NeuroImage 16, 538–550 (2002)
doi:10.1006/nimg.2002.1107, available online at http://www.idealibrary.com on IDEAL®

# Combining Brains: A Survey of Methods for Statistical Pooling of Information

Nicole A. Lazar,* Beatriz Luna,† John A. Sweeney,‡ and William F. Eddy*

*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; †Department of Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15213; and ‡Department of Psychiatry, University of Illinois, Chicago, Illinois 60605

Received April 24, 2001

**More than one subject is scanned in a typical functional brain imaging experiment. How can the scientist make best use of the acquired data to map the specific areas of the brain that become active during the performance of different tasks? It is clear that we can gain both scientific and statistical power by pooling the images from multiple subjects; furthermore, for the comparison of groups of subjects (clinical patients vs healthy controls, children of different ages, left-handed people vs right-handed people, as just some examples), it is essential to have a "group map" to represent each population and to form the basis of a statistical test. While the importance of combining images for these purposes has been recognized, there has not been an organized attempt on the part of neuroscientists to understand the different statistical approaches to this problem, which have various strengths and weaknesses. In this paper we review some popular methods for combining information, and demonstrate the surveyed techniques on a sample data set. Given a combination of brain images, the researcher needs to interpret the result and decide on areas of activation; the question of thresholding is critical here and is also explored.** © 2002 Elsevier Science (USA)

***Key Words:*** **combining tests; meta-analysis; multiple subjects; multiple voxels; thresholding.**

## 1. INTRODUCTION

Functional neuroimaging in its various forms is a powerful tool for the community of cognitive scientists and for clinicians interested in neuropathologies. Using methods such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), researchers have been able to greatly increase their understanding of healthy and abnormal brain function. Consideration of a simple fMRI experiment might help to clarify the sorts of data that are of interest. Suppose a neuroscientist is interested in learning about eye movement. To study this process, the scientist might put a subject in an MRI scanner, and have the subject alternate between, say, periods of fixating on a cross-hair and periods of performing memory guided saccades (that is, a target would appear in the peripheral vision; after the target was extinguished, the subject would be required to direct both eyes to the place where the target had been). While the subject does this, the experimenter might acquire a set of 5-mm-thick slices covering the entire brain every 3 s. By comparing images obtained when the subject was performing the task with those when the subject was fixating, the areas of the brain "activated" during eye movement can be localized for this individual. Typically each "slice" of the brain is acquired in a grid form, for example, in 3 by 3-mm voxels of data, so that behavior-related changes in brain activity in these voxels can be evaluated. By identifying where these task-related changes occur, a map of brain activity can be created.

A major limitation to the usefulness of these techniques has been the difficulty in integrating the results from different subjects, since these results are not single numbers, but instead whole images. There are two aspects to this difficulty. The first is spatial, that is, the brains of different people are of different sizes and shapes and have idiosyncratic structure, and it is therefore necessary to find some way of equating them, or putting them on a similar spatial scale. The standard solution to this problem has become warping the brain images of the subjects onto a common atlas, using Talairach coordinates (Talairach and Tournoux, 1988). The second issue is statistical, and involves methods of combining the functional information in the data in a reasonable fashion. Searching for a way to present the data from a neuroimaging study, researchers will often include one or two "representative" images, that show the effect of interest, or resort to *ad hoc* approaches for combining images, such as taking a simple average of the statistical maps for individual subjects. While there are circumstances where this might not be a bad thing to do (Cochran, 1954), averaging should not be blindly applied as the default combination technique. In general, the *ad hoc* methods don't allow for the accumulation of evidence arising from multiple subjects who display similar patterns of activation.



1053-8119/02 $35.00

This paper addresses only the statistical question, looking at methods for combining data from independent studies, where for our purposes here, each subject is a study. We will not discuss the advantages and disadvantages of the Talairach approach for combining brain structure, but will instead assume that the researcher has processed the data so that image sets from different subjects are comparable, and hence combinable. We do, however, caution the reader to note the tremendous loss in real spatial resolution incurred by the use of the Talairach method (Woods, 1996, gives examples of the distortions that can result from this approach).

Our goal, then, is to explore existing statistical techniques for combining data in the functional neuroimaging context. We don't think that there will be any argument with the goal itself; there is more power in data based on multiple subjects than there is in data based on one—this is the general statistical principle that we can learn more from a (well-chosen) large sample than from a small sample. Combining the data from many subjects will also result in a stronger signal; that is, if there is a reliable effect to be seen, pooling the data in a suitable way should make it more obvious. Finally, there is the issue of being able to generalize from one sample to a larger population. We extrapolate our inference from the specific subjects that we scanned, to others like them. As we pool together the data from more subjects, we are better able to make this generalization.

The rest of the paper proceeds as follows. Section 2 reviews common statistical techniques for combining information, and examines the problem of multiple tests, especially as related to the thresholding of brain images. Section 3 demonstrates the various proposals on a data set and compares the results. In the final section, we make recommendations for the optimal combination of functional brain maps across subjects.

## 2. SOME PROPOSALS FOR FUNCTIONAL NEUROIMAGING

In this section, we review a number of ideas suggested in the statistics literature for combining information across studies, where we think of each individual subject as a study. All of these methods have in common the aim of using the information acquired in different studies to get a clearer picture of the scientific question of interest, and to draw power from the accumulation of evidence. As pointed out in Hedges (1992), there are two main approaches to this problem—combining hypothesis tests and combining estimates of treatment effects. Our discussion follows this dichotomy.

### *2.1. Combining Tests*

Suppose we have $k$ independent tests of a particular null hypothesis, with values $T_1, T_2, \ldots, T_k$ of the test statistics and corresponding $P$ values of $P_1, P_2, \ldots, P_k$. Recall that the $P$ value of an observed test statistic $T$ is the probability of seeing a value as extreme as or more extreme than $T$, if the null hypothesis is true (hence it will be a small number, close to 0, for significant test statistics). In what follows, we assume that tests are one-sided, since we are looking for areas of activation. The value $k$ is the number of subjects from a functional neuroimaging study, and the methods that we describe in this section are applied on a voxel-by-voxel basis, to each of the $v$ voxels in Talairach space. Since all of the individual maps have been translated into a common coordinate system, combining in this way is statistically sensible.

Many proposals of reasonable (according to some statistical or intuitive criterion) methods for combining these independent sources have been made over the years. Perhaps the most popular is due to Fisher (1950), which uses the test statistic

$$T_F = -2 \sum_{i=1}^{k} \log P_i$$

(where log denotes the natural logarithm) and compares this to a $\chi^2$ distribution with $2k$ degrees of freedom. Large values of $T_F$ relative to the reference $\chi^2$ distribution lead to rejection of the null hypothesis of no effect (*i.e.*, activation in imaging studies). It is instructive to consider the effect that an individual $P$ value has on the summary statistic $T_F$. If $P_i$ is near 1 for some $i$, then that term of the sum is close to zero and the statistic is nearly unchanged, although the degrees of freedom increase by two. On the other hand, if $P_i$ is close to zero, then a small change in $P_i$ changes $T_F$ by $-2/P_i$.

