# Correspondence

## Permutation Testing Made Practical for Functional Magnetic Resonance Image Analysis

Matthew Belmonte* and Deborah Yurgelun-Todd

***Abstract*—We describe an efficient algorithm for the step-down permutation test, applied to the analysis of functional magnetic resonance images. The algorithm's time bound is nearly linear, making it feasible as an interactive tool. Results of the permutation test algorithm applied to data from a cognitive activation paradigm are compared with those of a standard parametric test corrected for multiple comparisons. The permutation test identifies more weakly activated voxels than the parametric test, always activates a superset of the voxels activated by this parametric method, almost always yields significance levels greater than or equal to those produced by the parametric method, and tends to enlarge activated clusters rather than adding isolated voxels. Our implementation of the permutation test is freely available as part of a widely distributed software package for analysis of functional brain images.**

***Index Terms*—fMRI, permutation test, resampling, software.**

### I. Introduction

The application of neuroimaging techniques has become increasingly prevalent as a method for characterizing the neural substrate of cognitive and emotional processing. Functional magnetic resonance imaging (fMRI) using blood oxygenation level-dependent (BOLD) contrast produces large spatial arrays of BOLD time series. The anatomical localization of brain activations is derived from statistical procedures that assess each of these series individually and construct a spatial map of the results. Although sensitive statistical methods for handling this multiple-comparison problem have been proposed, their implementation has not kept pace with advances in imaging technology and computational power that allow such large volumes of data to be generated and processed. This paper presents a practical implementation of the permutation test, a computationally intensive method that improves on more traditional, parametric estimates of significance in the context of multiple comparisons [2].

In general, statistical methods applied to regional brain activation aim to answer two fundamental questions. The first of these is the omnibus question: does the experimental manipulation have any significant effect as measured by the test statistic? The second question is the more relevant one for imaging research: given that the omnibus test is satisfied, what specific regions or coordinates give rise to the effect?

One of the most straightforward methods of addressing this question of localization is to treat each voxel as an experiment in its own right, repeating a single parametric test in a voxel-by-voxel manner. Although this method controls Type I error within any particular voxel considered by itself, it fails to control Type I error over the image as a whole. For example, with a typical $\alpha$ level of 0.05, one of every twenty voxels will be identified as activated. These spurious signals impair localization of real brain activation by cluttering the image.

The convention in many fMRI studies has been to address the problem of false positives by using a stringent $\alpha$, often computed as a Bonferroni correction for the number of comparisons. While this strategy does reduce the number of false positives to an acceptable level, it also eliminates genuinely activated voxels whose signals happen to be weak.

The shortcoming of Bonferroni correction and like methods becomes more apparent when one considers the spatial structure of fMRI data. The strategy of independent testing with $\alpha$ adjustment assumes that the fMRI time series at each voxel are uncorrelated. This assumption is not met, for several reasons:

A COLON, NOT A PERIOD.

1) fMRI techniques measure not neural activity *per se*, but BOLD contrast. Thus, the observed signal represents some convolution of neural activity with local vascular structure, which may extend into neighboring voxels.
2) Activated brain regions may encompass multiple neighboring voxels.
3) Anatomical connections between distant regions may produce correlated activities in those regions.
4) Physical limitations of MR methods give the output of the scanner an appreciable point spread function, blurring neighboring voxels into each other.

In addition to the issue of spatial dependencies among observations, there is the assumption, implicit in most parametric tests, that the observations are drawn from a normal distribution. The neurophysiological and vascular processes that lead to BOLD contrast are not well understood, and departures from the normal distribution may exist. Such violations of the parametric assumption exert their greatest effect in the distribution's tails—exactly the regions most important for significance testing.

### II. The Resampling Method

The overly conservative nature of the Bonferroni correction was noted by Blair and Karniski [2]. As an alternative, they proposed the permutation test. The permutation test is one of a family of methods known collectively as resampling procedures [6]. Taking advantage of the speed of modern computing systems, these methods construct an explicit, nonparametric model of the actual distribution from which a set of observations has been drawn.

The reasoning behind the permutation test can be developed from an examination of more traditional tests. A common parametric method of fMRI data analysis involves correlating the observed time series at each voxel with an ideal time series. This ideal series can be viewed as an indicator variable that describes the experimental condition *at* the time of each observation. The simplest form of such a series is a square wave whose value is one during the experimental condition and zero during the control condition.

