NeuroImage **13,** 920–930 (2001)
doi:10.1006/nimg.2001.0764, available online at http://www.idealibrary.com on IDEAL®

Estimation of the Number of "True" Null Hypotheses in Multivariate Analysis of Neuroimaging Data

F. E. Turkheimer,*,† C. B. Smith,* and K. Schmidt*,[1]

*Laboratory of Cerebral Metabolism, National Institute of Mental Health, Bethesda, Maryland 20892; and †MRC Cyclotron Unit, Hammersmith Hospital, DuCane Road, London W12-0NN, United Kingdom

Received August 8, 2000

The repeated testing of a null univariate hypothesis in each of many sites (either regions of interest or voxels) is a common approach to the statistical analysis of brain functional images. Procedures, such as the Bonferroni, are available to maintain the Type I error of the set of tests at a specified level. An initial assumption of these methods is a "global null hypothesis," i.e., the statistics computed on each site are assumed to be generated by null distributions. This framework may be too conservative when a significant proportion of the sites is affected by the experimental manipulation. This report presents the development of a rigorous statistical procedure for use with a previously reported graphical method, the *P* plot, for estimation of the number of "true" null hypotheses in the set. This estimate can then be used to sharpen existing multiple comparison procedures. Performance of the *P* plot method in the multiple comparison problem is investigated in simulation studies and in the analysis of autoradiographic data. © 2001 Academic Press

Key Words: **PET; autoradiography; multiple comparisons; *P* plot.**

INTRODUCTION

Data derived from imaging modalities, such as autoradiography or positron emission tomography (PET), are usually presented as measures of blood flow or metabolic activity in a number of sites of interest in the brain, where each site is either an anatomical region of interest (ROI) or a voxel in the data-volume (Ford *et al.*, 1991). Statistical analysis of images acquired in various experimental conditions is often performed by collecting results of univariate tests (Student's *t* tests or their equivalent) computed on each site of the images and by retaining as significant those ROIs or voxels in which the magnitude of the test statistic is above a certain threshold (Ford *et al.*, 1991; Friston *et al.*, 1991; Worsley *et al.*, 1992). The choice of threshold is usually influenced by the goal of minimizing the number of false positives that may be generated by the whole set of tests; i.e., the larger the number of tests, the higher is the expected number of false positives, and so the threshold is raised accordingly. This framework is often designated as the multiple comparison problem (for a review see Hochberg and Tamhane, 1987). Unfortunately, such a procedure, although simple in principle, decreases the probability of detecting "true" differences among experimental conditions as the number of tests increases (Hochberg and Benjamini, 1990). The present report describes an alternative approach to the multiple comparison problem that introduces some additional modeling of the set of collected statistics.

The report is organized as follows: Section 1.1 introduces notation, defines the problem of multiple comparisons, and gives an overview of the available approaches to its solution; Section 1.2 briefly specifies the problem for neuroimages; Section 1.3 gives some insight on the limitation of these procedures; Section 1.4 introduces a graphical method for the estimation of the number of "true" null hypotheses from the set of statistics; and Section 1.5 details its application in the multiple comparison problem. Section 1.6 contains technical details for the computation of the plot. Finally, the theoretical framework is applied to simulation studies and to the analysis of ROI data obtained from autoradiographic studies in rodents.

THEORETICAL FRAMEWORK

1.1. Problem and Notation

Consider a univariate statistic t and let $f(t, \theta)$ be its probability density function (p.d.f.) with parameter vector θ. In the case of the normal distribution $N(\mu, \sigma^2)$, θ is the two element vector $\theta = [\mu, \sigma^2]$, where μ and σ^2 are, respectively, the mean and variance of the density.

In the usual hypothesis-testing framework, it is of interest to estimate the likelihood of the data's having been generated under a null hypothesis; this corre-

[1] Software available from the corresponding author upon request.

1053-8119/01 $35.00

sponds to determining the probability p that the test statistic belongs to the null distribution. For a two-tailed test,

$$p = 1 - \int_{-t}^{t} f(x, \theta_0)\, dx, \tag{1}$$

where t is the observed value of the test statistic, and θ_0 is the vector of parameters of the null p.d.f. Examples of null distributions are the standard normal distribution $N(0, 1)$ and Student's t distribution, $T^{df}(0)$, where df indicates the degrees of freedom. In a univariate test, the observed value t allows rejection of the null hypothesis if its corresponding P value is smaller than a threshold value α; this threshold is usually arbitrarily set to some conventional value, e.g., 0.05, 0.01, or 0.001. Small values of P provide strong evidence that the null hypothesis is not true. α corresponds to the Type I error or to the chosen probability risk of false rejection of the null hypothesis.