Another early suggestion is that of Tippett (1931), to look at

$$T_T = \min_{i=1}^{k} P_i$$

and reject the null hypothesis if this value is smaller than $1 - (1 - \alpha)^{1/k}$ for a level $\alpha$ test. Wilkinson (1951) gives a generalization of this procedure, namely looking at the $r^{\text{th}}$ smallest $P$ value and rejecting the null hypothesis if this is smaller than a constant that depends on $k$, $r$, and $\alpha$. The other extreme member of this family is the test suggested by Worsley and Friston (2000)

$$T_W = \max_{i=1}^{k} P_i,$$

rejecting $H_0$ if $T_W$ is less than $\alpha^{1/k}$. Again, considering the same extreme cases, if some $P_i$ is near 1, then $T_W$ is

near 1, and that statistic is useless. $T_T$ will not be affected. If a single $P_i$ is close to zero, then the $i^{th}$ subject has no effect on the value of $T_W$, but renders $T_T$ less informative; if all the $P_i$ are close to zero then a small change in the largest of them changes $T_W$ by the same amount. If $T_W$ rejects the null hypothesis, then one can conclude that each individual subject would have rejected the null hypothesis. This is a very stringent requirement, probably too stringent for use in typical neuroimaging experiments.

Stouffer *et al.* (1949) defined

$$T_s = \sum_{i=1}^{k} \frac{\Phi^{-1}(1 - P_i)}{\sqrt{k}}$$

as a combined test statistic, where $\Phi^{-1}$ is the inverse normal cumulative distribution function. The null hypothesis is rejected for large values of $T_s$, as determined by critical points from the standard normal distribution. When $P_i = 1$, $T_s = -\infty$ and the statistic carries no information. If $P_i$ is near zero, then a small change in $P_i$ causes $T_s$ to change by $-[\sqrt{k\phi}\,(\Phi^{-1}(1 - P_i))]^{-1}$.

Another *P* value method was put forth by Mudholkar and George (1979): calculate

$$T_M = -c\sum_{i=1}^{k} \log\left(\frac{P_i}{1 - P_i}\right),$$

where $c = \sqrt{3(5k + 4)/k\pi^2(5k + 2)}$, and reject the hypothesis of no effect if $T_M$ is large, in reference to critical points of the *t* distribution with $5k + 4$ *df*. $T_M$ can be rewritten as $T_M = -c\sum_{i=1}^{k} \log P_i + c\sum_{i=1}^{k} \log(1 - P_i)$, highlighting its relation to the $T_F$ statistic defined above. For $P_i = 1$, $T_M = \infty$. If $P_i$ is near zero, then a small change in $P_i$ causes $T_M$ to change by $-c/P_i - c/(1 - P_i)$.

An *ad hoc* method that is in common use is to average the *t* statistics computed for the individual subjects. Define

$$T_A = \sum_{i=1}^{k} \frac{T_i}{\sqrt{k}}.$$

The null hypothesis is rejected for large values of $T_A$, as determined by critical points from the standard normal distribution. Under the null hypothesis, $T_A$ is approximately equal to $T_S$. Strictly speaking, $T_A$ is not a method for combining *P* values, but because of its similarity to Stouffer's method we have included it here.

Many methods based on combining *P* values and test statistics have been proposed (e.g., Pearson, 1933; see Lancaster, 1961 for a review); we have described here some of the simpler techniques. Discussion of methods for combining tests in the rest of the paper will focus on the six procedures presented: the methods due to Fisher *et al.*, Stouffer *et al.*, Mudholkar and George, and average *t* values.

This approach to combining information is appealing, since the tests that are pooled together don't have to be similar in any way, they can be based on different kinds of measurements. The methods are therefore quite general in their applicability. Furthermore, the combinations suggested by Fisher (1950) and by Mudholkar and George (1979) satisfy an optimality criterion, Bahadur efficiency (Bahadur, 1967, 1971), related to effective use of data as the number of subjects increases. (See also Littell and Folks, 1973; Berk and Cohen, 1979). On the other hand, as noted by several authors (Iyengar, 1991; Hedges, 1992; Jones, 1995), by combining tests in this way we do not obtain any information on the size, direction (from a two-sided test) or consistency of effects across the different studies. In our application, we will be interested in one-sided tests, specifically those corresponding to hypotheses regarding activation (deactivation being of somewhat lesser interest), so at least one of these drawbacks is eliminated here; it will be possible to make statements about the direction of effects, because this will be built into the analysis. Also, interpretation of the combined statistic may be difficult. Our underlying null hypothesis is that in every study, there was no effect. If we reject the null hypothesis based on a combined test procedure, what does this mean? Note that the null could be rejected on the basis of a nonzero effect in one study.

### 2.2. Combined Estimation

When all of the studies that are to be combined have similar designs and measure the outcome of interest in a similar fashion, several models exist for performing what is known in the medical and social sciences as "meta-analysis." These methods allow for inference about the direction, magnitude and consistency of effects, unlike the combining techniques outlined in the previous section. The most common meta-analysis models are the one-factor fixed effects and random effects. We now consider each of these in turn. Readers are also referred to Woods (1996) for a clear explication of these models in the imaging context.

If the studies are homogeneous in design, so that it is reasonable to think that they are measuring the same effect, a suitable statistical model is that

$$y_i = \theta + \varepsilon_i,$$

where $y_i$ is the effect observed in study $i$ (the effect, not the *t* statistic!), or one subject in an fMRI study, $\theta$ is the common mean effect, and $\varepsilon_i$ is the error in the $i$th study. The errors are usually taken to be independent

and normally distributed, with mean 0 and variance $V_i$. Under this fixed effects model, a straightforward estimate for the overall $\theta$ is

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i},$$

a weighted average of the $y_i$, in which the weights $w_i$ are inversely proportional to the estimated variance of each study. The estimate $\hat{\theta}$ is approximately Gaussian distributed, with mean $\theta$ and estimated variance $1/\sum w_i$. Based on this, it is possible to test whether $\theta$, the common mean, is zero, by defining the usual $t$-type statistic,

$$T_x = \frac{\hat{\theta}}{\sqrt{1/\sum w_i}}$$

and rejecting the null hypothesis for large values of $T_X$. When the variances of the subjects are homogeneous and the errors are independent Gaussians, the fixed effects model and Stouffer's method coincide. That is, $T_S$ is basically an unweighted version of $T_X$. An extreme subject might be one that shows the same level of activity in both conditions, and almost no variation. For this model, such a subject will have a very large weight and an effect size of 0, so will contribute nothing to the numerator of $\hat{\theta}$, but the denominator will be very large, and hence $T_X$ will be close to 0.

The random effects model may be adopted on the basis of knowledge that the studies are heterogeneous (e.g., they used different experimental designs or sampling protocols) or as a result of rejecting the hypothesis of homogeneity of the effect $\theta$ for the fixed effects model. Even when identical experimental designs and protocols are used, intersubject variability is often much greater than intrasubject variability. Furthermore, investigators are usually interested in making inference about a hypothetical population, from which the particular subjects were drawn, rather than restricting conclusions to the sample. For all of these reasons, the random effects model is preferred to the fixed effects model. The differences between the observed $y_i$ are now assumed to come from experimental error, as before, and from real differences among the studies, so that the true effect in study $i$ is a sample from a "hyperpopulation" of treatment effects. The model has the form

$$y_i = \theta_i + \varepsilon_i$$

$$\theta_i = \theta + e_i,$$

where the $\varepsilon_i$ are usually taken to be normal with mean 0 and variance $V_i$, the $e_i$ are taken to be normal with mean 0 and variance $\sigma_\theta^2$, and all the $e_i$, $\varepsilon_i$ are independent. Other distributional assumptions are possible, but the normal case is the most convenient, and hence most widely used. When $\sigma_\theta^2 = 0$, this model reduces to the fixed effects model.

