◆ Human Brain Mapping 28:1075–1088 (2007) ◆

The Bootstrap and Cross-Validation in Neuroimaging Applications: Estimation of the Distribution of Extrema of Random Fields for Single Volume Tests, with an Application to ADC Maps

Roberto Viviani,[1]* Petra Beschoner,[1] Tina Jaeckle,[2] Peter Hipp,[3] Jan Kassubek,[4] and Bernd Schmitz[2]

[1]*Department of Psychiatry III, University of Ulm, Germany*
[2]*Department of Radiology, University of Ulm, Germany*
[3]*Department of Radiotherapy, University of Ulm, Germany*
[4]*Department of Neurology, University of Ulm, Germany*

Abstract: We discuss the assessment of signal change in single magnetic resonance images (MRI) based on quantifying significant departure from a reference distribution estimated from a large sample of normal subjects. The parametric approach is to build a test based on the expected distribution of extrema in random fields. However, in conditions where the variance is not uniform across the volume and the smoothness of the images is moderate to low, this test may be rather conservative. Furthermore, parametric tests are limited to datasets for which distributional assumptions hold. This paper investigates resampling methods that improve statistical tests for signal changes in single images in such adverse conditions, and that can be used for the assessment of images taken for clinical purposes. Two methods, the bootstrap and cross-validation, are compared. It is shown that the bootstrap may fail to provide a good estimate of the distribution of extrema of parametric maps. In contrast, calibration of the significance threshold by means of cross-validation (or related sampling without replacement techniques) address three issues at once: improved power, better voxel-by-voxel estimate of variance by local pooling, and adaptation to departures from ideal distributional assumptions on the signal. We apply the cross-validated tests to apparent diffusion coefficient maps, a type of MRI capable of detecting changes in the microstructural organization of brain parenchyma. We show that deviations from parametric assumptions are strong enough to cast doubt on the correctness of parametric tests for these images. As case studies, we present parametric maps of lesions in patients suffering from stroke and glioblastoma at different stages of evolution. *Hum Brain Mapp 28:1075–1088, 2007.*

Key words: random fields; bootstrap; cross-validation; diffusion-weighted imaging; ADC maps

INTRODUCTION

There is a parallel evolution of MRI scanning techniques and computer-aided diagnostic approaches. Emerging MR techniques like diffusion-weighted imaging are widely used in the setting of acute stroke, but it is increasingly recognized that they may be of value for the diagnosis of other cerebral diseases [Moritani et al., 2005]. However, computed images from modern MR sequences like apparent diffusion coefficient (ADC) maps provide more information than the eye

*Correspondence to: Roberto Viviani; Department of Psychiatry III, University of Ulm, Leimgrubenweg 12, 89075 Ulm, Germany.
E-mail: roberto.viviani@uni-ulm.de

Received for publication 3 February 2006; Revision 11 July 2006; Accepted 28 July 2006

DOI: 10.1002/hbm.20332

Published online 31 January 2007 in Wiley InterScience (www.interscience.wiley.com).

can easily catch. This, and their quantitative nature, makes them particularly indicated for computer-aided diagnostic approaches.

In the field of cognitive neurosciences, the statistical parametric mapping (SPM) approach has been widely used with success for the detection of small significant changes in the signal obtained with positron emission tomography and functional magnetic resonance imaging techniques [Friston, 1996; Worsley et al., 2002]. The application of this statistical approach in the clinic, however, has been limited. Among the few existing examples are voxel-based studies to detect focal cortical dysplasia in patients with seizure disorders [Colliot et al., 2006; Kassubek et al., 2002; Woermann et al., 1999a, b]. For these statistical approaches to become more widely available in clinical practice they need be validated and extended to address the specific needs of clinical applications.

This paper is specifically concerned with the implementation of single volume tests for the study of images obtained for clinical purposes. There are several differences from the usual SPM [Friston et al., 1991; Worsley et al., 1992, 1996] or nonparametric mapping applications [SnPM: Holmes et al., 1996; Nichols and Holmes, 2001; for a comprehensive review, see Lange, 1999]. In the functional imaging setting in which parametric or nonparametric approaches were developed, the test statistic derives from a statistical model of the effects of the experimental conditions. By contrast, in the present setting the signal is not determined by the manipulations of an experimental paradigm. Rather, one posits that the signal detects tissue abnormalities by identifying signal deviating from a reference interval that is observed in the healthy brain. Furthermore, to be applicable in clinical investigations, the statistical inference must deliver usable results even if only one image from an individual patient is analyzed. These differences may be summarized by noting that in single volume tests the general linear model is replaced by the simpler model of a reference distribution.

In our study we will focus on resampling techniques, which will be used to estimate the empirical distribution of the extrema of random fields, and hence of the reference interval against which abnormal signal is identified. There are two reasons for this choice. First, the signal in structural images may not satisfy the distributional assumptions of a parametric test. This is especially a concern if a single image is tested, as in this case data are not made more normal by averaging. Second, resampling procedures are usually more efficient than parametric procedures based on random field theory such as SPM. In particular, if it cannot be assumed that the variance is uniform across the image (as is almost invariably the case in MRI), and the extent of the spatial correlation (or smoothness) of the signal is not high, random field theory tends to deliver too conservative tests [Nichols and Hayasaka, 2003; Worsley, 2005]. For this reason, increase of the degrees of freedom of the variance estimate by local pooling [Worsley et al., 2002] or the adoption of nonparametric tests based on permutation techniques such as SnPM [Holmes et al., 1996; Nichols and Holmes, 2001] have been advocated when voxel-level control of error rates is required in the linear model. Permutation, however, is not applicable in the single-image testing situation considered here. Hence, two alternative resampling approaches are explored here: the bootstrap [Efron and Tibshirani, 1993] and cross-validation [Stone, 1974].

The comparison of cross-validation and the bootstrap constitutes the methodological motivation of this study. In the words of Casella and Berger [2002], the bootstrap "is perhaps the single most important development in statistical methodology in recent times" (p. 517). Its application to neuroimaging, however, has not yet been investigated. Even if the utility of the bootstrap has been proven in a variety of situations, its applicability to specific problems cannot be taken for granted. The performance of the bootstrap as an estimator outside a number of well-known situations must be investigated in each case, since very often specific techniques must be employed to harness its power [Casella and Berger, 2002]. In some cases, the bootstrap is known to perform poorly [Bickel and Freedman, 1981]. In fact, it will be shown here that two bootstrap resampling schemes fail to provide satisfactory estimates of the empirical distribution of the extrema of random fields. In contrast, we will show that cross-validation provides a better approximation of the empirical distribution. In the discussion, these findings will be explained in the light of the relevant theoretical results in the field of nonparametric resampling methods [Politis et al., 1999], demonstrating that the type of resampling carried out by cross-validation succeeds when the bootstrap fails. To make use of the cross-validation estimate, we will adopt a semiparametric approach in which a random field model of the data is complemented with cross-validation to calibrate the error rate of the significance test [Loh, 1987]. This is motivated by the relative small size of the cross-validated sample.

As a case study, we will consider the application of quantitative brain mapping to ADC maps. Diffusion-weighted imaging is an emerging MRI technique to detect changes in the microstructure of brain tissue by measuring the distance traveled by diffusing water molecules. Destruction of tissue or accumulation of water in the intercellular space (edema) increase water diffusion. Shift of water into the intracellular compartment prior to cellular death, by contrast, decreases it. The signal becomes sensitive to diffusion when two additional linear gradients are added to the sequence between the excitation pulse and acquisition [Stejskal and Tanner, 1965]. The gradients have opposite effects, so that they cancel each other if the spinning hydrogen atoms remain in the same place. Because the phase offset induced by the gradients is location-dependent, the second gradient will not undo the effect of the first gradient completely if the hydrogen atom has moved. This results in a loss of signal due to dephasing. By repeating the measurement with gradients of different intensity and in different directions, the ADC of water can be calculated on the basis of the physical laws regulating the diffusion process and their effect on relaxation [Le Bihan, 1991]. In providing data on water

diffusion, ADC maps provide information on the microstructural integrity of brain tissues. ADC signal increases are most marked when the interstitial water content increases, since water molecules can then move more freely. Typical conditions giving rise to ADC increased signal are edema (as in inflammatory conditions or surrounding neoplastic lesions) or destruction of tissue (as in advanced cicatritial processes). The most common cause of ADC reduced signal is fresh cerebral infarction, when water molecules migrate into metabolically damaged cells. Reduced ADC can also be seen in the early stages of cerebral hemorrhage and following axonal damage [for a comprehensive treatment, see Moritani et al., 2005]. We will provide case studies of both increased and decreased signal changes.