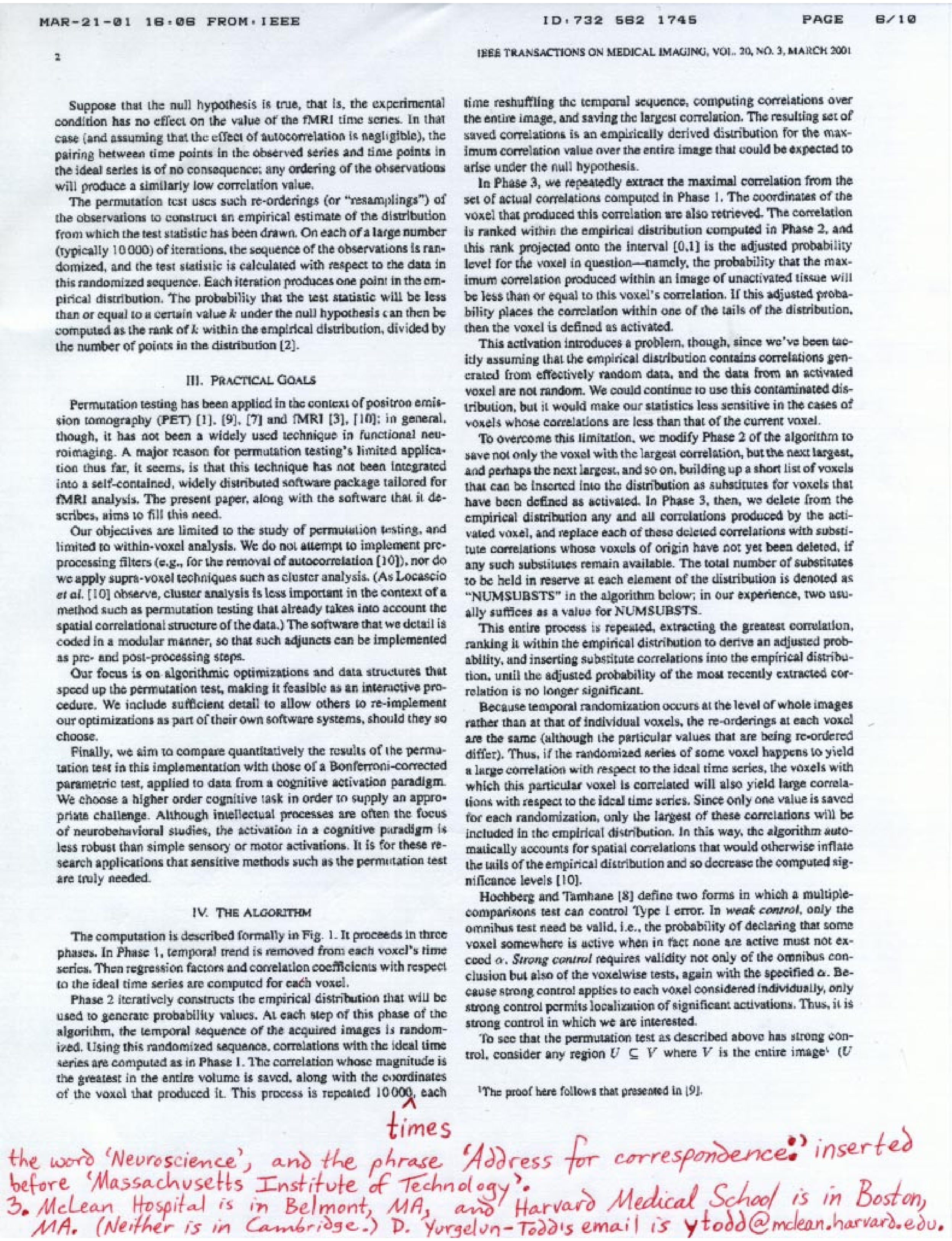
Manuscript received October 20, 2000; ① revised January 8, 2001. This work was supported by grants from the National Alliance for Autism Research (NAAR) and from the National Alliance for Research on Schizophrenia and Depression (NARSAD) and by a Boston University Graduate Student Research Fellowship (GSRF). Computational facilities were provided by the MIT Student Information Processing Board. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. W. Vannier. *Asterisk indicates corresponding author.*

*M. Belmonte is with the Cognitive Neuroimaging Laboratory of McLean Hospital Brain Imaging Center and the Boston University Program in Behavioral Neuroscience, ② Massachusetts Institute of Technology, room 14E-303, Cambridge, MA 02139-4307 USA (e-mail: belmonte@mit.edu).

D. Yurgelun-Todd is with the Cognitive Neuroimaging Laboratory of McLean Hospital Brain Imaging Center and Harvard Medical School, Cambridge, MA ③ 02139-4307 USA(e-mail: belmonte@mit.edu).

Publisher Item Identifier S 0278-0062(01)02783-5.

1. (This is merely a question, not necessarily a correction.) No revision to this manuscript was requested; is it still the practice to note a 'revised' date?
2. My physical mailing address is separate from the two institutions that were involved in this work. The current wording suggests that the Boston University Program in Behavioral Neuroscience is part of MIT, which is plainly false. I suggest a period (.) following OVER



Suppose that the null hypothesis is true, that is, the experimental condition has no effect on the value of the fMRI time series. In that case (and assuming that the effect of autocorrelation is negligible), the pairing between time points in the observed series and time points in the ideal series is of no consequence; any ordering of the observations will produce a similarly low correlation value.