As for the fixed effects model, an estimate for the expected value $\theta$ of the $\theta_i$ is a weighted average of the observed effects,

$$\hat{\theta}^* = \frac{\sum_{i=1}^{k} w_i^* y_i}{\sum_{i=1}^{k} w_i^*},$$

where $w_i^* = 1/(s_i^2 + \hat{\sigma}_\theta^2)$ and $s_i^2$ is an estimate of $V_i = E[(y_i - \theta_i)^2]$. Now there are two sources of uncertainty, whose sum has to be estimated; as a result, random effects models are more complicated than fixed effects models. Researchers have proposed different ways of estimating $\sigma_\theta^2$. One suggestion given in Hedges (1992) is

$$\hat{\sigma}_\theta^2 = S^2 - \frac{\sum S_i^2}{k},$$

where $S^2$ is the sample variance of $y_1, y_2, \ldots, y_k$. This estimator has the drawback that it could be negative, in which case the standard recommendation is to truncate to 0. Other estimators that don't have this undesirable property also exist, but these are more complicated, requiring an iterative computational procedure (see, for example, Rao, 1971; Rao and Kleffe, 1988). Any method of estimating the variance components, together with $\hat{\theta}^*$, can be used to build a test for the hypothesis that $\theta = 0$. For the random effects model, an unusual subject such as the one described for the fixed effects will influence the estimate of $\sigma_\theta^2$, but will otherwise not contribute to the outcome.

The standard errors for $\hat{\theta}$ the fixed effects estimate, tend to be smaller than those for $\hat{\theta}^*$, the random effects estimate, since the latter takes into account the variability across studies − the variance of $\hat{\theta}$ is

$$\frac{1}{\sum_i 1/V_i},$$

whereas the variance of $\hat{\theta}^*$ is

$$\frac{1}{\sum_i 1/(V_i + \sigma_\theta^2)}.$$

For $\sigma_\theta^2 = 0$, the fixed effects model, the two variances coincide, but otherwise, the random effects estimate has larger variance. In other words, the fixed effects assumption is an optimistic scenario. This is another reason that, even though they are harder to work with, it is generally

advised to use random effects, rather than fixed effects models, in meta-analysis (National Research Council, 1992). Holmes and Friston (1998) propose a convenient approach to the random effects problem where the variance is the same across subjects. They fit a simple model with a box-car reference function to each subject. The estimated activations are presented as images, which are then assessed using a one-sample $t$ test.

A major problem when performing standard meta-analyses is publication bias (also known as "the file drawer problem"; Rosenthal, 1979). This refers to the phenomenon that, in general, only studies with results which are statistically significant at, say, level $\alpha = 0.05$ (or smaller) are published. Studies with nonsignificant results often are not even submitted for publication, instead being filed away. This complicates a meta-analysis, as researchers won't have access to the unpublished studies; as a result, findings could look more significant than they really are. Various corrections for this problem have been suggested. It is worth emphasizing that when combining image data, we can avoid the serious issue of publication bias, as "studies" are individual subjects, all of whom are available to the scientist. One should not eliminate subjects, as this can artificially understate the experimental variance.

### 2.3. On Testing Multiple Voxels

Regardless of the method that is used to combine functional images across subjects, when constructing statistical brain maps, it is important to take account of the large number of voxels involved, and to adjust significance levels accordingly. This is the problem of thresholding, that is, finding the appropriate cutoff value for declaring that a voxel is active. A common method for adjusting for multiple tests in neuroimaging is the Bonferroni correction (see, for instance, Miller, 1981). Formally, the procedure may be used when the researcher is interested in performing a number of statistical tests, which are specified in advance. Suppose there are $v$ such tests of interest; in a neuroimaging study, $v$ would be the number of voxels in a statistical map (all the voxels in the brain in all slices acquired). The test statistic is calculated in the usual way. What changes is the critical point for deciding that a given test (voxel) is significant (active). For a $t$ test, instead of setting a threshold of around $t = 2$, for a significance level of $\alpha = 0.05$, some adjustment needs to be made for the fact that $v$ (independent) tests, each at level 0.05, are being carried out, in order that the overall statements of significance will still be appropriate. Each test is performed at level $\alpha/v$ (for one-sided tests; at level $\alpha/2v$ for two-sided tests) instead of $\alpha$; this guarantees that the probability of one or more false positives (tests incorrectly declared as significant) is less than $\alpha$. In imaging studies, $v$ will typically be very large, which in turn means that the thresholds used for deciding significance of a voxel, using this method, will be very large, larger than they need to be in many situations (Benjamini and Hochberg, 1995; Forman *et al.,* 1995).

A further possible complication in neuroimaging studies is that the data are available in two coordinate systems—the original acquisition space, and Talairach space. It would be wrong to calculate the Bonferroni correction on the original data and then apply this criterion to the transformed voxels in Talairach coordinates, or vice versa, because the number of voxels changes when we go from one set of coordinates to the other and the correlation structure of the data is also affected (the switch to Talairach space induces dependence among voxels beyond that already present in the data).

The Bonferroni correction works on the principle of controlling the familywise error rate at level $\alpha$, that is, of controlling the probability that even one of the null hypotheses under consideration is falsely rejected. This procedure in particular has been criticized because it lacks power, especially when the number of comparisons is large. Indeed, the inability of Bonferroni-corrected $t$ maps to reveal important effects in functional neuroimaging data has been the starting point for a collaboration between neuroscientists and statisticians (Eddy, 1997). A different approach is taken by Benjamini and Hochberg (1995), via the false discovery rate. In this method, the expected proportion of falsely rejected hypotheses, in other words, the ratio of the number of wrongly rejected hypotheses to the total number of rejected hypotheses, is controlled, leading to a gain in statistical power. The technique is easily implemented. For testing $v$ hypotheses $H_1, H_2, \ldots, H_v$, the first step is to order the $P$ values corresponding to the hypotheses from smallest to largest. Write the ordered values as $P_{(1)}, P_{(2)}, \ldots, P_{(v)}$, with $H_{(i)}$ denoting the null hypothesis with $P$ value $P_{(i)}$. Let $q$ be the desired rate of false discovery, that is, the rate of false discovery that the researcher is willing to tolerate, and let $r$ be the largest $i$ for which

$$P_{(i)} \leq \frac{i}{v}\frac{q}{c(v)},$$

where $c(v)$ differs according to the correlation structure of the tests. For independent tests, or when the tests follow a technical condition (positive dependence), one instance of which is when the noise in the data is Gaussian with nonnegative correlation across voxels, $c(v) = 1$ (Benjamini and Yekutieli, 2001); to accommodate an arbitrary joint distribution of $P$ values, use

$$c(v) = \sum_{i=1}^{v} 1/i \approx \log v + \gamma,$$

where $\gamma \approx 0.577$, Euler's constant. Then reject $H_{(1)}, H_{(2)}, \ldots, H_{(r)}$. An interesting aspect of this latter pro-

cedure is that it does not require that the test statistics be independent or even of the same kind. Since the voxels in an individual subject's brain are almost certainly not independent, this is a desirable feature.

When all null hypotheses are true, the false discovery rate is the same as the familywise error rate, hence the controlling parameter $q$ may be chosen at conventional levels for significance testing, although for many applications values of $q$ in the range of 15–20% might be deemed reasonable (Benjamini, personal communication). Furthermore, the method is adaptive, in the sense that the chosen thresholds change (become more or less conservative) with the strength of the signal. This would solve the difficult problem of finding thresholds that work for all subjects under all conditions—instead of trying to find such a threshold, which is likely to be arbitrary and *ad hoc*, the researcher can keep the tolerated level of false discoveries at a constant across subjects and experiments, and the appropriate thresholds will be determined by the data. For a more in-depth examination of this thresholding technique, with applications to neuroimaging, see Genovese *et al.* (2001).

