing a lesion.” Further, the number of suprathreshold voxels at this threshold were already rather low, suggesting that in the usual setting when exploring results thresholds of $z < 10$, age effects will likely not systematically alter interpretation, or the requirements concerning the size or composition of the control population.

Group size, as estimated from Figure 2, also has a substantial influence on the base-rate of suprathreshold voxels, with CSF and white matter (and, to a lesser extent,

tissue composition and GM; see Table IV) changing most. Suprathreshold CSF voxels are reduced by more than 40% at lower thresholds when using larger groups than when using smaller groups. This is the effect observable in Figure 2 and discussed above, where a wider range of normal (captured in the control population) will result in fewer voxels considered abnormal (in a given subject under study). Remarkably, the junction image and the partial volume estimate feature are rather robust even with smaller groups, that is, they do not profit as much from increasing group size. However, it must be borne in mind that this may indicate both a higher specificity or a lower sensitivity.

Effects of Field Strength and “Truly Versus Pseudonormal”

When assessing the base rate of suprathreshold voxels in healthy children scanned at 3T, differences from healthy children from the NIH study (Evans et al., 2006) seem marginal, but systematic. In the locally acquired data from healthy children scanned at 3T, there are consistently more suprathreshold voxels than in the healthy children scanned at 1.5T (see Fig. 4). At this point, it should be noted that subjects scanned at 1.5T versus subjects scanned at 3T may well differ not only with regard to the scanner’s field strength (for example, 3T scanners likely are newer machines, which may yield “better” images over older 1.5T scanners independently of the difference in field strength). These confounds are difficult to disentangle, we may therefore assess several sources of variance simultaneously. However, the difference is on the order of only a few percent, so that at a z -score of 3, there are 1.8% more suprathreshold voxels than in the average case of a subject from 1.5T (the rate is even lower for WM, but consistently higher for CSF, again suggesting caution when interpreting results from this feature). None of the differences between tissue classes reached significance. However, even in the light of this, our results are reassuring: while a slightly higher rate of suprathreshold voxels must be expected when making comparisons to reference data across field strengths, this order of magnitude does not seem to impair the general applicability of the approach to high-field data, as suggested previously (Huppertz et al., 2008).

With regard to using imaging data from children scanned for clinical reasons whose MR scans are read as normal, there are several arguments against this, based on both theoretical and practical grounds (Courchesne and Plante, 1996; Rivkin, 2000; Wilke and Holland, 2008). There is now a broad consensus that imaging data from healthy controls should be acquired through an open, ideally population-based approach (Evans et al., 2006; Paus, 2010). However, there may be situations where this, although highly desirable, may not be feasible, for example, when aiming to recruit control subjects for younger patients

investigated under general anesthesia (Ogg et al., 2009; Smith et al., 2011; Yuan et al., 2009). For settings like this, it may be unavoidable to recruit subjects scanned for another reason, and knowing how the use of pseudonormal controls systematically differs from reference data from truly healthy controls is important.

Our results indicate that there are small but systematic differences between such populations (Fig. 5) that differ by tissue classes. There are slightly fewer suprathreshold voxels in all features at lower z -scores, but slightly higher values in CSF at higher thresholds. Again, none of the differences between tissue classes reached significance, neither when compared with the truly normal controls scanned at 3T (subgroup II, cf. Fig. 5) nor when compared with the truly normal controls scanned at 1.5T (subgroup I, data not shown). The small magnitude of the difference and the lack of significance may seem to indicate that there is no practical problem with this approach (arguing against objections on theoretical grounds; Evans et al., 2006; Rivkin, 2000; Wilke and Holland, 2008). However, we still urge caution based on the fact that the pattern of the difference does not appear to be random and likely only failed to reach significance due to our small sample size ($n = 10$ in each group). As investigating this aspect more closely would consequently require many more subjects (both truly and pseudonormal), we cannot currently base a solid recommendation on our current findings. We still expect such an approach to introduce a systematic bias, however, and recommend avoiding it whenever possible.