The permutation test uses such re-orderings (or "resamplings") of the observations to construct an empirical estimate of the distribution from which the test statistic has been drawn. On each of a large number (typically 10 000) of iterations, the sequence of the observations is randomized, and the test statistic is calculated with respect to the data in this randomized sequence. Each iteration produces one point in the empirical distribution. The probability that the test statistic will be less than or equal to a certain value $k$ under the null hypothesis can then be computed as the rank of $k$ within the empirical distribution, divided by the number of points in the distribution [2].

## III. Practical Goals

Permutation testing has been applied in the context of positron emission tomography (PET) [1], [9], [7] and fMRI [3], [10]; in general, though, it has not been a widely used technique in functional neuroimaging. A major reason for permutation testing's limited application thus far, it seems, is that this technique has not been integrated into a self-contained, widely distributed software package tailored for fMRI analysis. The present paper, along with the software that it describes, aims to fill this need.

Our objectives are limited to the study of permutation testing, and limited to within-voxel analysis. We do not attempt to implement preprocessing filters (e.g., for the removal of autocorrelation [10]), nor do we apply supra-voxel techniques such as cluster analysis. (As Locascio *et al.* [10] observe, cluster analysis is less important in the context of a method such as permutation testing that already takes into account the spatial correlational structure of the data.) The software that we detail is coded in a modular manner, so that such adjuncts can be implemented as pre- and post-processing steps.

Our focus is on algorithmic optimizations and data structures that speed up the permutation test, making it feasible as an interactive procedure. We include sufficient detail to allow others to re-implement our optimizations as part of their own software systems, should they so choose.

Finally, we aim to compare quantitatively the results of the permutation test in this implementation with those of a Bonferroni-corrected parametric test, applied to data from a cognitive activation paradigm. We choose a higher order cognitive task in order to supply an appropriate challenge. Although intellectual processes are often the focus of neurobehavioral studies, the activation in a cognitive paradigm is less robust than simple sensory or motor activations. It is for these research applications that sensitive methods such as the permutation test are truly needed.

## IV. The Algorithm

The computation is described formally in Fig. 1. It proceeds in three phases. In Phase 1, temporal trend is removed from each voxel's time series. Then regression factors and correlation coefficients with respect to the ideal time series are computed for each voxel.

Phase 2 iteratively constructs the empirical distribution that will be used to generate probability values. At each step of this phase of the algorithm, the temporal sequence of the acquired images is randomized. Using this randomized sequence, correlations with the ideal time series are computed as in Phase 1. The correlation whose magnitude is the greatest in the entire volume is saved, along with the coordinates of the voxel that produced it. This process is repeated 10 000 times, each time reshuffling the temporal sequence, computing correlations over the entire image, and saving the largest correlation. The resulting set of saved correlations is an empirically derived distribution for the maximum correlation value over the entire image that could be expected to arise under the null hypothesis.

In Phase 3, we repeatedly extract the maximal correlation from the set of actual correlations computed in Phase 1. The coordinates of the voxel that produced this correlation are also retrieved. The correlation is ranked within the empirical distribution computed in Phase 2, and this rank projected onto the interval [0,1] is the adjusted probability level for the voxel in question—namely, the probability that the maximum correlation produced within an image of unactivated tissue will be less than or equal to this voxel's correlation. If this adjusted probability places the correlation within one of the tails of the distribution, then the voxel is defined as activated.

This activation introduces a problem, though, since we've been tacitly assuming that the empirical distribution contains correlations generated from effectively random data, and the data from an activated voxel are not random. We could continue to use this contaminated distribution, but it would make our statistics less sensitive in the cases of voxels whose correlations are less than that of the current voxel.

To overcome this limitation, we modify Phase 2 of the algorithm to save not only the voxel with the largest correlation, but the next largest, and perhaps the next largest, and so on, building up a short list of voxels that can be inserted into the distribution as substitutes for voxels that have been defined as activated. In Phase 3, then, we delete from the empirical distribution any and all correlations produced by the activated voxel, and replace each of these deleted correlations with substitute correlations whose voxels of origin have not yet been deleted, if any such substitutes remain available. The total number of substitutes to be held in reserve at each element of the distribution is denoted as "NUMSUBSTS" in the algorithm below; in our experience, two usually suffices as a value for NUMSUBSTS.

This entire process is repeated, extracting the greatest correlation, ranking it within the empirical distribution to derive an adjusted probability, and inserting substitute correlations into the empirical distribution, until the adjusted probability of the most recently extracted correlation is no longer significant.

Because temporal randomization occurs at the level of whole images rather than at that of individual voxels, the re-orderings at each voxel are the same (although the particular values that are being re-ordered differ). Thus, if the randomized series of some voxel happens to yield a large correlation with respect to the ideal time series, the voxels with which this particular voxel is correlated will also yield large correlations with respect to the ideal time series. Since only one value is saved for each randomization, only the largest of these correlations will be included in the empirical distribution. In this way, the algorithm automatically accounts for spatial correlations that would otherwise inflate the tails of the empirical distribution and so decrease the computed significance levels [10].