The paper is organized as follows. In the method section we will first briefly formulate the statistical testing framework for the detection of abnormal signals based on the departure from a reference interval, and describe the resampling schemes investigated in the study. The methods section will also contain details on the simulations, the software used, and the collection of MR data. The results section is divided into two main parts. In the first part we will examine the performance of the resampling techniques with artificial data. In this part we demonstrate the problems incurred by the bootstrap, the efficacy of cross-validated calibration, and its superiority to parametric techniques based on random fields theory. In the second part of the results section we will present the results of applying random field testing to ADC maps of subjects affected by stroke and brain tumor. The main findings are summarized in the discussion section.

MATERIALS AND METHODS

Specification of the Model

After preprocessing, the voxel-by-voxel signal y deriving from normal tissue is modeled by

$$y = \mu + \sigma\varepsilon, \tag{1}$$

where μ and σ are the voxel-by-voxel signal mean and standard deviation, and ε is a zero-mean, unit variance error term (to simplify notation, we will omit the index referring to the voxel). The null hypothesis, therefore, is that $y = \mu$. For parametric inference, ε is further assumed to be distributed as a stationary Gaussian random field across the volume. In this case, we may write

$$y \sim \mathrm{N}(\mu, \sigma^2). \tag{2}$$

By contrast, the alternative hypothesis is that signal deriving from abnormal tissue may take values that are very different from the mean signal μ. The statistic giving the departure of the observed value y from the reference distribution is

$$t = \frac{y - \hat{\mu}}{\hat{\sigma}\sqrt{(N + 1)/N}}. \tag{3}$$

where the voxel-by-voxel signal mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ are estimated from a reference sample of N individuals that is representative of the population of interest. Across the whole volume, the voxel-by-voxel t values form the parametric map $\{t\}$ which, under the null hypothesis and the distributional assumptions on the error term, is a lattice approximation of a t random field with $N - 1$ degrees of freedom, approximating the Gaussian random lattice as $N \to \infty$.

To implement a test, we need to compute a threshold $\vartheta(\alpha)$ to identify outlying values of t with significance level α, corrected for the multiple tests over the volume. In parametric random field theory tests, $\vartheta(\alpha)$ is given by the expected Euler characteristic of the t random field (Worsley et al., 1992, 1996). In the resampling approach adopted here, $\vartheta(\alpha)$ is estimated from the empirical distribution of $\max(\{t\})$ (the largest value of the parametric map $\{t\}$) in the reference sample.

Bootstrap Resampling Schemes

At a given threshold ϑ, the achieved significance level $\alpha_{\mathrm{ASL}}(\vartheta)$ of a test on a random field is given by

$$\alpha_{\mathrm{ASL}}(\vartheta) = \mathrm{Prob}_{H_0}\{\max(\{t\}) \geq \vartheta\}, \tag{4}$$

where $\{t\}$ is the random field variable distributed according to the null hypothesis H_0. The bootstrap will be used to generate an empirical distribution of $\max(\{t\})$ by resampling this statistic from a dataset for which the null hypothesis H_0 holds. The bootstrap estimate of $\alpha_{\mathrm{ASL}}(\vartheta)$ is then the proportion of random fields in the resampled empirical distribution where $\max(\{t\}) \geq \vartheta$. The purpose of this estimate is to later set the threshold ϑ so that the estimated achieved significance level is equal to the nominal size of the test.

In this study two bootstrap resampling schemes for the estimate of the distribution of $\max(\{t\})$ will be considered. The first scheme, simple bootstrap, is a generic bootstrap scheme which has also been proposed to select an estimator indexed by a parameter. In this context, it is also known as bootstrap adaptation. In this method, the estimate of the achieved significance level is calculated on the bootstrap sample. The second scheme, which we will call for convenience leave-out bootstrap, is similar to cross-validation in its main application to prediction problems. In the leave-out bootstrap, the estimate of the achieved significance level is calculated on the volumes that were not used to form the bootstrap sample. These schemes are discussed in the context of calibration of significance tests in Efron and Tibshirani [1993].

In the simple bootstrap N volumes at a time are sampled, with replacement, from the original sample, also of size N. Therefore, this bootstrap sample will differ from the original sample in that some volumes will not be contained in it, and other volumes will have been selected more than once. Let $\hat{\mu}^{*b}$, $\hat{\sigma}^{*b}$ be the mean and standard deviation estimated voxel-by-voxel from the b bootstrap sample,[1] $b = 1, 2, \ldots B$ (where B is the number of bootstrap samples). Let x_i be the signal after the preprocessing steps; the bootstrap standardized score z_i^{*b} for each volume i of the original reference sample is given by

$$z_i^{*b} = \frac{x_i - \hat{\mu}^{*b}}{\hat{\sigma}^{*b}\sqrt{(N+1)/N}}, \qquad b = 1, 2, \ldots B, \quad i = 1, 2, \ldots N, \tag{5}$$

and the proportion of the volumes with score over the threshold by

$$\alpha^{*b}_{\text{A-BOOT}}(\vartheta) = N^{-1}\sum_{i=1}^{n} I\{\max(\{z_i^{*b}\}) \geq \vartheta\}, \tag{6}$$

where I is the index function, $I(\epsilon) = 1$ if the expression ϵ is true, $I(\epsilon) = 0$ otherwise. The proportions from the bootstrap samples are averaged to give the bootstrap estimate of the achieved significance level:

$$\hat{\alpha}_{\text{A-BOOT}}(\vartheta) = B^{-1}\sum_{b=1}^{B} \alpha^{*b}_{\text{A-BOOT}}(\vartheta). \tag{7}$$

In the leave-out bootstrap one samples again, with replacement, N volumes from the original sample, also of size N. As before, this bootstrap sample will differ from the original sample in that some volumes will not be contained in it, and other volumes will be represented more than once. Here, however, only the volumes not included in the bootstrap sample will be used to compute the proportion of volumes with overthreshold scores. One will have, at each bootstrap sample b, Q_b volumes in the sample and R_b volumes out of it, $Q_b + R_b = N$. Let $\hat{\mu}^{*b}$, $\hat{\sigma}^{*b}$ be the mean and standard deviation estimated from the volumes in the bootstrap sample b. The leave-out bootstrap standardized score for each volume not included in the bootstrap sample r is given by

$$z_r^{*b} = \frac{x_r - \hat{\mu}^{*b}}{\hat{\sigma}^{*b}\sqrt{(N+1)/N}}, \qquad b = 1, 2, \ldots B, \quad r = 1, 2, \ldots R_b, \tag{8}$$

and the proportion of volumes with score over the threshold by

$$\alpha^{*b}_{\text{LO-BOOT}}(\vartheta) = R_b^{-1}\sum_{r=1}^{R} I\{\max(\{z_r^{*b}\}) \geq \vartheta\}, \tag{9}$$

which are averaged to give the bootstrap estimate of the achieved significance level:

$$\hat{\alpha}_{\text{LO-BOOT}}(\vartheta) = B^{-1}\sum_{b=1}^{B} \alpha^{*b}_{\text{LO-BOOT}}(\vartheta). \tag{10}$$

Cross-Validation

Cross-validation [Stone, 1974] is a resampling scheme whose fundamental difference from the bootstrap is that sampling is carried out without replacement. In addition, cross-validation schemes typically include prescriptions for the formation of the samples, instead of leaving it to random selection. The common application of cross-validation is the estimate of the risk incurred by a nonparametric fit on the basis of which the extent of smoothness of the fit is chosen [Green and Silverman, 1994; Wahba, 1990].

In simple cross-validation, one observation at a time is left out of the sample used in the estimate. Hence, if there are N volumes in the original reference sample, there are as many distinct cross-validation samples: $B = N$. Let $\hat{\mu}^{(i)}$, $\hat{\sigma}^{(i)}$ be the mean and standard deviation estimated from the reference sample after excluding the volume i from it. For each volume, the leave-one-out standardized score $z^{(i)}$ is given by

$$z^{(i)} = \frac{x_i - \hat{\mu}^{(i)}}{\hat{\sigma}^{(i)}\sqrt{N/(N-1)}}, \qquad i = 1, 2, \ldots N. \tag{11}$$

The cross-validation estimate of the achieved significance level $\hat{\alpha}_{\text{CV}}(\vartheta)$, analogous to the bootstrap estimate, is the proportion of volumes containing at least one leave-one-out score over the cut-off threshold ϑ:

$$\hat{\alpha}_{\text{CV}}(\vartheta) = N^{-1}\sum_{i=1}^{N} I(\max(\{z^{(i)}\}) \geq \vartheta). \tag{12}$$

Note that here the size of the cross-validation sample is $N - 1$, so that $\hat{\alpha}_{\text{CV}}(\vartheta)$ is biased since it refers to references sample of this size. If N is large, cross-validation may be carried out by leaving out an entire block of observations at a time instead of just one (n-fold cross-validation). The set of the original reference sample is divided into a number of blocks. Each block of observations in turn is left out to calculate the cross-validation estimates of the

[1] The bootstrap standard deviation estimate $\hat{\sigma}^{*b}$ is estimated from the bootstrap sample using the functional estimator $\hat{\sigma} = \sqrt{N^{-1}\sum_i^N (x_i - \hat{\mu})^2}$, instead of the unbiased estimator $\hat{\sigma} = \sqrt{(N-1)^{-1}\sum_i^N (x_i - \hat{\mu})^2}$ used elsewhere in this paper [Efron and Tibshirani, 1993, p. 298]. The use of the functional estimator is justified by the presence of duplicate observations in the bootstrap sample.

voxel-by-voxel mean and variance. The estimate of the achieved significance level $\hat{\alpha}_{CV}(\vartheta)$ is calculated as earlier. This makes the cross-validation estimate of the achieved significance level even more biased, since after leaving out the block, the estimate is now based on a reference sample of size Q, $Q < N$. However, in the field of nonparametric function estimation the estimate is known for carrying less variance. In the results section, simulations will be presented showing that this bias-variance tradeoff also holds for the calibration of random fields.