A different slant on the thresholding problem, presented by Forman *et al.* (1995), takes explicit advantage of the specifics of the neuroimaging setting. These authors note that an area of real activation is likely to be characterized by more than a single active voxel; rather, one should observe clusters of contiguous active voxels. On the other hand, the probability is small that a contiguous cluster of voxels will cross threshold by chance, that is, in the absence of any true signal. This difference in expected behavior can be exploited by calculating the probability that various sizes of clusters will happen by chance and that such clusters will be detected when there is a true signal. Forman *et al.* describe a simulation technique for determining the probability of a false detection of a voxel, on a per-voxel basis, for different cluster sizes and values of $\alpha$, under assumptions of no spatial correlation and of the existence of spatial correlation. They also show that it is possible to pick combinations of cluster size and significance level that work like Bonferroni-type corrections, without the accompanying loss of statistical power. A consequence of the use of this technique is that clusters smaller than the chosen cluster size threshold will not be detectable. However, the increased power to discover larger groupings of contiguous active voxels is a worthy tradeoff.

Several other approaches to the thresholding question have been proposed in the imaging literature. We mention a few here, with the caveat that they all involve extra mathematical or computational resources. Genovese *et al.* (1997) take a test–retest approach, in which experiments are repeated more than once. The authors suggest several models for quantifying consistency of effects across the repetitions, what they call the test–retest reliability. Based on these models, and the estimates they yield of making different classification errors, it is possible to set thresholds for maps acquired from experimental paradigms similar to those under which the test–retest runs were obtained. This approach requires extra data collection, a cost that the researcher may not be able to bear.

Worsley (1994) has advocated the use of random field theory for the analysis of statistical maps in neuroimaging (see also Worsley *et al.*, 1996). Regions of activation are determined according to excursion sets, which are areas in the random field where a given threshold value is exceeded. The observed excursion sets can be compared to those predicted by theory, in order to determine significance. It is quite difficult to calculate the probabilities exactly, however it is possible to approximate them. This approach has the advantage of naturally taking account of the spatial nature of the data when thresholding the images and the disadvantage of being sensitive to the assumptions.

The statistical theory of permutation tests drives the work of Holmes *et al.* (1996). As with all permutation tests, the computational burden is heavy. It is also harder to incorporate spatial structure with this approach.

In what follows, we choose to look at simple thresholding methods for comparing the various combining techniques. Our focus in this study is on understanding the similarities and differences of the approaches for combining subjects in a neuroimaging studies. Toward this goal, thresholds that are easily determined are most expedient.

## 3. A DATA EXAMPLE

### *3.1. Description of an fMRI Experiment*

Eleven non-mentally-retarded autistic adults ($k = 11$), of whom two were female, performed a memory-guided visual saccade task. The average age of the subjects was 32.3 years, with a standard deviation of 9.3. All subjects were Caucasian; two of the patients were left-handed. While in the scanner, subjects carried out a test that required them to remember the spatial location of peripheral targets over time. The task was as follows. Initially, the subjects fixated on a cross hair in the center of the field of view for one second, after which a target appeared peripherally for 75 ms in one of four locations. During the presentation of the peripheral target, and for an additional 4925 ms, subjects remained fixated on the cross hair. The fixation point was then turned off and subjects had 1500 ms to look to the remembered location before the fixation target reappeared for the next trial. The control condition was a visually guided saccade task. Tasks were presented in a block design where subjects performed 10.5 cycles of memory guided saccade trials alternating with visually guided saccade trials every

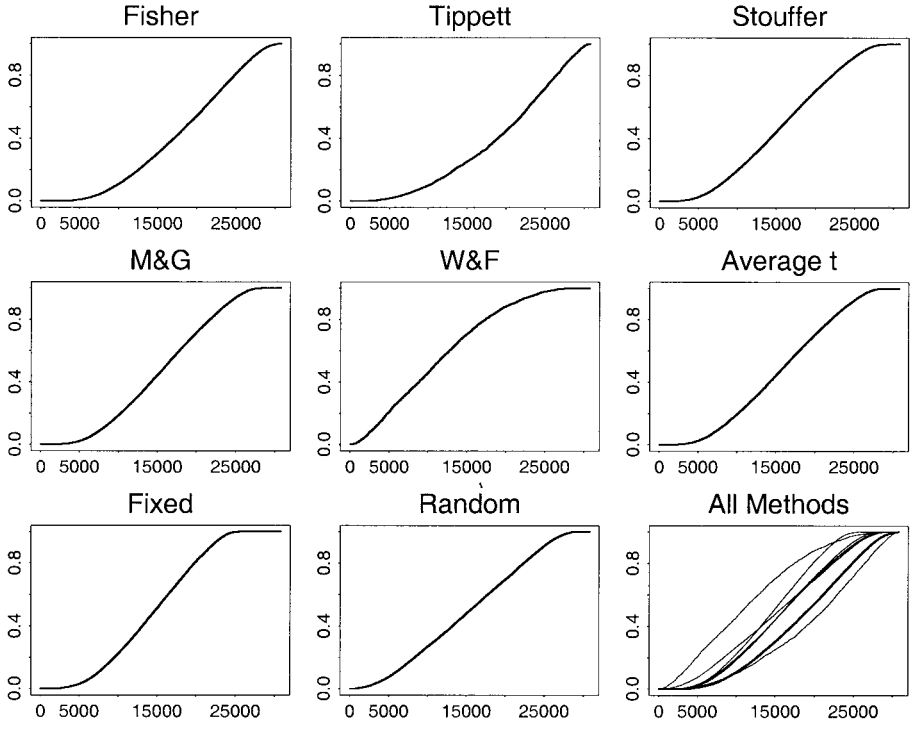


FIG. 1. Ordered $P$ values for each of the methods, Fisher, Tippett, Stouffer, Mudholkar, and George, Worsley, and Friston, average $t$ map, fixed effects model, random effects model. For reference, the thick line near the bottom in the last panel of the figure is the curve for Fisher's combining method.

30 s. In the half cycle at the end, only the memory guided saccade task was performed. Subjects were cued to a switch in trial type by a change in size of the fixation target. The neuroscientific rationale for this procedure is described in Sweeney *et al.* (1996).

The study was performed on a 3.0 Tesla Signa whole-body MR scanner (General Electric Medical Systems, Milwaukee, WI) with echo-planar imaging capability (Advanced NMR Systems, Inc., Wilmington, MA). Gradient-echo echoplanar imaging was performed using a commercial head RF coil. Acquisition parameters were: TE = 25 ms; TR = 5.0 s; single shot; full $k$-space; 128 × 64 acquisition matrix with a field of view (FOV) 40 × 20 cm. In order to cover the whole brain and cerebellum, 23 3-mm-thick oblique slices with a 2-mm gap were aligned to the base of the genu and splenium of the corpus callosum, generating isotropic 3.125 × 3.125 voxels. Subjects' heads were placed in an RF head coil packed with cushions in order to minimize head motion.

### 3.2. Comparison of Combining Methods

In this section we present the results of eight methods for combining information on the data from the eleven autistic adults. We examine the behavior of the following procedures:

- Fisher's method;
- Tippett's method;
- Stouffer's method;
- Mudholkar and George's method;
- Worsley and Friston's method;
- average $t$ method;
- fixed effects model;
- random effects model.

All methods were calculated on the data of the eleven subjects after these had been warped into Talairach space. In this comparison we do not deal explicitly with the time series nature of the data. Rather, we compare the results in each of the two conditions. Clearly, a more complete analysis would use the time course data directly; we plan to pursue this in future work. A slice from the middle of the head was chosen to represent each technique, the same slice in all eight cases.