Consider now the problem of simultaneously testing N univariate null hypotheses on N sites of interest, e.g., on N ROIs or voxels. Let $t_1, t_2, \ldots, t_N$ be the observed values of the test statistic and $p_1, p_2, \ldots, p_N$ the corresponding probabilities. In order to control the multiplicity effect when considering a family of comparisons simultaneously, classical multiple comparison procedures seek to control the probability of committing *any* Type I error. The control of this Family-Wise Error (FWE) implies that the probability of false rejection of *at least one* null hypothesis should be less than the error level α. This framework is common in neuroimaging data analysis (Ford *et al.*, 1991). A simple procedure that controls the FWE at level α is the Bonferroni procedure that allows the rejection of the i^{th} null hypothesis in a set of N tests if

$$p_i \leq \alpha/N. \tag{2}$$

The Bonferroni procedure can be improved by realizing that once one of the N null hypotheses has been rejected it cannot be considered *true* anymore and the number of possible true hypotheses remaining will be now $(N - 1)$. Holm (1979) incorporated this concept into a more powerful step-down procedure that maintains control of the FWE.

Suppose, without loss of generality, that the P values previously defined are ordered so that $p_1 < p_2 < \ldots < p_N$. Then, according to Holm's procedure, the i^{th} null hypothesis, H_i, is rejected when, for all $j = 1, \ldots, i$

$$p_j < \alpha/(N - j + 1). \tag{3}$$

Thus one starts from H_1 and tests whether the inequality $p_1 < \alpha/N$ holds; if so then H_1 is rejected and H_2 is tested by examining whether the inequality $p_2 < \alpha/(N - 1)$ holds, and so on. The iterative procedure ends whenever Inequality (3) is not satisfied, and all remaining hypotheses are accepted.

More recently Hochberg (1988) introduced a step-up version of Holm's procedure that follows from the closure principle of Marcus *et al.* (1976). The procedure extends from highest to lower P value by accepting H_i, if, for all $j = i, \ldots, N$,

$$p_j > \alpha/(N - j + 1). \tag{4}$$

Whenever Inequality (4) is not satisfied the hypothesis H_i and all hypotheses with lower P values are rejected.

The Hochberg procedure controls the FWE and is uniformly more powerful than the Holm procedure since it rejects any hypothesis rejected by the latter and applies the same correction quotient in a step-up fashion (Dunnet and Tamhane, 1992).

These Bonferroni-type procedures require no assumptions about the data and the measured variables. Further developments have sharpened the step-down and step-up approaches by incorporating the dependence structure among variables either in the case of normally distributed data with known covariance (Dunnet and Tamhane, 1991, 1992, 1995) or in the case of unknown covariance through randomization testing (Westfall and Young, 1993, pp. 116–117; Troendle, 1996).

1.2. Analysis of Neuroimages and the Issue of Spatial Correlation

In the analysis of neuroimages we focus on two approaches: ROI analysis and voxel-by-voxel analysis. Both methods are intended to detect localized changes in a physiological parameter of interest under controlled FWE. Both analyses consist of a filtering step and a thresholding procedure.

In ROI analysis there are expectations on the location and extent of changes in the physiological parameter as the regions usually correspond to anatomically defined areas. Therefore, the images are sampled with a carefully placed template of ROIs. This is the filtering operation; it is equivalent to the application of a mean filter of a defined shape, i.e., the shape of the ROI, placed in a specific location. Data are then used to derive appropriate statistics and perform the tests, e.g., t tests in the case of a group comparison. ROIs are clearly spatially independent in autoradiographic data, and in PET spatial correlation among ROIs should be negligible unless the ROIs are very small and adjacent. Therefore, Bonferroni or Bonferroni-like procedures can be employed to control the FWE efficiently.

In voxel-by-voxel analysis there is little information on location and extent of the expected change in the physiological parameter of interest. One may select a smoothing filter of a particular resolution, e.g., 12 mm, apply it to the entire volume, and compute statistics for all sites of the volume. Since adjacent sites are spatially correlated, efficient control of the FWE may be achieved by using either parametric procedures derived from Gaussian random field theory (Friston *et al.*, 1991; Worsley *et al.*, 1992) or nonparametric procedures when the underlying statistical distribution is unknown (Holmes *et al.*, 1996). The procedure can be repeated for a number of different resolutions and the error rate can be controlled accordingly in a parametric (Worsley *et al.*, 1996) or nonparametric fashion (Poline and Mazoyer, 1994).