To isolate the property of cross-validation and bootstrap resampling schemes that is responsible for their different success in resampling extrema, we will need to carry out simulations with a version of cross-validation that resembles the bootstrap, and that is obtained by considering all possible blocks of size Q, giving each combination of observations equal chances to be selected to form a block (this scheme differs from the bootstrap in that sampling within the block is done without replacement). There are then $\binom{N}{Q}$ possible ways to form a block, and if this number is too large, one may sample from the possible combinations at random (resampled cross-validation).

It is important to note that unless the size of the original sample N is very large, Eq. (12) can deliver only imprecise estimates of $\alpha_{CV}(\vartheta)$ in simple or n-fold cross-validation, since the number of the resampled thresholds is equal or smaller than N. Resampled cross-validation could provide a large cross-validation sample if $Q \ll N$, but this makes the resulting estimate very biased (resampled cross-validation is introduced here only for the purpose of comparing it with the bootstrap). For this reason, cross-validation is here complemented with a calibration procedure [Loh, 1987].

In what follows, we investigate the calibration of t random fields by adjusting their degrees of freedom. Typically, calibration is carried out on the nominal alpha level of the test [as in Loh, 1987]; here we chose to adjust the degrees of freedom. Once the degrees of freedom of a random field are calibrated, all significance levels within the range of the calibration become available. The calibration of the t field also improves the power of the test when the variance is estimated by local pooling, which exploits the fact that the variance in adjacent voxels may be similar, even if not uniform across the whole volume. Local pooling of the variance estimate becomes automatically available to the testing procedure, since resampling methods take care of the resulting reduced standard error of the voxel-by-voxel variance estimate [Nichols and Holmes, 2001]. When adjustment of the degrees of freedom of the field does not suffice to move the cut-off thresholds to the desired levels, it may be combined with calibration of the nominal alpha levels.

Let $\vartheta_{RFT,\lambda}(\alpha)$ be the threshold from the expected Euler characteristic of a t random field with degrees of freedom λ giving a test with theoretical significance level α. Using cross-validation, empirical estimates of the achieved significance level $\hat{\alpha}_{CV}(\vartheta_{RFT,\lambda}(\alpha))$ are computed from the reference sample. Let $\mathbf{a}$ be a vector of nominal alpha levels in the neighborhood of the target significance level, and $\hat{\mathbf{a}}_{CV,\lambda}$ the corresponding empirical cross-validation estimates of the achieved significance levels of the random fields theory thresholds for degrees of freedom λ. A simple way to obtain the cross-validation estimate of the required calibration is by choosing the value of λ that minimizes the square distance between the two vectors:

$$\hat{\lambda}_{CV} : \arg \min_{\lambda} \left\| \hat{\mathbf{a}}_{CV,\lambda} - \mathbf{a} \right\|^2. \quad (13)$$

This minimization problem is approached by calculating $\hat{\mathbf{a}}_{CV,\lambda}$ for a grid of λ values and then choosing the λ value for which $\left\| \hat{\mathbf{a}}_{CV,\lambda} - \mathbf{a} \right\|^2$ is smallest. This procedure is iterated a few times with smaller and smaller grids centered on the chosen λ. The threshold $\hat{\vartheta}_{RFT,\hat{\lambda}_{CV}}(\alpha)$ for a calibrated test with size α is given by the expected Euler characteristic of a t random field with $\hat{\lambda}_{CV}$ instead of $N - 1$ degrees of freedom. An identical procedure is applicable to bootstrap estimates of achieved significance values.

In the construction of the vector $\mathbf{a}$ one is tempted to use the small nominal α levels that will be used in practice. However, if only few observations are available, it will be difficult to evaluate the empirical error rate at such small levels. In a sample of 15, for example, the smallest positive error rate above zero is $1/15 \approx 0.0667$. This problem is compounded by the square error metric. Since there can be no negative error rates, the fit will be lenient with any underestimation of the effective rate at small nominal α levels, while overestimation will move the cross-validation estimate downwards. It is therefore important to use an effective range of values in the composition of the vector $\mathbf{a}$. In simulations with artificial data distributed according to the null we found that the cross-validation estimate improves when including α levels as high as 0.9. Unfortunately, the improvement obtained by using high α levels might not carry over to the real data if the distributional assumptions are violated, since the departure from the theoretical α levels will not be uniform across a wide α levels range. Thus, if many volumes are available, more accurate results may be obtained with narrow ranges. However, if few volumes are available, one is forced to use a relatively wide range for the vector $\mathbf{a}$, compensating the scarcity of data with more extensive use of the parametric assumptions. In this study, the range of the values of the vector $\mathbf{a}$ was restricted to the interval [0.01–0.2] in the calibration of the ADC maps, and the relatively narrow interval [0.01–0.5] was used in the simulations in order not to depart excessively from real-life conditions.

Simulations: Estimation of Achieved Significance Levels

All simulations are based on the generation of artificial volumes of size $32 \times 32 \times 32$ voxels obtained by convolving a Gaussian kernel (full width-half maximum, FWHM 4 voxels) with a standard normal variate. To avoid edge effects, images were padded at the sides with zero values for 3

times the FWHM size of the kernel. To make the results comparable to those of the existing studies in the literature, we will make use of volumes of size typical of EPI acquisitions [Nichols and Hayasaka, 2003].

In each trial, artificial volumes distributed as t random fields were generated using Eq. (3) for a reference sample size of $N = 15$ and a smoothing kernel of FWHM = 4 voxels (these values were chosen to keep computations feasible and remain within the parameter range explored by Nichols and Hayasaka [2003]). The parametric thresholds according to random field theory ϑ_{RFT} were calculated for a range of nominal α levels between 0.5 and 0.01. In each trial, the achieved α levels were estimated by creating 600 new artificial images and computing the proportion of these having a maximum above the theoretical threshold ϑ_{RFT}. Furthermore, in each trial the bootstrap methods presented in the previous sections [Eqs. (7) and (10), $B = 600$] and simple cross-validation [Eq. (12)] were used to estimate the achieved significance levels $\hat{\alpha}_{\mathrm{A-BOOT}}(\vartheta_{\mathrm{RFT}})$, $\hat{\alpha}_{\mathrm{LO-BOOT}}(\vartheta_{\mathrm{RFT}})$, and $\hat{\alpha}_{\mathrm{CV}}(\vartheta_{\mathrm{RFT}})$ of the parametric random field theory thresholds from the 15 volumes of the reference sample of the trial. Trials were repeated 60 times to create the box plots and histograms of the figures.

Simulations: Calibration of Artificial Random Fields Using Cross-Validation

As in the previous simulations, artificial volumes distributed as t random fields using a reference sample of size $N = 15$ were generated in each of 60 trials. Again, parametric thresholds were calculated for a range of nominal α levels between 0.5 and 0.01. This time, also the cross-validated calibrated thresholds $\hat{\vartheta}_{\mathrm{RFT},\hat{\lambda}_{\mathrm{CV}}}(\alpha)$ [Eq. (13) and explanation thereby] were estimated in each trial, and the achieved α levels were computed from 600 new artificial images for both parametric and calibrated thresholds. The process was repeated for smoothing kernels of different size.

Software

All code implementing the algorithms and the simulations presented here was developed on MATLAB 6.1 R12 (The Mathworks, Natick, MA) installed on a Pentium PC running Windows 2000 (Microsoft, Redmond, WA). Our implementation is written as an SPM toolbox (SPM99/SPM2, Wellcome Department of Cognitive Neurology, London; online at http://www.fil.ion.ucl.ac.uk, Frackowiak et al., 1997). We made use of the software in that package in original or adapted form for realignment, stereotactic normalization, segmentation, smoothing, and estimation of random field theory thresholds.