Using Multiple Features in Automated Morphometric Analysis Approaches

It was shown before that using different imaging features may enhance the identification of structural abnormalities in MR images (Huppertz et al., 2005; Bruggemann et al., 2007, 2009; Wilke et al., 2011) as they may highlight different aspects of the pathology under investigation. Some of these features will be highly correlated, which is to be expected based on the fact that, within the brain, the segmentation algorithm only has three tissue classes available to which any given voxel can be assigned (gray and white matter and cerebrospinal fluid; Ashburner and Friston, 2005). Hence, an increase in one class must by default be accompanied by a decrease in one or more of the other classes, which is exploited in the tissue composition feature introduced earlier (Wilke et al., 2011). The junction image introduced by Huppertz et al. (2005) is aimed at finding unusually high concentrations of voxels of intermediate signal intensity, an imaging feature often observed in the vicinity of malformations of cortical development (Bernasconi et al., 2011; Colombo et al., 2009; Kassubek et al., 2002). The partial volume estimate (Tohka et al., 2004) is an attempt to capture the effects of different tissue classes contributing to one voxel, which will mainly

occur at both normal and abnormal tissue borders. Hence, it can be assumed that these six features may offer slightly different and complementary views on the same problem. Which feature will be most sensitive to detect a given lesion will very much depend on the characteristics of the lesion itself (Wilke et al., 2011), and while some may turn out to be more sensitive or specific, no attempt was made to find the definitive threshold for each feature that will delineate “right from wrong.” Instead, what we aimed at here was finding the base rate of suprathreshold voxels that can be expected as a function of different parameters.

Two constraints were introduced, based on the hypothesis that more spurious positive findings are less likely to overlap and/or to be present in larger clusters. From our simulations, it becomes clear that both forcing more than one feature to be above a given threshold and introducing an extent threshold are very effective in reducing suprathreshold voxels (Fig. 6). The effect of requiring the effect to be present in two, three, or four features, as expected, is substantially stronger as it exerts stronger constraints on the observable effects. This becomes particularly apparent at higher z -scores, which is important as these will be decisive when exploring a given patient’s dataset (Kassubek et al., 2002). Consequently, results present in several features at higher thresholds are more likely to truly represent pathology.

Performance in Patients

Although not the primary goal of this study, algorithm performance in detecting brain abnormalities in individual clinical patients was evaluated in a preliminary investigative fashion. These results are included here as they illustrate a number of interesting aspects of the approach in general, and its strengths and weaknesses in particular. For one, it is remarkable that our algorithm identifies a higher percentage of suprathreshold voxels among the patients (subgroup IV) relative to the normal controls (subgroup II) scanned at the same field strength on the global level (Fig. 7). On the individual level, our approach was also able to correctly identify obvious lesions. For example, the polymicrogyria is identified in Figure 8 (top row), at very high z -scores ($z > 10$) in >4 maps (Fig. 8, top row). The higher rate of false suprathreshold voxels in the CSF class is also apparent visually for this patient. However, in another example of a patient with schizencephaly shown in Figure 8 (middle row, $z = 4$) the abnormality is most strongly depicted in just this class, underlining that the usability of every feature will depend on the lesion’s imaging characteristics. Finally, Figure 8 (bottom row) illustrates a shortcoming of our approach in a patient with a subcortical band heterotopia that could not be identified to its full extent in regions of high interindividual variability (such as the gray/white matter interface; Rademacher et al., 1993). Gaussian smoothing further reduces the sensitivity of the algorithm to pathology in such regions (see

also next section). It must therefore be stressed again that an automatically derived result can currently only serve to highlight brain areas that the algorithm, in its specific implementation and its specific settings, considers abnormal. The output of our algorithm is therefore initially presented “as is” (all features, no thresholds, no constraints) and allows for an interactive manipulation of the available constraints (z-score, cluster size, and number of features), with immediate visual feedback. In this way, it may help to identify brain regions that are worthy of a closer second look by an experienced neuroradiologist (or, alternatively, confirm existing suspicions regarding the configuration of a given region). Detailed evaluation of the application of the described approach to clinical datasets, in particular with more subtle lesions such as focal cortical dysplasia (Barkovich et al., 2012; Blumcke et al., 2011), is a subject for further research.