More recently a method that relies on the wavelet transform, an orthogonal transform for nonstationary signals, has been introduced for the multiscale analysis of neuroimages (Turkheimer *et al.*, 2000). In analogy with the ROI and voxel-by-voxel analyses, the image is passed through a battery of mirror filters that generate a new representation of the data wherein each coefficient represents an object of a specific resolution and position in the original image. In this approach statistical analysis is performed by thresholding in wavelet space. It can be shown that in this particular functional space, which has the same size as the original volume, one can easily achieve complete decorrelation among wavelet coefficients both between and within resolutions (Turkheimer *et al.*, 2000). Therefore, for this method a simple Bonferroni procedure suffices for the efficient control of FWE.

Given these considerations, this article concentrates on the problem of efficient control of FWE in the case of data that are not spatially correlated. Application to neuroimages is then natural both in the case of ROI analysis, and, though transformation to wavelet space, to voxel-by-voxel analysis.

1.3. Limitations of Classical Multiple Comparisons Procedures

The use of procedures for the control of the FWE preserves the significance level α for the set of comparisons but decreases substantially the power of detecting a true effect, i.e., it preserves the significance level at the expense of increasing the Type II error (Hochberg and Benjamini, 1990; Benjamini and Hochberg, 1995). The concern for this issue has raised a number of approaches that divert attention from the control of the FWE and, instead, deal with different philosophies that range from the control of some function of Type I and Type II errors (Spiøtvoll, 1972; Soric, 1989; Benjamini and Hochberg, 1995) to the total avoidance of correction for multiple comparisons (Rothman, 1990; Saville, 1990).

In the following section an alternative approach to the multiple comparison problem will be introduced that is motivated by the following observations. First, it is important to note that all the procedures described above are designed to control the number of false rejections in a set of N null hypotheses *supposing that all are true.* In other words, these methods control the number of statistics over a certain threshold generated by N independent sites (Bonferroni-type procedures) with the assumption that *they are all generated by null distributions.* This is a result of the extension to the multivariate problem of the univariate hypothesis testing framework that has the premise of a separate experiment for each inference assumed to be stated prior to the experiment (Hochberg and Benjamini, 1990).

The assumption of all the univariate hypotheses being null can result in a test's being too conservative when, for example, there is measurable evidence of a treatment effect from a global test such as the Hotelling T^2 or others. This argument had been incorporated into the approach of Duncan (1965) where the significance level of the pair-wise testing among means depends in a inverse way on the analysis of variance F statistic. The same issue arises when most of the significance levels computed in a multiple testing experimental design cluster around low P values, such as 0.1 or 0.05, which suggests that the assumption that all the H_i are true is quite unrealistic.

Let now N_0 be the number of *true* null hypotheses or, equivalently, the number of statistics generated by null distributions in the set. Consequently $(N - N_0)$ is the number of sites in which statistics are generated by non-null distributions. Of course N_0 is usually unknown, but it can be observed that if an estimate of it were available, less conservative corrections could be used (Schweder and Spjøtvoll, 1982; Hochberg and Benjamini, 1990).

1.4. The P Plot Graphical Method

An interesting approach to the multiple comparison problem considers the estimation of the number of *true* null hypotheses N_0 from the N tested and uses this estimate to sharpen the Bonferroni-type procedures previously described (Hochberg and Benjamini, 1990; Schweder and SpjØtvoll, 1982).

Suppose that N hypotheses are tested and that the corresponding P values are calculated. It is known that the P value is uniformly distributed over [0, 1] when the hypothesis under test is true. The P plot graphical method consists of plotting the ordered values $q_i = 1 - p_i$, sorted in ascending order, versus their rank. On this plot the points corresponding to the true hypotheses (large values of p_i) should roughly fall along a straight line passing through the origin and the points corresponding to the false hypotheses (small values of p) should deviate rightward. The slope of the straight line

fitted to the points with large values of p should give an estimate of N_0, say $\hat{N}_0$, that is computed as

$$\hat{N}_0 = (1/\hat{\beta}) - 1 \tag{5}$$

where $\hat{\beta}$ is the estimate of the slope. This procedure is valid and the estimate unaltered also if the p are correlated (Schweder and SpjØtvoll, 1982). Some examples of the plot are illustrated in Fig. 1.

1.5. The P Plot and Multiple Comparison Procedures

A number of different applications of the *P* plot have been suggested. The technique can be applied for informal inference to give an overall view of the collection of univariate tests as a valuable indicator of the number of variables affected by the treatment (Schweder and SpjØtvoll, 1982).