Image Acquisition

All MRI data were obtained with a 1.5-Tesla Magnetom Symphony (Siemens, Erlangen, Germany) whole-body MRI system equipped with a head volume coil. All subjects were scanned at the Department of Radiology of the University of Ulm as part of the clinical routine or an elective examination, and were patients under the care of one or more of the authors at the Department of Psychiatry III of the same university. No data were acquired specifically for the purpose of this study, and the data acquisition protocol was that of clinical routine. Image size was 256 × 256 × 19 voxels, voxel size 0.9 × 0.9 × 6.5 mm (slice thickness, 6 mm). The images were acquired with TR 3,300 ms, TE 96 ms, a flip angle of 90°, a bandwidth of 1,220 Hz/pixel, and a field of view of 240 × 240. Diffusion at three b-factors was measured: 0, 200, and 1,000. For the computation of ADC maps we used the software provided by Siemens. The ADC signal is given by the formula $\mathrm{ADC} = -\log(y_k/y_0)/k$, where k is the value of the b-factor, and y_k, y_0 are the signal values taken with the b-factor and without additional gradient (Moritani et al., 2005). Rearranging, one obtains $\log(y_0/y_k) = k \times \mathrm{ADC}$, which shows the transformed signal on the left to be a linear function of the b-factor through the ADC signal. Hence, the ADC signal is recovered by regressing the transformed signal on the b-factors. The ADC may be computed in each voxel according to the formula $\mathrm{ADC} = (\mathbf{b}'\mathbf{b})^{-1}\mathbf{b}'\mathbf{t}$, where $\mathbf{b}$ is the column vector of the b-factors (200, 1000)′, and $\mathbf{t}$ the vector of the natural logarithm of the ratio of the signal without additional gradient and the signal at the b-factors.

The reference sample was formed by selecting 40 consecutive patients admitted to the psychiatric wards of the University Clinic and who presented no demonstrable organic or severe psychiatric illness (a diagnosis of schizophrenia, bipolar illness, drug abuse, severe depression, obsessive-compulsive disorder as well as psychotic conditions of any kind were exclusion criteria). Internistic or neurological conditions also led to the exclusion from the reference sample. Patients had to be between the age of 18 and 40 (T2-weighted images often present age-related hyperintensities in the white matter that we did not want to be present in the reference sample). The clinical scanning routine included images in other modalities which were used to exclude subjects presenting anomalies. This reference sample was designed to represent a population of inpatients without abnormal brain tissue. The sample was composed of 9 males, 31 females. Average age was 35.2 (standard deviation 10.8). The primary diagnosis of 23 patients was mild to moderate depression, 11 patients were diagnosed as suffering from an anxiety disorder, 5 from a personality disorder, and 1 had no psychiatric illness.

To obtain the analyses in the case studies, all ADC maps were resampled to obtain voxels of size 2.25 × 2.25 × 6.5 mm, and registered to a T2 template through a 12-parameter affine normalization, followed by 16 iterations of a nonlinear normalization using 1 × 2 × 1 basis functions [Ashburner and Friston, 1999]. After segmentation, the selection of the segment of interest was obtained by setting the threshold for white and gray matter to 0.2 and excluding voxels classified with 0.15 or larger to CSF. A mask was formed by selecting all voxels for which at least 30 observations were available. This number was chosen so that the mask included most cerebral matter (see Fig. 1). The remaining observations were considered as missing and imputed by generating them from a normal distribution with parameters estimated from the existing observa-

tions [stochastic imputation, Little and Rubin, 2002, pp. 64–66]. Males contributed 112,615 voxels on average in way of observations (std. dev. 1,894), females 113,878 (std. dev. 2,513, not significant). The volumes were subsequently smoothed using an isotropic Gaussian kernel of FWHM = 8 mm and centered. The voxel-by-voxel mean and variance were estimated as described in the previous section. The variance reference image was subsequently smoothed with a Gaussian kernel of FWHM = 4.5 × 4.5 × 13 mm (2 × 2 × 2 voxels). Random field theory parameters were estimated using SPM software and the mask generated by the segmentation process. Smoothness was estimated from data. Thereafter, the appropriate degrees of freedom were estimated separately for tests for maxima and minima using simple cross-validation, giving values of 35 and 219, respectively.

The scans of patients suffering from neurological disease were taken for diagnostic purposes in the Department of Radiology of the University of Ulm. ADC maps and T2-weighted images were acquired as part of the clinical routine; the patient suffering from neoplastic disease was also examined with a contrast-enhanced T1-weighted scan sensitive to disruptions of the blood-brain barrier, as it is good clinical practice in such cases. These patients too were under the care of one or more of the authors, but were not inpatients of the psychiatric clinic. The ADC maps of these patients were processed as the scans of the reference sample. In particular, the same normalization procedure was used; to minimize the impact of circumscribed lesions, a smaller set of basis functions than the SPM default was used, but more nonlinear iterations were carried out. Registration was carried out with respect of the mean of the preregistered reference sample scans. Images were tested using the degrees of freedom previously established by cross-validation.

Case Studies

We present data on three patients suffering from cerebral infarction at different stages of evolution and glioblastoma, tested on the reference sample of $N = 40$. The data were provided without a consistent set of high-definition images that would have enabled the optimal normalization and segmentation procedures of the experimental setting. However, they are fairly typical of the conditions of the clinical setting and are analyzed here for illustrative purposes.

RESULTS

To substantiate the claims made in this study, we will adopt the following strategy. First, we will estimate the achieved α levels of random field theory thresholds for t random fields with the help of Monte Carlo simulations. As in previous studies [Nichols and Hayasaka, 2003], these thresholds will be shown to deliver conservative tests for low smoothness and degrees of freedom. We will then show that cross-validation (used as a subsampling strategy), but not the bootstrap, delivers correct estimates of these achieved α levels. Hence, we will show that calibration of t random field test through cross-validation delivers tests that are more powerful than the parametric counterparts in the situations where these latter are conservative. We will also show that cross-validation can improve test power when the estimate of the variance across the image is pooled across neighboring voxels. Finally, we will explore the reasons for the failure of the bootstrap, showing that sampling without replacement is the key for the success of cross-validation.

Estimation of Achieved Significance Levels

Here we show that simple cross-validation is an effective method for estimating the empirical distribution of maxima of t random fields and of the achieved significance levels for given thresholds. In contrast, the bootstrap fails in this task. For this reason, the bootstrap can neither be used to provide rejection regions computed from the empirical distribution nor to calibrate a t random field.

The box plot in the top left corner of Figure 2 shows the achieved α levels from the parametric thresholds. Note that the achieved α levels are lower than the nominal thresholds (shown as black dots for reference) since the random field theory thresholds refer to continuous random fields and not to lattice approximations as in digital images [Nichols and Hayasaka, 2003; Worsley, 2005]. Below it a histogram of the maxima of the t parametric maps observed in the new random fields generated to calculate the achieved significance levels. Because of the large number of samples, this histogram is a very good approximation to the real distribution of the maxima of random fields with these characteristics. Note that the whole first column on the left of Figure 2 refers to the distribution of maxima of the fields and the achieved α levels of the parametric threshold ϑ_{RFT}, while the remaining columns refer to estimates of these quantities. To emphasize this fact, the background of the first column is shaded.

The task of the resampling methods is that of estimating these achieved significance levels (the data summarized by the boxes) from the reference sample alone. One can see that both bootstrap methods (central columns) provided poor estimates of the real achieved significance levels. In the case of the simple bootstrap (centre left), the estimates tend to be above the real achieved significance levels. The leave-out bootstrap (centre right), which is more similar to cross-validation in the way it calculates the significance levels from data that were not used to standardize the fields, gives too high estimates of the achieved significance levels.