Figure 1 shows the $P$ values for each method, ordered from smallest to largest. The $x$ axis gives the order, from 1 to 30,751, of each voxel (each of the original slices is transformed in Talairach space to have dimension 161 × 191, or $v$ = 30751 voxels in all); the $y$ axis gives the corresponding $P$ values, which range from 0 to 1.

In order to put all of the methods on an equal basis, we transformed them so that under the null hypothesis of no activity, the $P$ values would be expected to fall on a diagonal line of slope 1. For $T_T$, under the null hy-

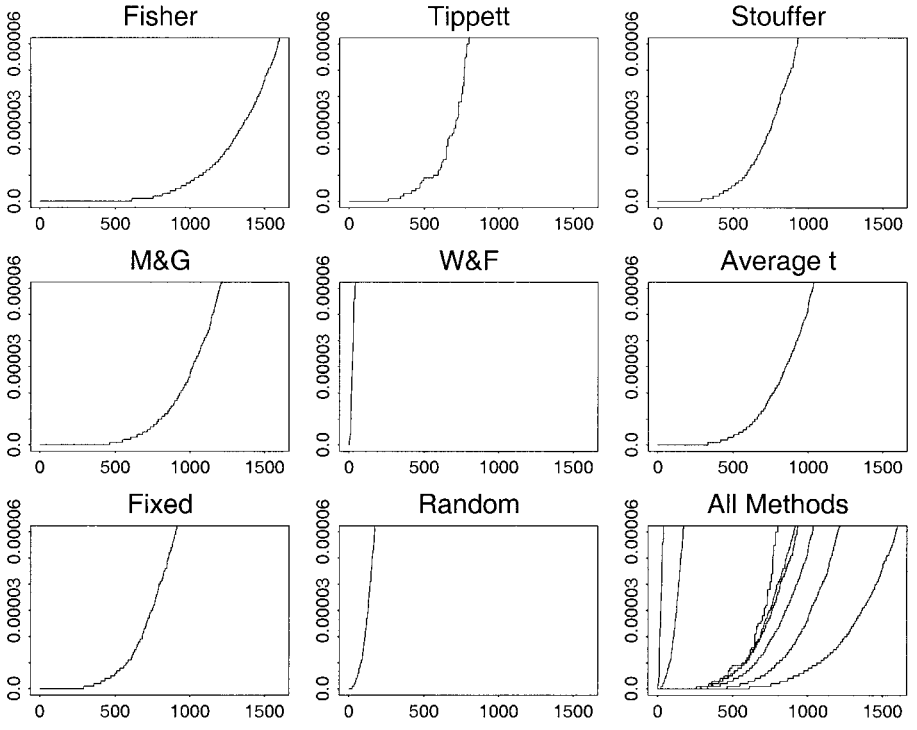


FIG. 2. Close-up of the first 1500 ordered $P$ values for each of the methods, Fisher, Tippett, Stouffer, Mudholkar, and George, Worsley, and Friston, average $t$ map, fixed effects model, random effects model. In the last panel the curves are ordered at the top edge, from left to right: Worsley and Friston, random effects; Tippett, fixed effects; Stouffer, average $t$; Mudholkar and George, Fisher.

pothesis, $(1 - T_T)^k$ would be expected to fall on a diagonal line; for $T_W$ the appropriate transformation is $1 - T_W^k$. For the other methods, calculating the significance levels, relative to the appropriate reference distributions, gives the required transformation. These are shown in Fig. 1. Departures from the diagonal are helpful for understanding the behavior of the different methods. Only about 18,000 of the 30,751 voxels in the plots represent brain; the rest are air, and these are usually clipped out. We do not clip voxels outside the brain. We find that they serve a very useful (statistical) diagnostic purpose. If there is much activation detected outside the brain, then the threshold is probably set too high (since small $P$ values are significant). If there is no activation outside the brain then the threshold is probably set too low.

Performing a one-tailed test, it is likely that the largest $P$ values, in the upper right corner of each plot, come from those voxels outside the head. Small values, at the bottom left corner of each plot, correspond to voxels for which the null hypothesis of no activation should be rejected. The question that arises from inspection of these plots, and that is at the crux of the thresholding problem discussed above, is "Where do we draw the line to decide that a given voxel is active?" If we are interested in comparing different methods, the question is complicated because it is further evident from inspection of the graph that this can be done in at least two ways, in addition to the various more complex methods described in Section 2.3. We can either pick a significance level, that is, draw a horizontal line across the plot, and all voxels below the line will be declared active (this is equivalent to doing a Bonferroni correction); or we can pick some number of voxels, that is, draw a vertical line across the plot, and all voxels to the left of the line would be considered active (this is similar to using a false discovery rate procedure).

Even without doing this comparison formally, a number of characteristics of the methods are observable in Fig. 1. The curves for $T_T$ and $T_W$ are mirror images of each other, which is reasonable since the former looks at the smallest of the $P$ values and the latter looks at the largest. All of the curves are similar in appearance—they start off near 0, rising up in a more or less uniform fashion, before leveling off at 1. The methods differ in how long the initial plateau at 0 lasts, and how quickly the value 1 is reached, but qualitatively they are the same.

Figure 2 shows a close-up of the lower-left corner of Fig. 1, with just the 1500 most significant voxels. This allows for a more detailed comparison of the methods, and reinforces our comments on the general picture. Three groups are apparent from this close-up: the random effects model and $T_W$ are in one group, with curves that increase quickly and steeply away from the plateau at 0. Next, in a rough ordering, are $T_T$, $T_S$, and $T_X$

**TABLE 1**

The $P$ value of the $r$th Ordered Voxel for Each of the Eight Methods and Various Values of $r$

| Method $r$ | Fisher | Tippett | Stouffer | Mudholkar | Worsley | Average $t$ | Fixed | Random |
|---|---|---|---|---|---|---|---|---|
| 64 | 0 | 0 | 0 | 0 | 0.00011 | 0 | 0 | 0.000006 |
| 128 | 0 | 0 | 0 | 0 | 0.000488 | 0 | 0 | 0.000032 |
| 256 | 0 | 0 | 0 | 0 | 0.001826 | 0 | 0 | 0.000145 |
| 512 | 0 | 0.000001 | 0.000009 | 0.000001 | 0.004907 | 0.000004 | 0.000007 | 0.000688 |
| 1024 | 0.000008 | 0.000118 | 0.000084 | 0.000031 | 0.015938 | 0.000059 | 0.000094 | 0.003318 |
| 2048 | 0.000194 | 0.000992 | 0.001246 | 0.000579 | 0.05361 | 0.000985 | 0.001396 | 0.012700 |
| 4096 | 0.003647 | 0.010898 | 0.014564 | 0.009697 | 0.155351 | 0.012954 | 0.016054 | 0.051341 |
| 8192 | 0.058092 | 0.064678 | 0.12207 | 0.109229 | 0.369203 | 0.118996 | 0.13533 | 0.193222 |

*Note.* Using Fisher's method to combine the data from the 11 subjects, the 64th most significant voxel has a $P$ value of approximately 0, while using the random effects model, the 64th most significant voxel has a $P$ value of around 0.000006.

(which are very close to each other), and $T_A$, which begin to rise from 0 after about 250 voxels. Finally come $T_M$ and $T_F$, which start to rise at about 500 voxels. An interesting aspect of $T_M$ is that, while its curve starts to rise from the plateau at around 500 voxels, as does that of $T_F$, the rate of increase of the curve itself is closer to that of the second group.