Methodological Considerations

As with any approach, there is always room for improvement. The larger image inhomogeneities induced by high-field or parallel imaging (de Zwart et al., 2006; Moser et al., 2012) pose additional challenges to comparing direct image properties (such as the intermediate signal intensity voxels exploited by the junction image; Huppertz et al., 2005). This is also suggested by the rather flat appearance of this map in Figure 8 (all rows) in these patients scanned at 3T. Although the unified segmentation routine (Ashburner and Friston, 2005) has been suggested to be robust enough for yielding comparable results of both standard and high-field MRI data (Huppertz et al., 2008), it may be necessary to include an additional image inhomogeneity correction step to deal with stronger inhomogeneities. In order not to introduce additional variance in the comparison by using another software solution (e.g., Gispert et al., 2004; Hui et al., 2010; Sled et al., 1998), a feasible approach may be to remove inhomogeneities in a first step before performing the actual segmentation. For example, running the segmentation algorithm on the bias-corrected output of a first run should yield reduced sensitivity to image inhomogeneities.

Smoothing of the variance of the control population achieves a local pooling of the variance, but does reduce spatial resolution and thus, spatial specificity (see Fig. 8, bottom row). One way to address this would be to use a different approach to spatial normalization, such as the DARTEL procedure (Ashburner, 2007), aimed at achieving a closer overlap of structures and thus requiring less smoothing. However, in a systematic exploration of a number of factors, using more extensive normalization approaches was detrimental to the detection rate, as was using images modulated with the Jacobian determinant of the transformation matrix (Wilke et al., 2003b). We therefore abstained from exploring these factors again. Another alternative might be to increase signal-to-noise

by using spatially adaptive smoothing filters (Gerig et al., 1992; Manjon et al., 2010; Perona and Malik, 1990) which will decrease noise but are aimed at respecting tissue boundaries. This may allow less Gaussian smoothing and should improve spatial specificity due to the matched filter theorem (Jones et al., 2005). Finally, the conversion to obtain a z-score may be vulnerable in areas of very low variance, as discussed above. Introducing a lower threshold, or using a robust z-score (Frederix and Pauwels, 1999) could be used to this effect. Using a bootstrap approach may help to reduce the influence of outliers, especially when using a small group (see region LPS in Fig. 2), as shown before (Wilke et al., 2011; Wilke and Schmithorst, 2006).

Possible Limitations of this Study

It must be acknowledged that the scenarios established here do not allow disambiguation of the effects of a larger sample size from the effects of a larger age range. With a perfect (truly exhaustive) reference database, one could investigate 100 closely age-matched versus 100 not closely age-matched controls. This could theoretically be done using the current database, but in this case, the age range of the “closely matched” subjects would already be rather wide (>50 months in some cases), making comparisons difficult. However, comparing the performance in the 12 months/mixed gender approach with the 24 months/matched gender approach shows that the effect of widening the age range slightly (with an almost equal number of subjects) does not seem to have a substantial effect (scenarios ar_12_g1 vs. ar_24_g0 in Table IV). Given the variability of human brain development (Brain Development Cooperative Group, 2012; Gogtay et al., 2004; Paus, 2010; Wilke and Holland, 2003; Wilke et al., 2007), this is not overly surprising, as within this age range, a large variability will already be present in any case. However, given the substantial changes happening over time, the practical conclusion can only be to match children as closely as possible with regard to age, while aiming for group sizes that safely exceed a critical minimum of at least 25–30 as suggested by the simulations shown in Figure 2.

CONCLUSIONS

We have investigated the effect of control group size and composition on the results when comparing an individual child with a database of healthy controls. We found that group size substantially influences both the variance captured in the resulting tissue maps as well as an unexpected lack of influence of matching on gender. Data from high-field scanners produce slightly different base rates of suprathreshold voxels, as does using patients whose scans were read as normal. The effect of two approaches toward controlling suprathreshold voxels in a multidimensional approach were shown to be substantial and effective in

reducing this rate. Finally, specific strengths and weaknesses of such an approach were demonstrated in individual patients.