Hochberg and Tamhane [8] define two forms in which a multiple-comparisons test can control Type I error. In *weak control*, only the omnibus test need be valid, i.e., the probability of declaring that some voxel somewhere is active when in fact none are active must not exceed $\alpha$. *Strong control* requires validity not only of the omnibus conclusion but also of the voxelwise tests, again with the specified $\alpha$. Because strong control applies to each voxel considered individually, only strong control permits localization of significant activations. Thus, it is strong control in which we are interested.

To see that the permutation test as described above has strong control, consider any region $U \subseteq V$ where $V$ is the entire image[1] ($U$

[1] The proof here follows that presented in [9].

the word 'Neuroscience', and the phrase 'Address for correspondence:' inserted before 'Massachusetts Institute of Technology'.
3. McLean Hospital is in Belmont, MA, and Harvard Medical School is in Boston, MA. (Neither is in Cambridge.) D. Yurgelun-Todd's email is ytodd@mclean.harvard.edu.

```
PHASE 1: Compute the experimental correlations.
Let $V$ be an array of all the image samples, with spatial
coordinates as the three major indices and time as the
minor index.
$S_{\mathrm{IDEAL}} := \sum_{t=0}^{T-1} \mathrm{IDEAL}_t$
$SS_{\mathrm{IDEAL}} := (\sum_{t=0}^{T-1} \mathrm{IDEAL}_t^2) - \frac{S_{\mathrm{IDEAL}}^2}{T}$
CORR $:= \emptyset$
for $(x, y, z) \in$ VOLUME
    $V_{xyz} :=$ DETREND$(V_{xyz})$
    $S_{xyz} := \sum_{t=0}^{T-1} V_{xyzt}$
    $SS_{xyz} := (\sum_{t=0}^{T-1} V_{xyzt}^2) - \frac{S_{xyz}^2}{T}$
    $r :=$ corr(IDEAL, $V_{xyz}$) using $S_{\mathrm{IDEAL}}, SS_{\mathrm{IDEAL}}, S_{xyz}, SS_{xyz}$
    CORR := CORR $\cup \{(r, (x, y, z))\}$

PHASE 2: Compute the null distribution.
initialise SEQ to the series $[0, T-1]$
initialise NULLDIST to $\emptyset$
for $m \in [0, M-1]$
    SEQ := SHUFFLE(SEQ)
    initialise all correlation+coordinate lists $R$ to null
    for $(x, y, z) \in$ VOLUME
        for $t \in [0, T-1]$  PERMUTATION$[t] := V_{xyz\mathrm{SEQ}_t}$
        $r :=$ corr(IDEAL, PERMUTATION) using $S_{\mathrm{IDEAL}}, SS_{\mathrm{IDEAL}}, S_{xyz}, SS_{xyz}$
        if $|R| \geq$ NUMSUBSTS $\longrightarrow$
            let $r_{min}$ be nin$(\{|r'| \mid \exists (i, j, k)\ (r', (i, j, k)) \in R\})$
            if $|r| > r_{min} \longrightarrow$
                delete from $R$ the element that produced $r_{min}$
                insert $(r, (x, y, z))$ in sorted position in $R$
            fi
        [] $|R| <$ NUMSUBSTS $\longrightarrow$
            insert $(r, (x, y, z))$ in sorted position in $R$
        fi
    NULLDIST := NULLDIST $\cup R$

PHASE 3: Compute adjusted probabilities.
initialise $p_{xyz}$ to $\frac{1}{2}$ for all $x, y, z$
repeat
    Find $(r, (x, y, z)) \in$ CORR such that $\forall (r', (x', y', z')) \in$ CORR $|r'| \leq |r|\}$
    CORR := CORR $- \{(r, (x, y, z))\}$
      $p_{xyz} := \frac{\mathrm{RANK}(r, \mathrm{NULLDIST})}{|\mathrm{NULLDIST}|}$
      if $p_{xyz} \leq \frac{\alpha}{2}$ or $p_{xyz} \geq 1 - \frac{\alpha}{2} \longrightarrow$
          for all lists $R \in$ NULLDIST such that $\exists r$ head$(R) = (r, (x, y, z))$
              do head$(R) = (r, (i, j, k))$ such that $p_{ijk} \neq \frac{1}{2} \longrightarrow$
                  $R :=$ tail$(R)$
              od
      fi
  until $\frac{\alpha}{2} < p_{xyz} < 1 - \frac{\alpha}{2}$
```

ROMAN LETTER 'T',
NOT GREEK LETTER
UPSILON.

Fig. 1. The optimized permutation test algorithm.

may be as small as a single voxel.) For all $i$, $0 \leq i < M$, the $i$th element in the distribution of maximal correlations over this region must be less than or equal in magnitude to the $i$th element in the distribution of maximal correlations over the entire image $V$: $|R_U[i]| \leq |R_V[i]|$.