These general trends are confirmed by inspection of Table 1. In that table we show the significance levels corresponding to designating a fixed number of voxels to be active, with that number increasing from just 64 voxels, or 0.2% of all voxels, to 8192, or just over 25% of all voxels. For all of the methods, the most significant voxels are very highly significant, with $P$ values near 0. When averaging the $t$ maps, it does matter greatly which reference distribution is used to calculate the $P$ values. A "naive" comparison to the normal distribution with mean 0 and variance $k$, for instance, results in the 64$^{th}$ most active voxel in the average $t$ map having a $P$ value of around 0.02, which would not be significant for a Bonferroni-corrected test, at level $\alpha =$ 0.05. The effect of the averaging is to smooth out any signal that is not strong and consistent across subjects, hence the need to refer it to a distribution with smaller variance. All the methods, with the exception of $T_W$, can pick up patterns of activity that are manifested by some, but not all, subjects. The conjunction method of Worsley and Friston makes a rather extreme demand on the experiment, particularly in the light of the test-retest results of Genovese *et al.* (1997). In that paper the authors show that typical estimates of the probability that a truly active voxel will be active are on the order of 0.5; this means that using $T_W$, even if all the subjects were one and the same individual, the probability of detecting a truly active voxel is on the order of $2^{-k}$.

Figure 3 displays the 1024 most active voxels for each method. Looking at the same number of active voxels in each case provides us with a way of comparing the output of the different methods. The two methods based on extreme $P$ values, $T_T$ and $T_W$ give qualitatively different results than the other six procedures. In addition, Worsley and Friston's method produces more noise (activity outside the brain) than do the others. It must be kept in mind, however, that for this method in particular, the corresponding significance levels are high—the 1024$^{th}$ most active voxel has an uncorrected $P$ value of 0.016. The random effects model seems to be a hybrid, sharing elements with both $T_W$ and $T_F$ (for example). The images produced by Fisher *et al.* average $t$ and fixed effects are quite close to each other. The Tippett image looks significantly different from all of the rest, not so much in the areas that are picked out, as in the shape of the regions. We come back to this point in the later discussion.

Table 2 presents the second way of comparing the procedures, that is, the numbers of voxels that would be picked up by each of the methods, for a given significance level. This is equivalent to drawing a horizontal line across each of the plots in Fig. 1 at the same height. The first column in the table gives the significance level that was applied to the individual voxel tests. In parentheses, we give the corresponding familywise rate, based on doing a Bonferroni correction on 18,000 brain voxels in a single slice. Even though the significance levels appear to be very small, the large number of voxels being tested more than compensates, with the net result that the overall significance levels are not at all small. By choosing to display only a single slice, we have also been "Bonferroni friendly" in this table—in practice, the correction would be on the number of brain voxels in each slice, summed over all the slices in the brain.

Based on Table 2, it is evident that $T_F$ and $T_M$ are outlying because they pick up much more activity, especially for higher $\alpha$ levels. This is expected from the connection between the two statistics which was pointed out in Section 2.1. $T_S$ is close to the fixed effects model, also as expected, and to $T_A$; both $T_W$ and the random effects model pick out fewer active voxels than the others. The differences we observe between the fixed effects and the random effects models might seem

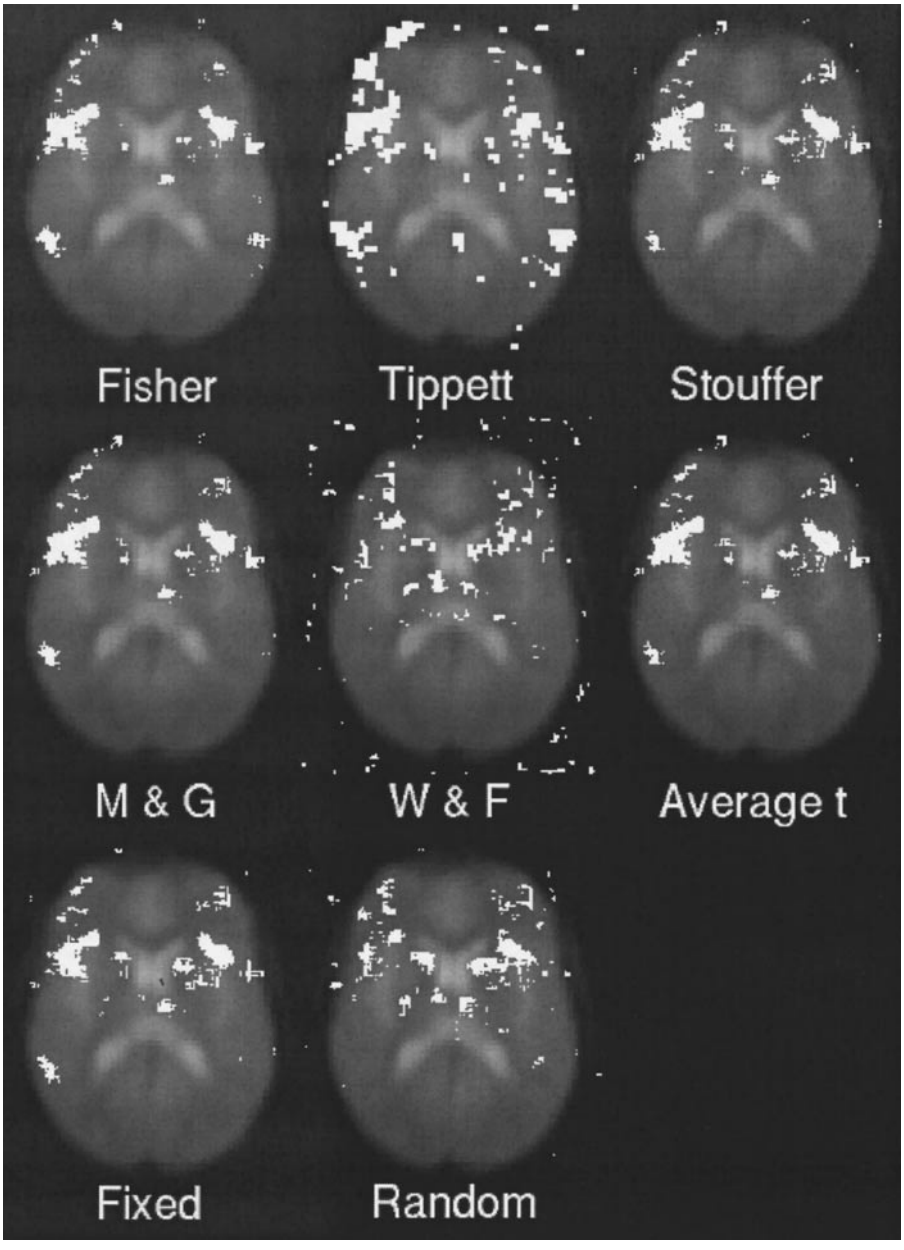


**FIG. 3.** Thresholded maps for all combining methods, for a threshold corresponding to the 1024 most active voxels. The top row shows Fisher, Tippett, Stouffer; the middle row shows Mudholkar and George, Worsley and Friston, average *t*; the bottom row shows fixed effects, random effects.

surprising at first glance. However, they are due to the fundamentally different nature of the two approaches. While the fixed effects model allows the investigator to draw conclusions regarding the specific subjects in the study (who are presumed to be of interest in and of themselves), the random effects model gives the researcher the ability to generalize findings to a larger population.

Figure 4 presents the results of one thresholding, for $\alpha = 0.0000028$, corresponding to a significance level of 0.05, with a Bonferroni correction for the 18000 brain voxels in a single slice. The cutoff value from the standard normal distribution appropriate to $\alpha = 0.0000028$ is approximately 4.5. As can be seen in the Fig. 4, there are some interesting differences among the methods, which are consistent for various levels of the threshold. One of the most striking effects is that Tippett's method looks much less spatially refined or detailed than any of the other procedures; this was also evident in Fig. 3. While the general pattern of activity is the same in this image as in the others, the overall impression is less satisfactory.