A clue as to why this happened is given by the histograms of the maxima of the fields $\max(\{z^{*b}\})$ generated by the bootstrap resampling scheme as in Eq. (5), shown just below the box plot of the estimates of the α levels. One can see that these distributions are very different from the real distribution of the maxima as displayed in the bottom left. The shape of the histogram of the simple bootstrap is strongly bimodal; the mode of the histogram of the leave-out bootstrap has a mode at around 10, instead of 6. By contrast, simple cross-validation (right column) does a fine job of estimating the achieved α levels. The histogram of the

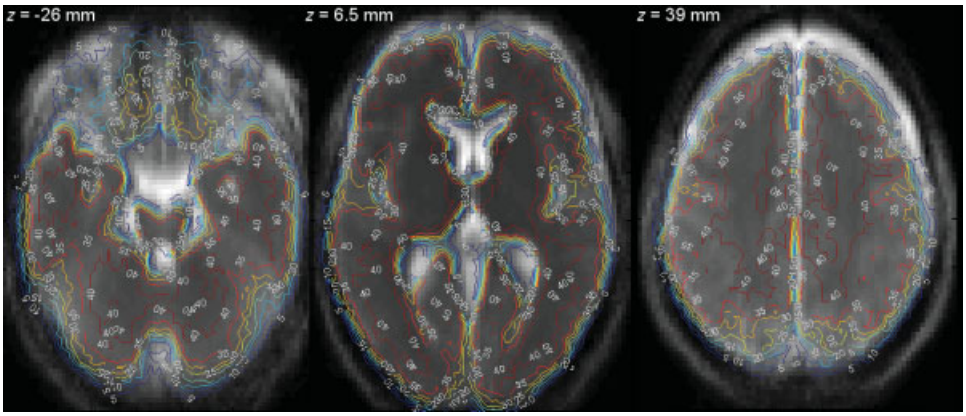

Figure 1.

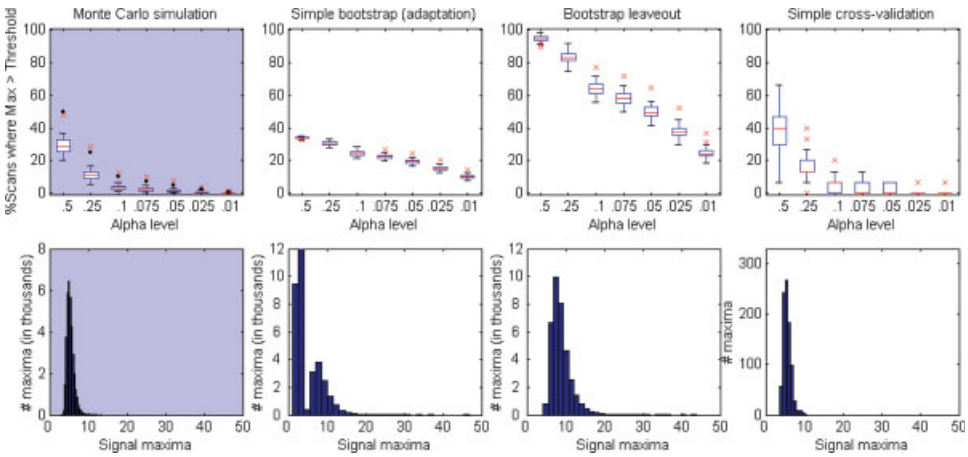

Figure 2.

maxima of the random field is similar in shape to the target histogram in the left column, and has the same mode.

The overall shape of the histograms of the scores of the maximum values in the bootstrap allows making some conjectures as to the causes of the problem. The bimodality of the histogram of the adaptation method suggests that fields that were included in the sample used to calculate the standardization parameters $\hat{\mu}^{*b}$, $\hat{\sigma}^{*b}$ give different maximum value scores than those that were left out. Thus, it seems that the leave-out property of the cross-validation resampling scheme is important for its success. It cannot be the only one, however, since if this were the case leave-out bootstrap would give good results. The next simulation shows that the decisive difference between the bootstrap and cross-validation is the form of resampling, with or without replacement.

This simulation creates a situation in which bootstrap and cross-validation differ only in the way resampling is carried out: with or without replacement. On the cross-validation side, one needs to be able to draw many samples to compute estimates based on a comparable number of obser-

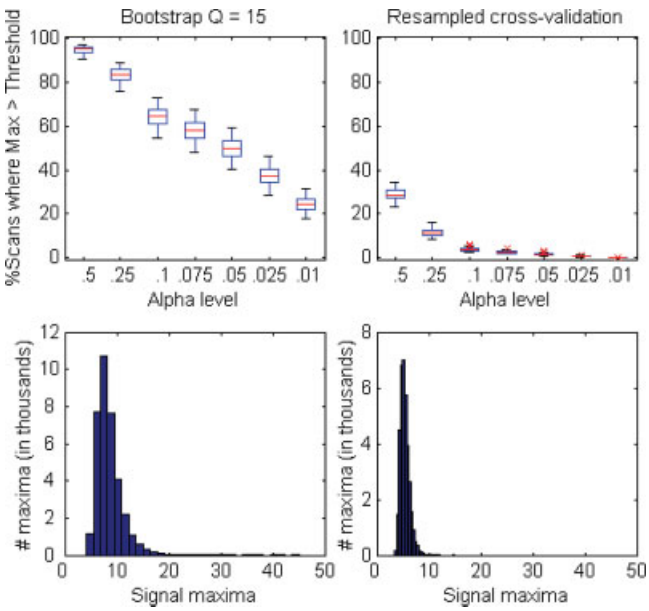

Figure 3.

vations as in the bootstrap. This suggests the use of resampled cross-validation. On the bootstrap side, one needs to set the size of the samples to the same as in resampled cross-validation. To accomplish this, artificial samples of $N = 40$ fields were created, but the task was set to estimating the achieved significance levels of reference samples of $Q = 15$ (this situation would not arise in practice, but is investigated here for theoretical purposes). The bootstrap $Q = 15$ of the left column of Figure 3 is a leave-out bootstrap in which bootstrap samples of 15 fields are chosen with replacement from a pool of 40. The resampled cross-validation in the right column of Figure 3 has same resample sizes of 15, but in the formation of the cross-validation blocks, resampling takes place without replacement. The task is again that of estimating the distribution of maxima and achieved α levels of the parametric thresholds shown in the left column of Figure 2. One can see that the bootstrap $Q = 15$ suffers from the same problem as ordinary leave-out bootstrap, while resampled cross-validation does a fine job of estimating the achieved significance levels and produces a maximum value scores histogram that is very similar to the correct one.

Since the bootstrap has a proven track record as a variance estimator, one may wonder why it fails in the present situation. Here, it is important to estimate the extreme tails of the empirical distribution accurately, a task in which the bootstrap is known to perform poorly [Efron and Tibshirani, 1993]. Intuitively, the reason for the success of sampling without replacement is that the estimation of the empirical distribution of the maxima depends on sampling the single maximum value, and the probability of doing so in each sample is higher if sampling occurs without replacement. One may expect the problem to be more acute if the original dataset is small. This intuition will be made more precise in the final discussion.

Calibration of Artificial Random Fields Using Cross-Validation

The purpose of the next simulations is to demonstrate the effectiveness of random field threshold calibration using cross-validation.

In the top row of Figure 4, one can see that the parametric thresholds display different degrees of conservativeness: the smaller the smoothing kernel, the more conservative the test, since the achieved α levels are further below their nominal values, displayed as black dots for reference. In the middle row of Figure 4, one can see that calibrated α levels are much closer to the nominal levels, delivering more powerful tests. Even if now the error rates are closer to the correct levels, they are slightly lower than the correct ones since they refer to a reference sample of $N - 1$ instead of N. However, since the degrees of freedom are now random, there is more variance in the achieved error rates than in the parametric case, so that in some trials the achieved error rates are too high. In the bottom row of Figure 4 one can see the gains of using cross-validated calibration in terms of lower thresholds (the parametric thresholds of random field theory are shown as pink dots for reference).

The major advantage of adopting calibrated thresholds results from exploiting spatially pooled variance estimates (see Fig. 5). The left column shows the effect of smoothing the variance estimate with an isotropic Gaussian kernel of FWHM = 2 voxels. One can see that the achieved error rates are substantially unaltered, but the thresholds reduction is

Figure 1.

The background brain image in grey is the average of the 40 normalized ADC maps of the reference sample. The contour plots in color show the extent of the brain classified as grey or white matter in a number of ADC maps. At one extreme, the dark red line delimits the voxels which were classified in all ADC maps (number of observations = 40). At the other extreme, the dark blue line delimits the voxels classified in at least 5 observations only. Lines in graded colors show intermediate numbers of observations. In the study, the number of observations was set to 30, corresponding to the orange contour line.

Figure 2.

Top row: box plots of achieved significance levels at thresholds specified by random fields theory (left), and their estimates obtained with the simple (centre-left), the leave-out bootstrap (centre-right), and cross-validation (right). The nominal levels of the random field theory tests are shown in the first column as black dots for reference. The boxes are drawn at the lower and upper quartile values, with a line at the median value. The whiskers extend from each end of the box covering the extent of the rest of the data, with the exception of outliers marked as "x". Bottom row: histograms of maxima in standardized volumes. In the left column, the boxes drawn from the artificial data represent the target distribution of the maxima of the random fields that the bootstrap methods and cross-validation attempt to estimate.

Figure 3.