ACKNOWLEDGMENTS

The authors thank Leonid Rozhkov, CCHMC, for technical assistance, and Hansel Greiner, CCHMC, and Benjamin Bender, UK Tubingen, for helpful discussions. The algorithm is available from the author. This manuscript reflects the views of the authors and may not reflect the opinions or views of the Brain Development Cooperative Group Investigators or the NIH. The contract numbers for the NIH MRI study of normal brain development were N01-HD02-3343, N01-MH9-0002, and N01-NS-9-2314, -2315, -2316, -2317, -2319 and -2320. A listing of the participating sites and a complete listing of the study investigators can be found at http://www.bic.mni.mcgill.ca/nihp/inf/icipating_centers.html. Healthy control subject imaging data for the subgroup II were obtained in Cincinnati; for more information on the sources of funding, see the grant sponsor information section on the title page.

REFERENCES

- Altaye M, Holland SK, Wilke M, Gaser C (2008): Infant brain probability templates for MRI segmentation and normalization. *NeuroImage* 43:721–730.
- Ashburner J (2007): A fast diffeomorphic image registration algorithm. *NeuroImage* 38:95–113.
- Ashburner J, Friston KJ (2000): Voxel-based morphometry - the methods. *NeuroImage* 11:805–821.
- Ashburner J, Friston KJ (2005): Unified segmentation. *NeuroImage* 26:839–851.
- Barkovich AJ, Guerrini R, Kuzniecky RI, Jackson GD, Dobyns WB (2012): A developmental and genetic classification for malformations of cortical development: Update 2012. *Brain* 135:1348–1369.
- Bernasconi A, Antel SB, Collins DL, Bernasconi N, Olivier A, Dubeau F, Pike GB, Andermann F, Arnold DL (2001): Texture analysis and morphological processing of magnetic resonance imaging assist detection of focal cortical dysplasia in extra-temporal partial epilepsy. *Ann Neurol* 49:770–775.
- Bernasconi A, Bernasconi N, Bernhardt BC, Schrader D (2011): Advances in MRI for 'cryptogenic' epilepsies. *Nat Rev Neurol* 7:99–108.
- Blumcke I, Thom M, Aronica E, Armstrong DD, Vinters HV, Palmini A, Jacques TS, Avanzini G, Barkovich AJ, Battaglia G, Becker A, Cepeda C, Cendes F, Colombo N, Crino P, Cross JH, Delalande O, Dubeau F, Duncan J, Guerrini R, Kahane P, Mathern G, Najm I, Ozkara C, Raybaud C, Represa A, Roper SN, Salamon N, Schulze-Bonhage A, Tassi L, Vezzani A, Spreafico R (2011): The clinicopathologic spectrum of focal cortical dysplasias: A consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. *Epilepsia* 52:158–174.
- Brain Development Cooperative Group (2012): Total and regional brain volumes in a population-based normative sample from 4 to 18 years: The NIH MRI Study of Normal Brain Development. *Cereb Cortex* 22:1–12.
- Bruggemann JM, Wilke M, Som SS, Bye AM, Bleasel A, Lawson JA (2007): Voxel-based morphometry in the detection of dysplasia and neoplasia in childhood epilepsy: Combined grey/white matter analysis augments detection. *Epilepsy Res* 77:93–101.
- Bruggemann JM, Wilke M, Som SS, Bye AM, Bleasel A, Lawson JA (2009): Voxel-based morphometry in the detection of dysplasia and neoplasia in childhood epilepsy: Limitations of grey matter analysis. *J Clin Neurosci* 16:780–785.
- Castellanos FX, Lee PP, Sharp W, Jeffries NO, Greenstein DK, Clasen LS, Blumenthal JD, James RS, Ebens CL, Walter JM, Zijdenbos A, Evans AC, Giedd JN, Rapoport JL (2002): Developmental trajectories of brain volume abnormalities in children and adolescents with attention-deficit/hyperactivity disorder. *JAMA* 288:1740–1748.
- Colombo N, Salamon N, Raybaud C, Ozkara C, Barkovich AJ (2009): Imaging of malformations of cortical development. *Epileptic Disord* 11:194–205.
- Courchesne E, Plante E (1996): Measurement and analysis issues in neurodevelopmental magnetic resonance imaging. In: Thatcher RW, Lyon GR, Rumsey J, Krasnegor N, editors. *Developmental Neuroimaging: Mapping the Development of Brain and Behavior*, ed 1. San Diego: Academic Press. p 43.
- Cuadra MB, Cammoun L, Butz T, Cuisenaire O, Thiran JP (2005): Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *IEEE Trans Med Imaging* 24:1548–1565.
- de Zwart JA, van Gelderen P, Duyn JH (2006): Receive coil arrays and parallel imaging for functional magnetic resonance imaging of the human brain. *Conf Proc IEEE Eng Med Biol Soc* 1: 17–20.
- Evans AC, Brain Development Cooperative Group (2006): The NIH MRI study of normal brain development. *NeuroImage* 30: 184–202.
- Fischl B, Dale AM (2000): Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci USA* 97:11050–11055.
- Frederix G, Pauwels EJ (1999): Automatic interpretation based on robust segmentation and shape-extraction. *Lect Notes Comp Sci* 1614:773–780.
- Friston K (2012): Ten ironic rules for non-statistical reviewers. *NeuroImage* 61:1300–1310.
- Friston KJ, Holmes AP, Worsley KJ (1999): How many subjects constitute a study? *NeuroImage* 10:1–5.
- Gaser C, Altaye M, Wilke M, Holland SK (2007): Unified Segmentation Without Tissue Priors. *NeuroImage* 36(Suppl 1):S68; see also <http://dbm.neuro.uni-jena.de/vbm/vbm5-for-spm5/use-of-tissue-priors-experimental/> (retrieved August 28, 2012).
- Gaser (2012): VBM8 Toolbox. Available online at <http://dbm.neuro.uni-jena.de>.
- Gerig G, Kikinis R, Kubler O, Jolesz FA (1992): Nonlinear anisotropic filtering of MRI data. *IEEE Trans Med Imaging* 11:221–232.
- Giedd JN, Snell JW, Lange N, Rajapakse JC, Casey BJ, Kozuch PL, Vaituzis AC, Vauss YC, Hamburger SD, Kaysen D, Rapoport JL (1996): Quantitative magnetic resonance imaging of human brain development: Ages 4–18. *Cereb Cortex* 6:551–560.
- Gispert JD, Reig S, Pascau J, Vaquero JJ, García-Barreno P, Desco M (2004): Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error. *Hum Brain Mapp* 22:133–144.