**TABLE 2**

The Number of Voxels Declared Significant for Each of the Eight Methods and Various Significance Levels

| Method $\alpha$ | Fisher | Tippett | Stouffer | Mudholkar | Worsley | Average $t$ | Fixed | Random |
|---|---|---|---|---|---|---|---|---|
| 0.000001 (0.018) | 699 | 301 | 335 | 514 | 3 | 388 | 338 | 31 |
| 0.000002 (0.036) | 788 | 362 | 386 | 585 | 7 | 446 | 381 | 39 |
| 0.000004 (0.072) | 886 | 443 | 459 | 668 | 12 | 524 | 453 | 52 |
| 0.000008 (0.144) | 1021 | 495 | 551 | 774 | 13 | 619 | 548 | 72 |
| 0.000016 (0.288) | 1197 | 634 | 650 | 895 | 13 | 724 | 646 | 99 |
| 0.000032 (0.567) | 1386 | 721 | 776 | 1033 | 23 | 867 | 759 | 128 |
| 0.000064 (> 1) | 1617 | 800 | 941 | 1223 | 42 | 1047 | 919 | 171 |

*Note.* Using Fisher's method to combine the data from the 11 subjects, and a significance level of 0.000004, 886 voxels will be declared active, whereas by using the average $t$ map, 524 voxels will be picked out as active. In parentheses are the corresponding familywise significance levels, based on 18,000 voxels. A $P$ value of 0.018, after Bonferroni adjustment for 18,000 voxels, gives a significance level of 0.000001 for each of the individual tests.

An easy way to begin to understand why Tippett's method exhibits this blockiness, is to consider the location where the minimum (over all the brains) $P$ value occurs. This location will be one of the "active" voxels. One of the subjects, subject $j$, has a $P$ value equal to this minimum. Within each subject, in Talairach coordinates, adjacent voxels are very highly correlated. Thus, within subject $j$ adjacent voxels will have values approximately equal to the minimum value and hence will be extreme as well.

For the same level $\alpha$, the qualitative impression given by all methods is the same; more or less the same regions are picked up as active, but the combining procedures do differ in how much activation is picked up and in the size of the activated areas. Fisher's method and the method of Mudholkar and George are comparable in the amounts of activation detected. Both of these find more active voxels than are found, using the same threshold, by Stouffer's method or either of the "meta-analysis" models, particularly random effects. The difference between random effects and the other methods in the number of active voxels for a given threshold might be due to the random effects model's reduced sensitivity to contributions from more unusual subjects. Worsley and Friston's conjunction test also shows very small amounts of activation, because it requires every subject to have a significant result. We believe this to be a very strong requirement.

## 4. CONCLUSIONS AND RECOMMENDATIONS

In this paper we have discussed eight methods for combining image data from multiple subjects and interpreting the results. Five of the methods are based on the combination of tests, more specifically of the $P$ values obtained from a statistical hypothesis test. The other three are based on the data themselves, in the form of means or $t$ tests.

A serious question that must be confronted in any effort to compare different ways of combining the information of multiple subjects is picking fair and objective thresholds. We have suggested some ideas here; a more in-depth discussion, with a particular focus on the False Discovery Rate of Benjamini and Hochberg (1995) can be found in Genovese, Lazar and Nichols (2001).

In addition, there is the practical issue of which method to use for the combining itself. It seems clear from the example described in the previous section that a simple averaging of the $t$ maps (or other statistical maps) summarizing each subject is going to be an efficient use of data only if compensation is made in the form of a reference distribution with smaller variance than the individual $t$ distributions. Otherwise the average will smooth away too much of the signal and the information available across subjects. For different reasons, Tippett's method is not to be recommended. The combined image is very blocky, and this, we believe, is a result of the transformation into Talairach coordinates. Tippett's method is one of only two methods we considered which does not apply some transformation to the combined values—aside from the conjunction (Worsley and Friston), the other $P$ value-based procedures take a logarithm (Fisher), an inverse cumulative distribution (Stouffer) or a logit transform (Mudholkar and George) and then do some sort of averaging. The methods based on combined estimation (fixed and random effects, average $t$) also are based on weighted averages. The effect of this weighting is to smooth over the realigned voxels. On the other hand, Tippett's method performs no such smoothing. The coarseness evident in the image is caused by the use of extreme $P$ values together with large correlation between neighboring voxels induced by the transformation into Talairach space. Worsley and Friston's approach apparently does not suffer from this problem, because it enforces a smoothing of its own, namely, that all voxels should show significant levels of activation.

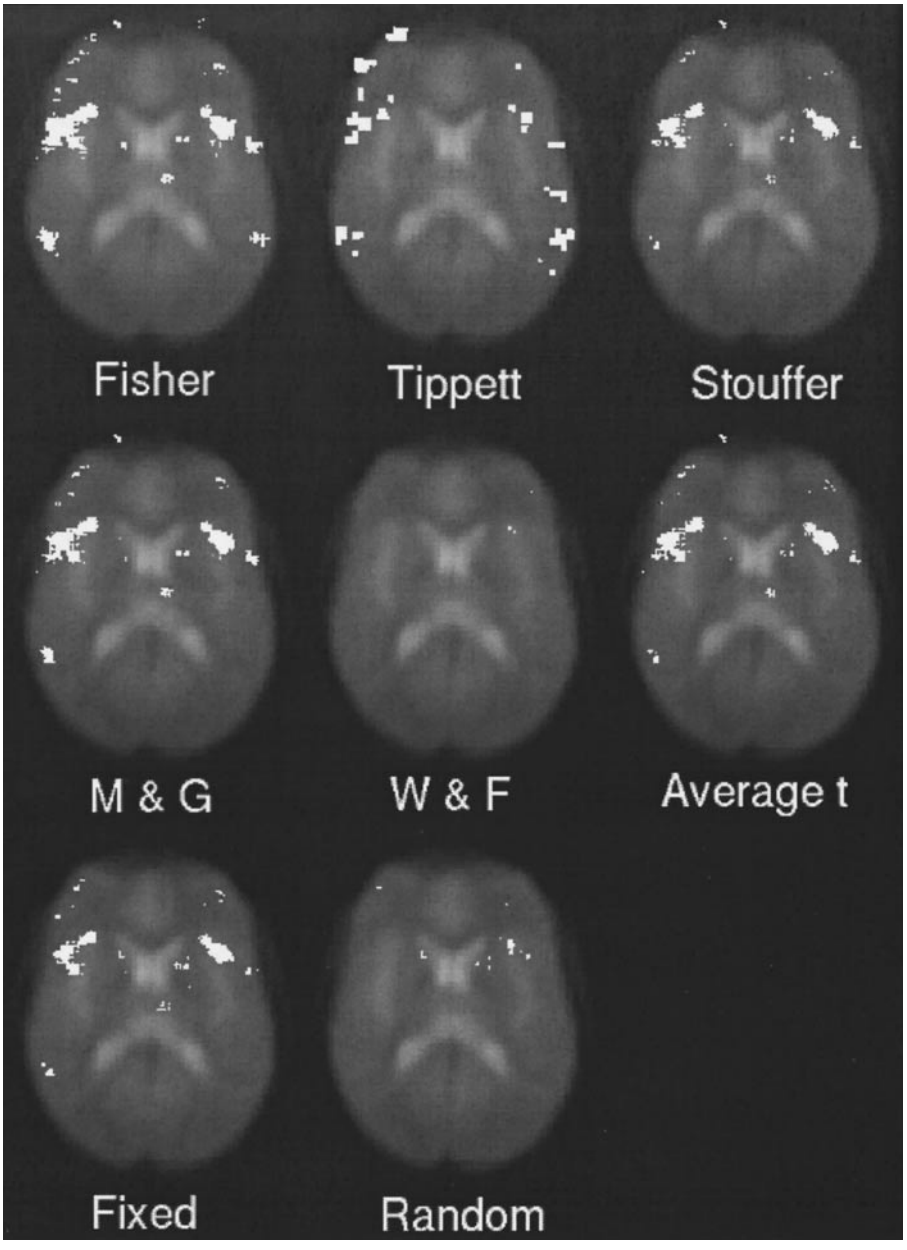


**FIG. 4.** Thresholded maps for all combining methods, for a threshold corresponding to a Bonferroni-corrected level 0.05 test, based on 18,000 brain voxels. The top row shows Fisher, Tippett, Stouffer; the middle row shows Mudholkar and George, Worsley and Friston, average *t*; the bottom row shows fixed effects, random effects.