So the rank of any particular correlation $r$ against the distribution $R_U$ must always be at least as far into the tail as the rank of $r$ against $R_V$. Therefore, the significance levels calculated for voxels within $V$ as a whole can never exceed those for voxels within $U$ by itself.

## V. Time Complexity

Let $T$ be the number of points in each time series, $N$ the number of voxels in the volume analyzed, and $M$ the size of the empirical distribution. The regression procedures used in Phase 1 are linear in $T$, and insertion of each of the resulting correlations into an ordered binary tree is $\log(N)$. Each of these steps is performed once for each voxel analyzed. Thus, Phase 1 is $O(N\,T + N\log(N))$.

In Phase 2, the step of randomization of the time series is again linear in $T$ [5]. Computing correlations for each voxel takes time proportional to $N \times T$, as above. Saving the maximal correlation in an ordered binary tree takes $\log(M)$ time. Each of these steps is performed once for each of the $M$ entries in the empirical distribution. Thus, Phase 2 is $O(M\,N\,T + M\log(M))$.

In Phase 3, extracting the maximal correlation from the binary tree that was constructed in Phase 1 takes $\log(N)$ time. Finding the rank of this correlation within the empirical distribution that was constructed in Phase 2 takes $\log(M)$ time. Deleting from the empirical distribution the correlations produced by the activated voxel again takes $\log(M)$ time. (The tree structure that stores the distribution is keyed both on correlation values and on coordinates, so that both the rank-ordering and coordinate-deletion operations can he performed in logarithmic time.) Each of these steps is performed once for each voxel whose activation is significant. This number of significantly activated voxels generally will be some fraction of $N$, the total number of voxels analyzed. Thus, Phase 3 is $O(N\log(N) + N\log(M))$.

DELETE THE TWO COMMAS THAT SURROUND 'THUS'.

The entire algorithm is, thus, $O(M\,N\,\Upsilon + N\log(N) + M\log(M)$ ~~$+ N\log(M))$~~. For any practical values of these parameters the first term dominates, and so the algorithm behaves linearly in $M$, $N$, and $T$. It is, thus, desirable to hold $M$, $N$, and $T$ to the minimal values necessary to produce adequate statistics. $T$ cannot be usefully reduced, else we would lose a great deal of information in the correlations. $M$ should not be taken much lower than $10^4$; otherwise the empirical distribution, and the resulting adjusted probabilities, would be too grainy. $N$, though, can be decreased without losing any of the information that we care about, by computing the set of voxels that represent brain tissue and applying the procedure only to those voxels. (We implement this selection of brain voxels using a combination of intensity thresholding and region-growing.) This restriction has the beneficial side effect of making the computed distribution more representative of the tissue under consideration.

AVOID HYPHENATING A HYPHENATED WORD

## VI. Comparision to Other Methods

Our system of excluding activated voxels from the empirical distribution is a variant of the "step-down" procedure, in which the empirical distribution is recomputed at every step of Phase 3 so as to exclude voxels that have been activated. Since complete recomputation of the distribution is computationally expensive, Holmes *et al.* [9] proposed a hybrid between this procedure and single-step permutation testing: rather than computing one adjusted probability and excluding one activated voxel on every iteration of Phase 3, their method iterates in "jumps" in which all voxels with adjusted probabilities less than $\alpha$ are declared activated and excluded *en masse*. While this hybrid procedure is more sensitive than the plain permutation test, it still fails to identify as many voxels as the complete step-down version of the algorithm. Although as observed by Holmes *et al.* it would be impractical to recompute the empirical distribution on every iteration of Phase 3, we note that such *de novo* recomputation can be avoided using the substitution procedure that we outline above, in which replacements for deleted elements are precomputed in Phase 2 and applied in Phase 3.

In a method developed for analysis of PET images, Heckel *et al.* [7] suggest ordering the sequence in which randomized permutations are used, in such a way as to minimize the number of changes between successive permutations. This method facilitates incremental computation of correlation values since the only time points that need to be examined are the ones whose corresponding points in the ideal time series differ from those of the previous permutation. The sums computed for the previous permutation can then be updated accordingly, reducing the total number of memory accesses. Although the problem of finding such a minimum-change ordering for a given set of permutations is NP-hard in the general case (indeed, it reduces fairly directly to the well-known Travelling Salesman Problem), Heckel *et al.* note that a good approximate solution can be computed in $M^2$ time. They note further that in typical implementations, constant factors are such that the savings in the $M\,N\,T$ term may more than make up for this extra $M^2$ term.

The savings produced by Heckel's optimization depend strongly on the computing hardware on which the algorithm is implemented. Modern developments such as high-speed cache memory, pipelined instruction processing, and vectorization make multiple memory references much less of a performance concern. On our test platform, a 500 MHz Alpha 21 164 processor (Compaq Computer Corporation, Houston, TX) with 8K Level 1 cache and 96K level 2 cache, using optimized code generated by the Digital UNIX C compiler, a test of this optimization made no appreciable difference in execution time. Heckel's method is a valuable option, but only for certain types of computing systems.