- Gogtay N, Giedd JN, Lusk L, Hayashi KM, Greenstein D, Vaituzis AC, Nugent TF, Herman DH, Clasen LS, Toga AW, Rapoport JL, Thompson PM (2004): Dynamic mapping of human cortical development during childhood through early adulthood. *Proc Natl Acad Sci USA* 101:8174–8149.
- Good CD, Johnsrude I, Ashburner J, Henson RN, Friston KJ, Frackowiak RS (2001): Cerebral asymmetry and the effects of sex and handedness on brain structure: A voxel-based morphometric analysis of 465 normal adult human brains. *NeuroImage* 14:685–700.
- Gothelf D, Penniman L, Gu E, Eliez S, Reiss AL (2007): Developmental trajectories of brain structure in adolescents with 22q11.2 deletion syndrome: A longitudinal study. *Schizophr Res* 96:72–81.
- Hoeksma MR, Kenemans JL, Kemner C, van Engeland H (2005): Variability in spatial normalization of pediatric and adult brain images. *Clin Neurophysiol* 116:1188–1194.
- Hui C, Zhou YX, Narayana P (2010): Fast algorithm for calculation of inhomogeneity gradient in magnetic resonance imaging data. *J Magn Reson Imaging* 32:1197–1208.
- Huppertz HJ, Grimm C, Fauser S, Kassubek J, Mader I, Hochmuth A, Spreer J, Schulze-Bonhage A (2005): Enhanced visualization of blurred gray-white matter junctions in focal cortical dysplasia by voxel-based 3D MRI analysis. *Epilepsy Res* 67:35–50.
- Huppertz HJ, Wellmer J, Staack AM, Altenmüller DM, Urbach H, Kröll J (2008): Voxel-based 3D MRI analysis helps to detect subtle forms of subcortical band heterotopia. *Epilepsia* 49:772–785.
- Ioannidis JPA (2005): Why most published research findings are false. *PLoS Med* 2:696–701.
- Jones DK, Symms MR, Cercignani M, Howard RJ (2005): The effect of filter size on VBM analyses of DT-MRI data. *NeuroImage* 26:546–554.
- Kassubek J, Huppertz HJ, Spreer J, Schulze-Bonhage A (2002): Detection and localization of focal cortical dysplasia by voxel-based 3-D MRI analysis. *Epilepsia* 43:596–602.
- Lenroot RK, Gogtay N, Greenstein DK, Wells EM, Wallace GL, Clasen LS, Blumenthal JD, Lerch J, Zijdenbos AP, Evans AC, Thompson PM, Giedd JN (2007): Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *NeuroImage* 36:1065–1073.
- Machlisen B, d'Agostino E, Maes F, Vandermeulen D, Hahn HK, Lagae L, Stiers P (2007): Linear normalization of MR brain images in pediatric patients with periventricular leukomalacia. *NeuroImage* 35:686–697.
- Manjón JV, Coupé P, Martí-Bonmatí L, Collins DL, Robles M (2010): Adaptive non-local means denoising of MR images with spatially varying noise levels. *J Magn Reson Imaging* 31: 192–203.
- Mehta S, Grabowski TJ, Trivedi Y, Damasio H (2003): Evaluation of voxel-based morphometry for focal lesion detection in individuals. *NeuroImage* 20:1438–1454.
- Moser E, Stahlberg F, Ladd ME, Trattnig S (2012) 7-T MR—From research to clinical applications? *NMR Biomed* 25:695–716.
- Muzik O, Chugani DC, Juhász C, Shen C, Chugani HT (2000): Statistical parametric mapping: Assessment of application in children. *NeuroImage* 12:538–549.
- Nichols TE (2012): Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage* 62:811–815.
- Nichols TE, Holmes AP (2002): Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum Brain Mapp* 15:1–25.