Top row: box plots of estimated achieved significance levels. Bottom row: histograms of maxima in standardized volumes. Left and central columns shows results from the "bootstrap $Q = 15$" and resampled cross-validation. Right column shows results from simple cross-validation. The results in the right column look different as they are calculated from a smaller data set, but otherwise agree closely with the population data of the left column of Figure 2.

much more substantial than without variance smoothing (bottom row). For tests at the 0.05 level, the reduction amounts to almost 3 points, compared with less than 1 point without smoothing.

The right column Figure 5 shows how the use of 5-fold cross-validation trades off some bias in the estimation of the error rates for a reduction in the variance, especially at nominal significance levels equal or less than 0.1. As a result, only in outlier trials are the achieved error rates higher than the bound specified by the nominal significance levels. The calibrated thresholds, although not as low as in simple cross-validation, are substantially lower than those that would be used without variance estimate pooling; they also show less variance across trials.

These simulations show that cross-validation is an effective means to calibrate random field testing. The improvements of cross-validation are most marked when it is used in combination with locally pooled variance estimates. As in nonparametric fits, n-fold cross-validation exhibits a bias-variance tradeoff that can be used to improve the confidence that the bounds specified by the nominal α level on Type I error rates are respected.

Calibration of ADC Maps

In this section we shall apply cross-validation to real ADC maps, with the purpose of showing that calibrated rejection regions are more accurate than those computed with the parametric random field theory test. ADC maps from $N = 40$ individuals satisfying the criteria for the reference sample were collected (see "Material and methods"). Figure 6, left, displays the superimposed histograms of the standardized reference sample. The histograms are somewhat more variable than what one might expect from normal data; notably, some of the right tails of the histograms appear to be slightly thicker and longer that the left tails. A permutation test of the maxima of the voxel-by-voxel skewness obtained by multiplying a random subset of the standardized volumes by -1 at each of 12,000 steps demonstrates only a trend towards positive skewness ($P = 0.098$). However, the visual impression of longer right tails is confirmed by a permutation test of the symmetry of the maxima and minima. Figure 6, right, shows the permutation distribution of the t statistic on the absolute extrema of the standardized volumes, after multiplication of a random subset of the volumes by -1. If the extrema were distributed symmetrically around the zero mean, this statistic would be close to zero. The value of the statistic of the observed volumes (indicated by a red mark on the abscissa), however, is located well to the right of the bulk of the permutation distribution, indicating that maxima are further away from the zero mean than the minima ($P < 0.0001$).

Even if the moderate skewness would normally raise no concern in a univariate context, its marked effect on the average difference between the extrema may be relevant here, since parametric random field tests are based on the distribution of extrema [Salmond et al., 2002]. While a comprehensive analysis of the specific distributional characteristics of ADC maps and their effects on parametric random field tests are outside the scope of this paper, we will note that the calibrated random field model is sensitive to the asymmetry of the extrema, while the symmetric rejection regions of the parametric test cannot be correct. When applied to the whole reference sample of $N = 40$ (using a vector of seven alpha levels between 0.2 and 0.01), simple cross-validation estimated the calibrated degrees of freedom to a value of 35 for maxima, 219 for minima, against the theoretical value of 39. The corresponding parametric thresholds for tests at the nominal level $P = 0.05$ are ± 5.43, and those calibrated by cross validation 5.53 and -4.70. These data suggest that at even moderate degrees of skewness the symmetric rejections regions of parametric test may lead to inaccurate results.

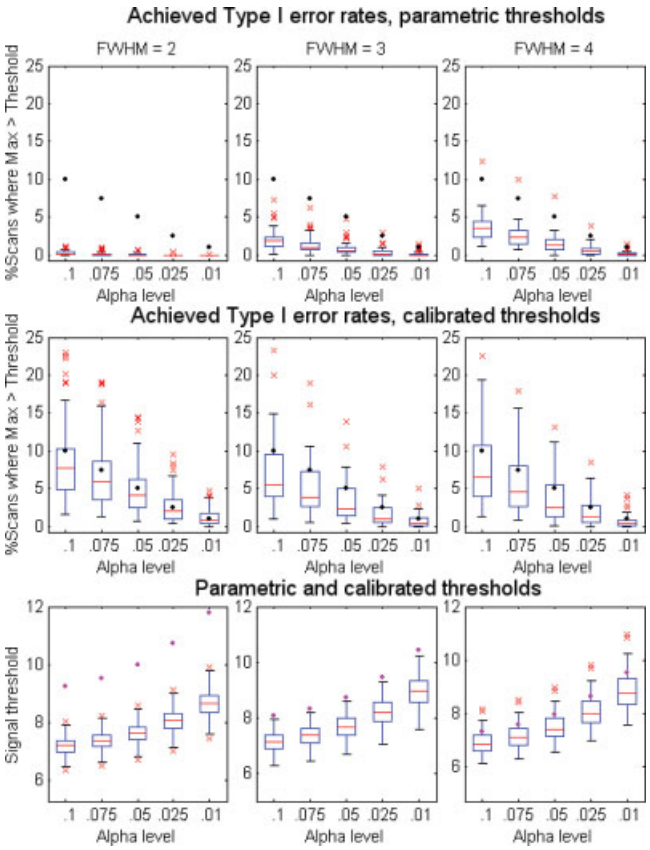

Figure 4.
Top row: box plots of achieved significance levels of random field tests compared with the nominal levels (black dots). The tests were carried out on reference samples of size $N = 15$ with increasing volume smoothness (from left to right).

Case Studies

Case study 1: Old cerebral infarction

The signal changes caused by infarction evolve with the progression of the lesion. In the initial stages, water diffu-

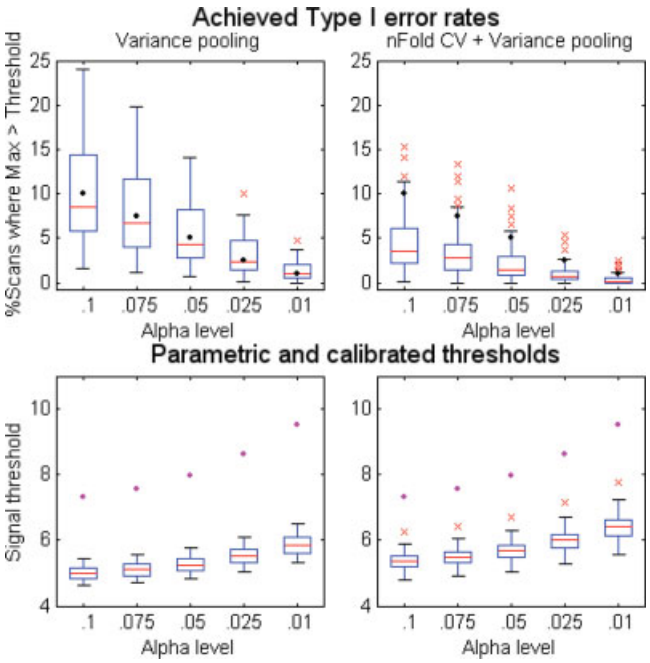

Figure 5.
Calibration of achieved significance levels. In the top row, box plots of the achieved significance levels. Bottom row shows box plots of the thresholds obtained by calibration. The thresholds obtained by parametric random field tests without calibration are shown as pink spots. On the abscissa, nominal significance levels are given.

sion decreases, presumably due to the reduced diffusion of the larger proportion of water molecules in the intracellular compartment prior to cellular death. At later stages, cicatricial tissue and the removal of necrotic parts increase the diffusion coefficients [Atlas, 2002]. Figure 7 displays transversal slices of T2-weighted structural images, normalized ADC maps, and the statistical maps thresholded at a significance level of 0.05 (two-tailed) obtained by testing the normalized ADC maps with the reference sample of $N = 40$ with the random field model calibrated by cross-validation.

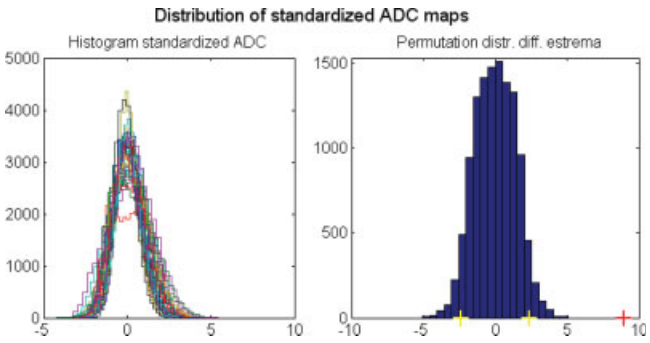

Figure 6.
In the first diagram on the left, histograms of the standardized signal of the 40 ADC maps used in this study are shown. The histograms are drawn as lines to allow superimposing them in a single diagram. On the right, permutation distribution of the average difference between the extrema of standardized volumes (see text for details) is shown. Yellow marks on the abscissa indicate the 5th and 95th percentiles of the permutation distribution of the extrema, and the red mark indicate the value of the observed sample.