DELETE THE TWO COMMAS THAT SURROUND 'THUS'

## VII. Test Cases

For an evaluation of the permutation test on actual fMRI data, we used images from eight normal, right-handed subjects collected during performance of the Stroop color-word interference task [12]. All subjects signed an informed consent approved by the McLean Hospital Institutional Review Board, and had no history of head injury, psychotropic medication, seizure disorder, substance abuse, or other neurological or psychiatric disorder. Functional scans were acquired on a Signa 1.5-T system (General Electric, Milwaukee, WI) modified by Advanced NMR Systems (Wilmington, MA). The ~~task~~ consisted of two 30-s blocks of the task, alternating with three 30-s periods of rest. During the task periods, subjects viewed color names projected onto a screen in front of the scanner in an incongruent color. Subjects were asked to say the name of the display color; to succeed, they had to suppress the tendency to read the color name (e.g., the word "BLUE" written in red). Stimuli were displayed in lines of six. During the 30-s activation period, six of these lines were presented for 4.5 s each, with 0.5-s interstimulus intervals.

REPLACE 'TASK' WITH 'PARADIGM'

Fifty T2*-weighted single-shot gradient-echo coronal images in each of 12 slices were acquired (effective ΥE: 40 ms; ΥR: 3 s; flip angle: 90 °; 64 × 64 matrix; in-plane resolution: 3.125 mm; slice thickness: 6 mm; and slice gap: 1 mm). Slices were perpendicular to the plane defined by the anterior and posterior commissures, and covered the region from the central sulcus to the tip of the frontal pole. The images were motion-corrected in $k$-space using the Decoupled Automated Rotation and Translation (DART) algorithm [11].

THE 'T' IN 'TE' AND 'TR' SHOULD BE THE ROMAN LETTER 'T', NOT THE GREEK LETTER UPSILON.

Brain tissue was distinguished from nonbrain areas of the image using an automated procedure that examined the median-filtered histogram of voxel intensities averaged over the entire time series, and selected the minimum value of this histogram within the interval between the brain and skull-air peaks. The intensity associated with this

1. THE CIRCLED LETTER SHOULD BE THE ROMAN LETTER 'T' (CAPITALISED AND ITALICISED), NOT THE GREEK LETTER UPSILON.

2. THE RIGHTMOST TERM IN THIS EXPRESSION IS UNNECESSARY AND SHOULD BE DELETED. THIS CHANGE WILL LEAVE THE EXPRESSION AS FOLLOWS: $O(MNT + N\log(N) + M\log(M))$

TABLE I

| SUBJECT | Age | Sex | $N_B$ | $N_P$ | $\frac{N_P}{N_B}$ | $NC_B$ | $NC_P$ | $N_{B>P}$ | $N_{P>B}$ $z_B^{B>P}$ | $z_B^{P>B}$ | $t_B, p_B$ | $z_P^{B>P}$ | $z_P^{P>B}$ | $t_P, p_P$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 31 | M | 534 | 82 | 15% | 52.65% | 51.28% | 8 | 433 | 3.44 | 2.86 | $t(439) = 3.06, p = 0.0022$ | 3.43 | 2.99 | $t(439) = 2.52, p = 0.012$ |
| 2 | 33 | F | 258 | 29 | 11% | 50.80% | 50.00% | 0 | 216 | — | 2.75 | | — | 2.97 | — |
| 3 | 21 | M | 446 | 41 | 9% | 55.90% | 54.17% | 0 | 345 | — | 2.79 | — | | 2.96 | — |
| 4 | 26 | F | 197 | 16 | 8% | 66.60% | 64.92% | 0 | 155 | — | 2.81 | | — | 2.94 | — |
| 5 | 24 | M | 395 | 29 | 7% | 42.00% | 42.30% | 32 | 233 | 3.74 | 2.76 | $t(263) = 10.87, p < 0.000001$ | 3.65 | 2.86 | $t(263) = 8.96, p < 0.000001$ |
| 6 | 19 | F | 177 | 25 | 14% | 67.29% | 65.67% | 41 | 103 | 3.42 | 2.53 | $t(142) = 12.46, p < 0.000001$ | 3.35 | 2.67 | $t(142) = 11.06, p < 0.000001$ |
| 7 | 22 | F | 297 | 78 | 26% | 57.47% | 52.09% | 0 | 230 | — | 2.81 | | — | 3.09 | — |
| 8 | 34 | F | 634 | 61 | 10% | 25.84% | 26.12% | 97 | 349 | 3.52 | 2.79 | $t(444) = 13.64, p < 0.000001$ | 3.45 | 2.92 | $t(444) = 11.40, p < 0.000001$ |