- Nugent AC, Luckenbaugh DA, Wood SE, Bogers W, Zarate CA Jr, Drevets WC (2013): Automated subcortical segmentation using FIRST: Test-retest reliability, interscanner reliability, and comparison to manual segmentation. *Hum Brain Mapp* 34:2313–2329.
- Ogg RJ, Laningham FH, Clarke D, Einhaus S, Zou P, Tobias ME, Boop FA (2009): Passive range of motion functional magnetic resonance imaging localizing sensorimotor cortex in sedated children. *J Neurosurg Pediatr* 4:317–322.
- Paus T (2010): Population neuroscience: Why and how. *Hum Brain Mapp* 31:891–903.
- Perona P, Malik J (1990): Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell* 12: 629–639.
- Peterson BS, Anderson AW, Ehrenkranz R, Staib LH, Tageldin M, Colson E, Gore JC, Duncan CC, Makuch R, Ment LR (2003): Regional brain volumes and their later neurodevelopmental correlates in term and preterm infants. *Pediatrics* 111:939–948.
- Rademacher J, Caviness VS Jr, Steinmetz H, Galaburda AM (1993): Topographical variation of the human primary cortices: Implications for neuroimaging, brain mapping, and neurobiology. *Cereb Cortex* 3:313–329.
- Rademacher J, Bürgel U, Zilles K (2002): Stereotaxic localization, intersubject variability, and interhemispheric differences of the human auditory thalamocortical system. *NeuroImage* 17:142–160.
- Raschle N, Zuk J, Ortiz-Mantilla S, Sliva DD, Franceschi A, Grant PE, Benasich AA, Gaab N (2012): Pediatric neuroimaging in early childhood and infancy: Challenges and practical guidelines. *Ann N Y Acad Sci* 1252:43–50.
- Reiss AL, Abrams MT, Singer HS, Ross JL, Denckla MB (1996): Brain development, gender and IQ in children. A volumetric imaging study. *Brain* 119:1763–1774.
- Rivkin MJ (2000): Developmental neuroimaging of children using magnetic resonance techniques. *Ment Retard Dev Disabil Res Rev* 6:68–80.
- Salmond CH, Ashburner J, Vargha-Khadem F, Connelly A, Gadian DG, Friston KJ (2002): Distributional assumptions in voxel-based morphometry. *NeuroImage* 17:1027–1030.
- Scarpazza C, Sartori G, De Simone MS, Mechelli A (2013): When the single matters more than the group: Very high false positive rates in single case Voxel Based Morphometry. *NeuroImage* 70:175–188.
- Schaer M, Eliez S (2007): From genes to brain: Understanding brain development in neurogenetic disorders using neuroimaging techniques. *Child Adolesc Psychiatr Clin N Am* 16:557–579.
- Schmithorst VJ, Wilke M, Dardzinski BJ, Holland SK (2005): Cognitive functions correlate with white matter architecture in a normal pediatric population: A diffusion tensor MRI study. *Hum Brain Mapp* 26:139–147.
- Seghier ML, Ramalackhansingh A, Crinion J, Leff AP, Price CJ (2008): Lesion identification using unified segmentation-normalisation models and fuzzy clustering. *NeuroImage* 41: 1253–1566.
- Sled JG, Zijdenbos AP, Evans AC (1998): A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging* 17:87–97.
- Smith CD, Chebrolu H, Wekstein DR, Schmitt FA, Markesbery WR (2007): Age and gender effects on human brain anatomy: A voxel-based morphometric study in healthy elderly. *Neurobiol Aging* 28:1075–87.