This patient, a woman of 43, suffered from an extensive right-hemisphere infarction after aneurysm clipping several years previously. Over the years, the damaged tissue underwent extensive reorganization, resulting in the formation of scar tissue, contraction of the brain parenchyma and compensatory enlargement of the ventricle on the same side (T2-weighted image on the left and normalized ADC maps in the centre). Enlargement of the ventricle is detected by the segmentation process, which classified as belonging to the CSF compartment large portions of tissue that are usually classified as white or gray matter. This additional CSF-classified voxels are displayed in gray in the parametric maps on the right, superimposed on an averaged T1-weighted template. The statistical analysis displays the enhanced diffusion of water in the cicatricial tissue in red and yellow.

Case study 2: Fresh cerebral infarction

The second case study consists of images of a fresh stroke. Here we expect areas with reduced diffusion coefficient. The statistical maps were thresholded at an alpha level of 0.05 (one-tailed).

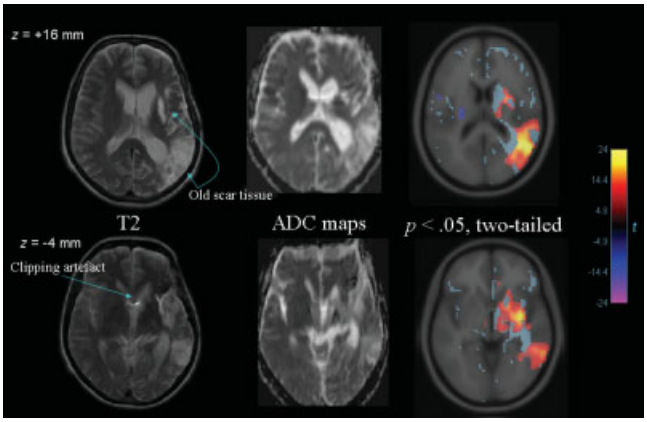

Figure 7.
Extensive right-hemisphere infarction in T2-weighted images (left column) and ADC maps (centre column). Right column, parametric maps of the lesion as overlay on the normal MNI T1 template. The transversal slices show voxel-by-voxel t scores thresholded at corrected significance level of $P < 0.05$, two-tailed. In yellow and red, voxels over threshold for maxima are shown; in blue, voxels under threshold for minima are given. In grey, voxels that were classified by the segmentation process as CSF in this ADC map and that were part of the reference images. All images are shown in the neurological convention with the patient's right to the right side.

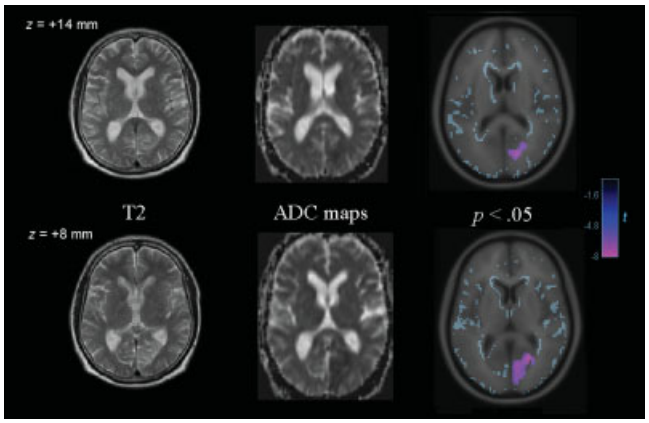

Figure 8.
Occipital-calcarine fresh stroke in T2-weighted images (left column) and ADC maps (centre column). Right column shows parametric maps of the lesion. The transversal slices show voxel-by-voxel *t* scores thresholded at corrected significance level of $P < 0.05$, one-tailed.

These images were taken from an 81-years-old male patient suffering from acute hemianopsia in the left hemifield. In the T2-weighted images on the left, taken 24 h after onset of symptoms, a fresh infarct is barely detectable as an area of hyperintense signal with ill-defined contours in the occipital-calcarine cortex (Fig. 8). The normalized ADC maps in the central column show reduced diffusion in the same zone. The reduction of the signal is visible in the statistical maps on the right.

Case study 3: Glioblastoma

The third case study consists of images obtained of a glioblastoma, a highly malignant and invasive tumor of the nervous tissue, where we will look for enhancements of the diffusion coefficients due to the neoplastic tissue and the surrounding edema. The figures were built as in Case study 1, with the exception that the T2-weighted structural images were replaced by T1-weighted images with contrast medium to better highlight the neoplastic lesions following any breakdown of the blood-brain barrier. The statistical maps were thresholded at an alpha level of 0.05 (one-tailed) (Fig. 9).

This patient (a male of 63 years of age) suffered from a glioblastoma (histologically confirmed). The primary lesion is located in the right-medial temporal lobe. The tumor compresses the right ventricle, leading to disturbances of CSF flow and dilatation of the temporal horn, as apparent in the T1-weighted images on the left. Even if located in areas usually occupied by white or gray matter, these voxels were correctly classified by the segmentation algorithm as CSF (in gray in the parametric maps on the right, which were superimposed on a normal average T1 template). The upper row shows images before radiotherapy. At this stage, the tumor did not show the contrast enhancement usually seen with this type of lesion, and the ADC values were isointense to the surrounding tissue. The statistical map, however, pinpoints a second lesion in the left temporal lobe, which is barely visible in the T1-weighted images. The bottom row shows a follow-up study 4 months after completion of radiotherapy. The tumor tissue now shows extensive contrast enhancement, possibly associated with radiation necrosis or ongoing tumor growth (left). The lesion is enlarged with further compression of the right lateral ventricle and dilation of the temporal horn. The statistical map on the right depicts the tumor as well as the surrounding edema in yellow and red.

DISCUSSION

There were both practical and methodological motivations behind the study, and we discuss them in turn

Application to Single-Image Tests

In this study a reference distribution model that makes contact with well-established medical practices in establishing departures from normal reference values was introduced, and applied to the assessment of images taken for clinical purposes. Cross-validated calibration was shown to increase the power of random field models testing when the variance cannot be estimated globally, extending the scope of resampling methods to applications where permutation cannot be applied. In simulations with artificial *t* random fields, random fields calibrated with cross-validation were shown to deliver more powerful tests than parametric methods when the smoothness of the image is low (see Fig. 4). Calibration of random field models also adapts the test to departures from ideal distributional assumptions on the

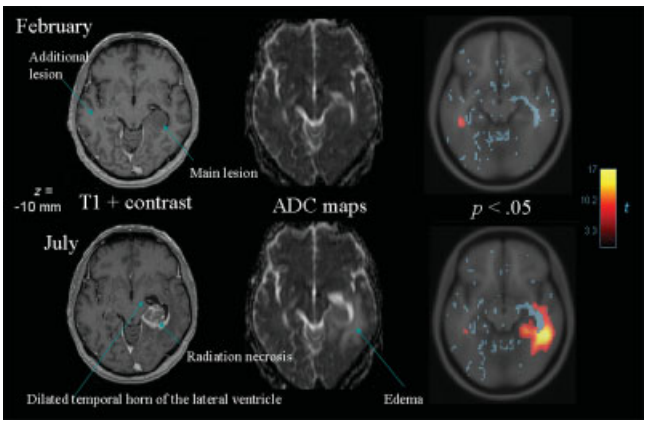

Figure 9.
Glioblastoma in T1-weighted images with contrast medium (left column) and ADC maps (centre column). Right column shows parametric maps of the lesion. The transversal slices show voxel-by-voxel *t* scores thresholded at corrected significance level of $P < 0.05$, one-tailed.

signal (at least as long as the real distribution is the same in all examined voxels), delivering guarantees on its validity that are otherwise not available in parametric tests. Evidence was presented that ADC maps in this study are characterized by an asymmetric distribution of extrema. This is an important practical aspect when testing the signal of single images, since here the effect of departures from normality assumptions are most marked.

An important advantage of the semiparametric approach presented here is the possibility of combining calibration with local pooling of the voxel-by-voxel variance estimates in t field models. The voxel-by-voxel mean and variance estimators may well be optimal in the univariate setting, but in this context the object of the estimate is the random field as a whole. Indeed, if the variance is spatially correlated, pooling its variance must reduce the risk of the estimate. The situation is somewhat analogous to variance function estimation in generalized regression, where variance function models or kernel methods have been shown to be superior to estimating the variance by using the replications at each observed point [Carroll and Ruppert, 1988; Davidian and Giltinan, 1995]. In the present context, we showed that in artificial volumes the cut-off thresholds can be lowered by about a third while maintaining the same achieved significance levels (see Fig. 5).