Of the remaining methods, we agree with the recommendation of the National Research Council (1992) that, if possible, random effects or fixed effects models should be used. Both of these methods involve going back to the original data—in our example, we needed to go to the level of mean and standard deviation maps for each subject in each of the two conditions of the experiment; furthermore, the mathematical manipulations, especially for the random effects model, are more complicated. The potential difficulties in fitting the random effects model, in particular in obtaining estimates of the variance components, are offset by the generalizability of the results to the populations from which subjects are drawn. As this is often the goal of neuroscience studies, the random effects approach is clearly advantageous (Woods, 1996, makes a similar argument). It is worth keeping in mind for both of the model-based approaches that there is more statistical information in the raw data than in a transformation of the data. In addition, it is relatively straightforward to expand the two hierarchical models to include the case where we want to compare two or more groups of subjects. We plan to address this in a future study.

If the raw data are not available at the level necessary for fitting the fixed or random effects models, there are small differences among the remaining methods. Stouffer's method is very close to the fixed effects model and might be preferred for that reason. For similar reasons, Worsley and Friston's method could be recommended on the basis of approximating the random effects model. Fisher's method has the advantages that it is well known, is easy to implement, and has certain statistical optimality properties. On the other hand, there is some evidence that it may be less effective for large numbers of subjects (Rosenthal, 1978). It is worth pointing out that Fisher's method has been implemented in studies by local researchers (for example, Luna *et al.*, 2001) and it appears to give satisfactory results.

## ACKNOWLEDGMENTS

The authors thank Krista E. Garver and Christine Krisky for help in processing the neuroimaging data. We are also very grateful to two anonymous referees for their careful and detailed comments. Their patience and perspicacity contributed greatly to our understanding. All calculations were done in the FIASCO package, using mri_rpn_math.

## REFERENCES

Bahadur, R. R. 1967. Rates of convergence of estimates and test statistics. *Ann. Math. Stat.* **38**: 303–324.

Bahadur, R. R. 1971. *Some Limit Theorems in Statistics*, Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* Series *B* **57**: 289–300.

Benjamini, Y., and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. Technical Report, Department of Statistics and Operations Research, Tel Aviv University. *Ann. Stat.*, in press.

Berk, R. H., and Cohen, A. 1979. Asymptotically optimal methods of combining tests. *J. Am. Stat. Assoc.* **74**: 812–814.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10**: 101–129.

Eddy, W. F. 1997. Functional magnetic resonance imaging is a team sport. *Stat. Comput. Graphics News.* 8: 17–20.

Fisher, R. A. 1950. *Statistical Methods for Research Workers*, 11th ed. Oliver and Boyd, London.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. 1995. Improved assessment of significant change in functional magnetic resonance imaging (fMRI): Use of a cluster size threshold. *Magn. Reson. Med.* **33**: 636–647.

Genovese, C. R., Lazar, N. A., and Nichols, T. 2001. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *In press.*

Genovese, C. R., Noll, D. C., and Eddy, W. F. 1997. Estimating test–retest reliability in fMRI I: Statistical methodology. *Magn. Reson. Med.* **38**: 497–507.

Hedges, L. V. 1992. Meta-analysis. *J. Edu. Stat.* **17**: 279–296.

Holmes, A. P., Blair, R. C., Watson, J. D. G., and Ford, I. 1996. Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Mapp.* **16**: 7–22.

Holmes, A. P., and Friston, K. J. 1998. Generalisability, random effects and population inference. *NeuroImage* **7(4)**: S754.

Iyengar, S. 1991. Much ado about meta-analysis. *Chance* **4**: 33–40.

Jones, D. R. 1995. Meta-analysis: Weighing the evidence. *Stat. Med.* **14**: 137–149.

Lancaster, H. O. 1961. The combination of probabilities: An application of orthonormal functions. *Austrálian J. Stat.* **3**: 20–33.

Littell, R. C. and Folks, J. L. 1973. Asymptotic optimality of Fisher's method of combining independent tests II. *J. Am. Stat. Assoc.* **68**: 193–194.

Luna, B., Minshew, N. J., Garver, K., Lazar, N. A., Thulborn, K. R., Eddy, W. F. and Sweeney, J. A. 2001. Neocortical system abnormality in autism: An fMRI study of spatial working memory. Submitted.

Miller, R. G. 1981. *Simultaneous Statistical Inference*, 2nd ed. Springer-Verlag, New York.

Mudholkar, G. S., and George, E. O. 1979. The logit method for combining probabilities. In *Symposium on Optimizing Methods in Statistics* (J. Rustagi, Ed.), pp. 345–366. Academic Press, New York.

National Research Council 1992. *Combining Information: Statistical Issues and Opportunities for Research*, Contemporary Statistics, Number 1. National Academy Press, Washington.

Pearson, K. 1933. On a method of determining whether a sample of a given size *n* supposed to have been drawn from a parent population having known probability integral has probably been drawn at random. *Biometrika* 25: 379–410.

Rao, C. R. 1971. Estimation of variance and covariance components—MINQUE theory. *J. Multivariate Anal.* **1**: 257–275.

Rao, C. R., and Kleffe, J. 1988. *Estimation of Variance Components and Applications*, 2nd ed. Wiley, New York.

Rosenthal, R. 1978. Combining results of independent studies. *Psychol. Bull.* 85: 185–193.

Rosenthal, R. 1979. The "file drawer problem" and tolerance for null results. *Psychol. Bull.* **86**: 638–641.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams, R. M. 1949. *The American Soldier: Vol. I. Adjustment During Army Life.* Princeton Univ. sity Press, Princeton.

Sweeney, J. A., Mintun, M. A., Kwee, S., Wiseman, M. B., Brown, D. L., Rosenberg, D. R. and Carl, J. R. 1996. A positron emission tomography study of voluntary saccadic eye movements and spatial working memory. *J. Neurophysiol.* 75: 454–468.

Talairach, J. and Tournoux, P. 1988. *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging.* Georg Thieme Verlag, Stuttgart.

Tippett, L. H. C. 1931. *The Method of Statistics*, 1st ed. Williams and Nor-gate, London.

Wilkinson, B. 1951. A statistical consideration in psychological research. *Psychol. Bull.* **48**: 156–158.

Woods, R. P. 1996. Modeling for intergroup comparisons of imaging data. *NeuroImage* **4(3)**: S84–S94.

Worsley, K. J. 1994. Local maxima and the expected Euler characteristic of excursion sets of $\chi^2$, $F$ and $t$ fields. *Adv. Appl. Probability* **26**: 13–42.

Worsley, K. J., and Friston, K. J. 2000. A test for a conjuction. *Stat. Probability Lett.* **47**: 135–140.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. 1996. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4**: 58–73.