- Thirion B, Pinel P, Me'riaux S, Roche A, Dehaene S, Poline JB (2007): Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage* 35:105–120.
- Tohka J, Zijdenbos A, Evans A (2004): Fast and robust parameter estimation for statistical partial volume models in brain MRI. *NeuroImage* 23:84–97.
- Unser M (1999): Splines: A perfect fit for signal and image processing. *IEEE Sign Proc Mag* 16:22–38.
- White SH (2000): Conceptual foundations of IQ testing. *Psychol Pub Pol Law* 6:33–43.
- Wilke M (2012a): An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. *NeuroImage* 59:2062–2072.
- Wilke M (2012b): An iterative jackknife approach for assessing reliability and power of FMRI group analyses. *PLoS One* 7: E35578.
- Wilke M, Holland SK (2003): Variability of gray and white matter during normal development: A voxel-based MRI analysis. *Neuroreport* 14:1887–1890.
- Wilke M, Holland SK (2008): Structural MR-Imaging studies of the brain in children: Issues and opportunities. *Neuroembryol Aging* 5:6–13.
- Wilke M, Schmithorst VJ, Holland SK (2003a): Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. *Magn Reson Med* 50:749–757.
- Wilke M, Kassubek J, Ziyeh S, Schulze-Bonhage A, Huppertz HJ (2003b): Automated detection of gray matter malformations using optimized voxel-based morphometry: A systematic approach. *NeuroImage* 20:330–343.
- Wilke M, Schmithorst VJ (2006): A combined bootstrap/histogram analysis approach for computing a lateralization index from neuroimaging data. *NeuroImage* 33:522–530.
- Wilke M, Krägeloh-Mann I, Holland SK (2007): Global and local development of gray and white matter volume in normal children and adolescents. *Exp Brain Res* 178:296–307.
- Wilke M, Schmithorst VJ, Holland SK (2002): Assessment of spatial normalization of whole-brain magnetic resonance images in children. *Hum Brain Mapp* 17:48–60.
- Wilke M, de Haan B, Juenger H, Karnath HO (2011): Manual, semi-automated, and automated delineation of chronic brain lesions: A comparison of methods. *NeuroImage* 56:2038–2046.
- Woermann FG, Free SL, Koepp MJ, Ashburner J, Duncan JS (1999): Voxel-by-voxel comparison of automatically segmented cerebral gray matter—A rater-independent comparison of structural MRI in patients with epilepsy. *NeuroImage* 10:373–384.
- Yoon U, Fonov VS, Perusse D, Evans AC; Brain Development Cooperative Group (2009): The effect of template choice on morphometric analysis of pediatric brain data. *NeuroImage* 45: 769–777.
- Yuan W, Mangano FT, Air EL, Holland SK, Jones BV, Altaye M, Bierbrauer K (2009): Anisotropic diffusion properties in infants with hydrocephalus: A diffusion tensor imaging study. *Am J Neuroradiol* 30:1792–1798.