Finally, the application of cross-validated calibration was illustrated by successfully detecting the pathological signal in ADC maps of patients affected by stroke and glioblastoma.

Methodology: Superiority of Cross-Validation Relative to the Bootstrap

It was shown here that the bootstrap, unlike cross-validation, cannot be applied to the estimation of the distribution of extrema of random fields (see Fig. 2). Simulations were made that indicate that the type of resampling (with replacement vs. without replacement) is responsible for the success of cross-validation procedures (see Fig. 3). We discuss here why cross-validation succeeds where the bootstrap fails.

Bickel and Freedman [1981] drew attention on the failure of the bootstrap to provide consistent estimates of statistics of extrema in a univariate setting. More generally, the resampled statistic must satisfy mild smoothness conditions to be successfully estimated by the bootstrap [see Beran and Ducharme, 1991 for a useful discussion].

Cross-validation is a specific instance of a more general class of resampling schemes, characterized by resampling without replacement. Politis et al. [1999] summarize a number of results demonstrating that resampling without replacement is a more robust estimation technique than the bootstrap; in particular, sampling without replacement provides consistent estimators under the conditions where the bootstrap was showed to fail. n-Fold cross-validation, for subsamples increasingly close to N, illustrates the sequential principle on which the asymptotic efficacy of sampling without replacement is based: each of the B subsamples, of size Q, are indeed subsamples of size Q from the true model for samples of size N. Thus, under very mild conditions, the sampling distribution of the chosen statistic gets closer to the true distribution of samples of size N as Q gets close to N [Politis et al., 1999, p. 40]. The bootstrap, on the other hand, attempts to approximate the model directly with resamples of size N drawn i.i.d. from the empirical sample (that is, without replacement), and this attempt succeeds under conditions that are more restrictive than in sampling without replacement.

Note, however, that the smaller the ratio Q/N, the closer the bootstrap-Q gets to sampling Q observations without replacement, since the probability that an observation is sampled twice becomes increasingly smaller. Thus, asymptotically when $Q/N \to 0$ as $N \to \infty$, the differences between the bootstrap-Q and resampled cross-validation vanish. More generally, it is known that the bootstrap becomes consistent when resampling proceeds on subsets rather than on the original dataset, since it is asymptotically equivalent to sampling without replacement. The rate at which this asymptotic result is reached, however, varies depending on the statistic under consideration, so that general results are not available [Politis et al., 1999, p. 51]. This means that while two methods compared in Figure 3, bootstrap-Q and resampled cross-validation, are asymptotically equivalent, the simulations demonstrate that the rate of convergence to the limiting distribution of extrema of random fields is slower for the bootstrap than for cross-validation. Because of asymptotic equivalence, at some point with increasing N the bootstrap technique too will deliver good results. The rationale for using cross-validation or related sampling without replacement techniques is then that they are safe techniques, which in the present context succeed even if bootstrap techniques fail.

As is usually the case with the bootstrap, results do not easily generalize. While our target here was single image tests, our results invite to caution when estimating the empirical distribution of extrema of random fields, especially if the dataset is small. However, since not all conceivable bootstrap schemes were analyzed here, further research should address the effectiveness of alternative algorithms. In other situations, a larger set of alternative resampling schemes may be available, such as resampling from chunks of adjacent residuals [Efron and Tibshirani, 1993]. Even if successful, however, alternative resampling schemes may be complex and therefore impractical, especially if compared with permutation tests.

The semiparametric approach presented here is justified by the small size of the reference sample, but can reasonably be expected to be more efficient if the reference sample is close to the distributional assumptions of the random field, i.e., if the random field model is approximately valid. In this case, estimation of the empirical null by resampling leads to gains in power, especially if the estimate of the variance is pooled across neighbouring voxels. Of course, if the reference sample is far from being normally distributed and is large, then one may and should abandon the random field model and resort to the appropriate quantiles of the

empirical distribution of the extrema obtained by sampling without replacement, or to a nonparametric estimate of $\vartheta(\alpha)$ viewed as a function of α in a reasonable domain interval.

REFERENCES

Ashburner J, Friston KJ (1999): Nonlinear spatial normalization using basis functions. Hum Brain Mapp 7:254–266.

Atlas SW (2002): Magnetic Resonance Imaging of the Brain and Spine, 3rd ed. (2 vols.). Philadelphia: Lippincott, Williams & Wilkins.

Beran R, Ducharme G (1991): Asymptotic Theory for Bootstrap Methods in Statistics. Montréal: Les Publications CRM, Université de Montréal.

Bickel PJ, Freedman DA (1981): Some asymptotic theory for the bootstrap. Ann Stat 9:1196–1217.

Carroll RJ, Ruppert D (1988): Transformation and Weighting in Regression. London: Chapman and Hall.

Casella G, Berger RL (2002): Statistical Inference, 2nd ed. Pacific Grove, CA: Duxbury.

Colliot O, Bernasconi N, Khalili N, Antel SB, Naessens V, Bernasconi A (2006): Individual voxel-based analysis of gray matter in focal cortical dysplasia. Neuroimage 29:162–171.

Davidian M, Giltinan DM (1995): Nonlinear Models for Repeated Measurement Data. London: Chapman and Hall.

Efron B, Tibshirani RJ (1993): An Introduction to the Bootstrap. London: Chapman and Hall.

Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Mazziotta JC (1997): Human Brain Function. London: Academic Press.

Friston KJ (1996): Statistical parametric mapping and other analyses of functional imaging data. In: Toga AW, Mazziotta JC, editors. Brain Mapping: The Methods. New York: Academic Press.

Friston KJ, Frith PF, Liddle PF, Frackowiak RSJ (1991): Comparing functional (PET) images: The assessment of significant change. J Cereb Blood Flow Metab 11:690–699.

Green PJ, Silverman BW (1994): Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. London: Chapman and Hall.

Holmes AP, Blair RC, Watson JDG, Ford I (1996): Nonparametric analysis of statistic images from functional mapping experiments. J Cereb Blood Flow Metab 16:7–22.

Kassubek J, Hupperz HJ, Spreer J, Schulze-Bonhage A (2002): Detection and localization of focal cortical dysplasia by voxel-based 3D MRI analysis. Epilepsia 43:596–602.

Lange N (1999): Statistical procedures for functional MRI. In: Moonen CTW, Bandettini PA, editors. Functional MRI. Berlin: Springer.

Le Bihan D (1991): Molecular diffusion nuclear magnetic resonance imaging. Magn Reson Q 7:1–30.

Little RJA, Rubin DB (2002): Statistical Analysis with Missing Data, 2nd ed. New York: Wiley.

Loh W-Y (1987): Calibrating confidence coefficients. J Am Stat Assoc 82:155–162.

Moritani T, Ekholm S, Westesson PL (2005): Diffusion-Weighted MR Imaging of the Brain. Berlin: Springer.

Nichols TE, Hayasaka S (2003): Controlling the familywise error rate in functional neuroimaging: A comparative review. Stat Methods Med Res 12:419–446.

Nichols TE, Holmes AP (2001): Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum Brain Mapp 15:1–25.

Politis DN, Romano JP, Wolf M (1999): Subsampling. Berlin: Springer.

Salmond CH, Ashburner J, Vargha-Khadem F, Connelly A, Gadian DG, Friston KJ (2002): Distributional assumptions in voxel-based morphometry. Neuroimage 17:1027–1030.

Stejskal EO, Tanner JE (1965): Spin-diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys 42:288–292.

Stone M (1974): Cross-validatory choice and assessment of statistical predictions. J R Stat Soc B 36:111–146.

Wahba G (1990): Spline Models for Observational Data. Philadelphia: Society for Industrial and Applied Mathematics.

Woermann FG, Free SL, Koepp MJ, Ashburner J, Duncan JS (1999a): Voxel-by-voxel comparison of automatically segmented cerebral grey matter—A rater-independent comparison of structural MRI in patients with epilepsy. Neuroimage 10:373–384.

Woermann FG, Free SL, Koepp MJ, Sisodiya SM, Duncan JS (1999b): Abnormal cerebral structure in juvenile myoclonic epilepsy demonstrated with voxel-based analysis of MRI. Brain 122:1202–1208.

Worsley KJ (2005): An improved theoretical *P* value for SPMs based on discrete local maxima. Neuroimage 28:1056–1062.

Worsley KJ, Marrett S, Neelin P, Evans AC (1992): A three-dimensional statistical analysis for CBF activation studies in human brain. J Cereb Blood Flow Metab 12:900–918.

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC (1996): A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73.

Worsley KJ, Liao C, Aston J, Petre V, Duncan GH, Morales F, Evans AC (2002): A general statistical analysis for fMRI data. Neuroimage 15:1–15.