$N_B$: Total number of voxels activated with the standard, Bonferroni-corrected test.
$N_P$: Additional voxels activated with the permutation test. (All voxels activated by the Bonferroni test were also activated by the permutation test.)
$N_P/N_B$: Percent increase in activated volume between Bonferroni test and permutation test.
$NC_B$: Percent of voxels activated with the Bonferroni test that were not part of clusters.
$NC_P$: Percent of voxels activated with the permutation test that were not part of clusters.
$N_{B>P}$: Number of voxels more active with the Bonferroni test than with the permutation test.
$N_{P>B}$: Number of voxels more active with the permutation test than with the Bonferroni test.
$z_B^{B>P}$: Mean Bonferroni-test z-score of voxels that were more active with the bonferroni test.
$z_B^{P>B}$: Mean Bonferroni test z-score of voxels that were more active with the permutation test.
$t_B$, $p_B$: Significance of difference in Bonferroni-test z-scores.
$z_P^{B>P}$: Mean permutation-test z-score of voxels that were more active with the Bonferroni test.
$z_P^{P>B}$: Mean permutation-test z-score of voxels that were more active with the permutation test.
$t_P$, $p_P$: Significance of difference in permutation-test z-scores.

histogram minimum was then used as a threshold to identify putative brain voxels. Finally, a region-growing algorithm identified the largest set of connected putative brain voxels, and labeled all areas within or enclosed by this region as brain. Minor corrections to the output of this procedure were implemented by hand for each data set.

The permutation test as described above was applied to each data set, restricted to the set of brain voxels. In a parallel procedure, the standard, Bonferroni-corrected test was also applied, with the correction factor calculated as the reciprocal of the number of brain voxels. In each case, a simple square wave was used as the ideal time series against which to compute correlations. The probability values output by the permutation test and by the Bonferroni procedure were transformed to z-scores for storage and further computation.

Table I gives descriptive statistics for each test applied to each data set, as well as comparative statistics between the two tests. For each of the two test procedures, the total number of voxels activated with a two-tailed $\alpha$ of 0.05 was calculated. In all cases the permutation test activated more voxels than the Bonferroni test, and in no case did the permutation test omit any voxels that were activated by the Bonferroni test. The increase in activated volume ranged from 7% to 26%.

The total number of activated voxels that were part of clusters was also calculated, where a cluster was defined as any group of more than one adjacent voxel, within a single slice, in which the sign of the activation was uniformly positive or uniformly negative. For the purposes of this computation, diagonal adjacency was allowed. Although the permutation test always increased the total number of activated voxels, as Table I shows, it always decreased the proportion of unclustered voxels. In other words, the voxels added by the permutation test tended to be part of activated clusters rather than occurring over widespread regions throughout the image (see Fig. 2 for an example).

For the set of voxels that were activated by both tests (which in all cases equaled the set of voxels activated by the Bonferroni test), we compared the activation levels from each test, and produced a count of the number of voxels that were more activated by one test than by the other. Voxels whose probability levels differed between the two tests by an amount less than the resolution of the permutation test ($10^{-4}$ in the case of our implementation) were considered equally activated and were, thus, excluded from these counts. As can be seen in Table I, the permutation test almost always produced higher levels of activation than the Bonferroni test. Binomial tests comparing the counts were all highly significant.

DELETE COMMAS SURROUNDING 'THUS'

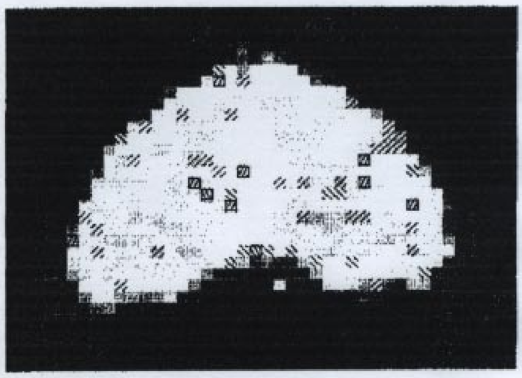


Fig. 2. Comparison of activations in the prefrontal cortex of Subject 1. Activated voxels are cross-hatched upwards, deactivated voxels downwards. Voxels activated only under the permutation test are highlighted with boxes. Left and right are reversed as per radiological convention. Note how the permutation test extends the left dorsolateral activation into the sulcus between the inferior and middle frontal gyri, and links previously unconnected deactivations in medial orbitofrontal cortex, as well as adding voxels close to other regions of activation.

We also wished to get some idea of the relative strengths of activations of the voxels identified by the two tests. In particular, we wished to determine whether the permutation test selectively identifies weakly activated voxels. To answer this question, for each data set we examined the z-scores (transformed from probability levels) of the set of voxels that were activated more strongly by one test than by the other. (In half of the cases, no voxels were activated more strongly by the Bonferroni test than by the permutation test; the table cells corresponding to these cases are, therefore, empty.) In each case two comparisons were performed, one on the set of z-scores derived from the Bonferroni test and one on the set of z-scores derived from the permutation test. In all cases analyzed, for both comparisons separately, the z-scores of the voxels that were more strongly activated by the permutation test were lower than those of the voxels that were more strongly activated by the Bonferroni test. Thus, the permutation test demonstrated an ability to identify weaker activations.

## VIII. Conclusions

The permutation test is a more sensitive analytic strategy than simple corrections for multiple comparisons, because it takes into account deviations from the normal distribution and spatial correlation in the data. This advantage was evident in the results from our test cases, in which the permutation test always activated a superset of the regions activated by the simpler method, tended to enlarge clusters rather than adding isolated voxels, almost always produced tail probabilities greater than or equal to those of the simpler test, and activated voxels whose signals were weaker than those picked up by the simpler test. Although the complete step-down version of the permutation test may at first seem computationally infeasible [9], some finesse with data structures can make the time bound of this algorithm nearly linear, and complete analysis of a data set takes only nine minutes of computer time using current technology (a 500-MHz Alpha 21 164 processor). We have implemented this optimized permutation test as part of AFNI [4], a freely available, widely used package of routines for analysis of functional brain images.

In an attempt to deal with one problem at a time, and in keeping with past work on resampling methods in functional imaging [9], [1], [3], [7], we have not attempted to account for autocorrelation in the observed time series. In cases in which the ideal time series represents a blocked design, the presence of autocorrelation within the observed time series tends to exaggerate estimates of significance slightly, since the empirical distribution is based on data whose autocorrelation has been removed by shuffling. One way to reduce this slight biasing effect would be to shuffle the observed time series in chunks, so that the order of chunks with respect to each other is randomized but the original order of the samples within each chunk is preserved. A better method, of course, would be to model the autocorrelation and remove it. Locascio *et al.* [10] present an elegant model of autocorrelation in fMRI time series, based on autoregressive and moving-average techniques originally developed for the analysis of economic data. As Locascio *et al.* observe, an implementation of such a model as part of a software package tailored specifically for the analysis of fMRI data (e.g., [4]) would be a useful tool.

We view the permutation test software presented here as one among several potential improvements and optimizations to correlation-based strategies for fMRI data analysis. We invite the addition of other optimization steps, and regard the system that we have described as a first step toward an even more sophisticated, freely available fMRI analysis package that takes advantage of currently available levels of computational speed.

ROMAN LETTER 'T', NOT GREEK LETTER UPSILON.

## References

[1] S. Arndt, T. Cizadlo, N. C. Andreasen, D. Heckel, S. Gold, and D. S. O'Leary, "Υests for comparing images based on randomization and permutation methods," *J. Cereb. Blood Flow Metab.*, vol. 16, pp. 1271–1279, 1996.

[2] R. C. Blair and W. Karniski, "Distribution-free statistical analyzes of surface and volumetric maps," in *Functional Neuroimaging: Technical Foundations*, R. W. Thatcher, M. Hallett, T. Zeffiro, E. R. Joay, and M. Huerta, Eds. San Diego, CA: Academic, 1994, pp. 19–28.

[3] M. J. Brammer, E. T. Bullmore, A. Simmons, S. C. R. Williams, P. M. Grasby, R. J. Howard, P. W. R. Woodruff, and S. Rabe-Hesketh, "Generic brain activation mapping in functional magnetic resonance imaging: A nonparametric approach," *Magn. Reson. Imag.*, vol. 15, pp. 763–770, 1997.

[4] R. W. Cox, "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Comput. Biomed. Res.*, vol. 29, pp. 162–173, 1996.

[5] R. Durstenfeld, "Algorithm 235: Random permutation [G6]," *Commun. Assoc. Computing Machinery*, vol. 7, p. 420, 1964.

[6] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia, PA: Soc. Ind. Appl. Math., 1982.

[7] D. Heckel, S. Arndt, T. Cizadlo, and N. C. Andreasen, "An efficient procedure for permutation tests in imaging research," *Comput. Biomed. Res.*, vol. 31, pp. 164–171, 1998.

[8] Y. Hochberg and A. Tamhane, *Multiple Comparison Procedures*. New York: Wiley, 1987.

[9] A. P. Holmes, R. C. Blair, J. D. G. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments," *J. Cereb. Blood Flow Metab.*, vol. 16, pp. 7–22, 1996.

[10] J. J. Locascio, P. J. Jennings, C. I. Moore, and S. Corkin, "Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging," *Human Brain Mapping*, vol. 5, pp. 168–193, 1997.

[11] L. C. Maas, B. D. Frederick, and P. F. Renshaw, "Decoupled automated rotational and translational registration for functional MRI time series data: The DART registration algorithm," *Magn. Reson. Med.*, vol. 37, pp. 131–139, 1997.

[12] C. M. MacLeod, "Half a century of research on the stroop effect: An integrative review," *Psychological Bulletin*, vol. 109, pp. 163–203, 1991.

analyses