However, the estimate of N_0, by itself, does not allow direct inferences on each hypothesis; in fact, knowing the number of true and false hypotheses in a collection of tests does not ascertain which hypotheses are true and which are false. Thus, the practice of rejecting the null hypotheses with the $(N - \hat{N}_0)$ smallest P values can cause an uncontrolled number of false rejections. In order to make formal probabilistic statements on each hypothesis, the estimate $\hat{N}_0$ can be used with one of the Bonferroni-type procedures with a consequent increase in its power. For example, in the case of the Bonferroni method, if the FWE has to be controlled at the overall level α then the level $\alpha/\hat{N}_0$ will be used for the individual tests instead of α/N with a significant decrease of Type II errors, particularly when $\hat{N}_0 \ll N$ (Schweder and SpjØtvoll, 1982).

Similar modifications apply to the stepwise methods (Hochberg and Benjamini, 1990). The Hochberg procedure, for example, can be modified as follows. Given the set of ordered P values and corresponding hypotheses previously defined, start the procedure by accepting those hypotheses for which $p_i > \alpha$, and let N_1 be their number. Consider then the inequality:

$$p_{(N-N_1)} < \frac{\alpha}{\min(N_0,\ N_1 + 1)} \tag{6}$$

If (6) is satisfied then $H_{(N-N_1)}$, and all the ordered hypotheses with smaller P values, are rejected. If the inequality is not satisfied then $H_{(N-N_1)}$ is accepted, N_1 is increased by 1 and the inequality (6) is reapplied to the following hypotheses (Hochberg and Benjamini, 1990).

1.6. A Procedure for the Application of the P Plot

Two problems hamper the use of the graphical technique. The first concerns the number of $q_i = 1 - p_i$ used in the estimation of the slope $\hat{\beta}$, the choice of which affects both the bias and the uncertainty of the estimate. Often the plot will not show a clear-cut break between the aligned q_i belonging to the "true" null hypotheses and those deviating from the straight line that presumably belong to "false" null hypotheses, but a rather gradual bend will be observable. The trade-off then is between considering a large number of q_i, which lowers the variance of the estimate, or computing the slope from only the initial points close to the

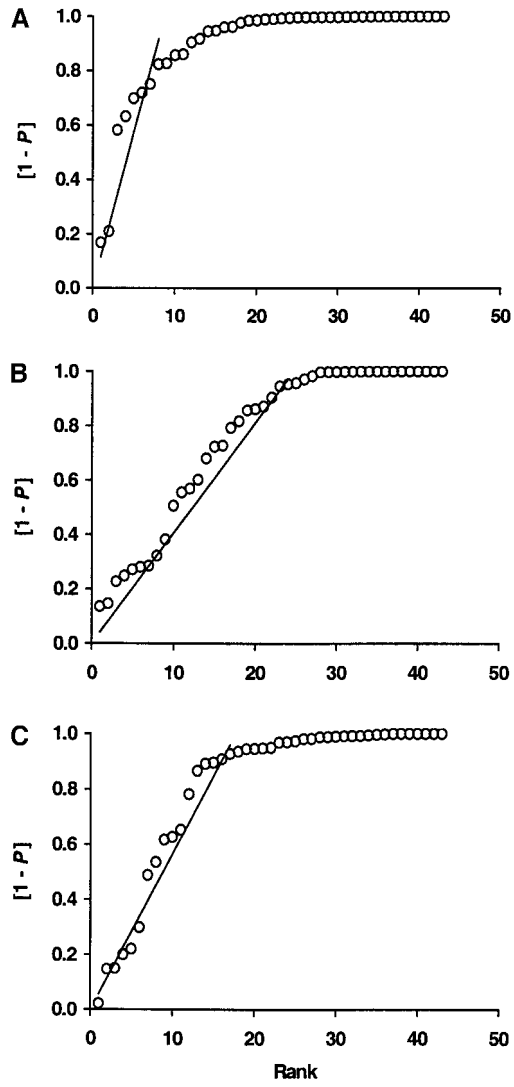

FIG. 1. *P* Plots of the contrasts among the groups of rats in which local rates of cerebral glucose utilization were measured with the quantitative autoradiographic [^{14}C]deoxyglucose method. Data are from the study of Entrei *et al.* (1999) and are shown in Table 1. In each contrast, a *P* value, p_i, was computed by use of a Student's t test on each of the 43 regions of interest, and the complementary probabilities, $q_i = 1 - p_i$, were rank ordered so that $q_1 \leq q_2 \leq \cdots \leq q_{43}$. The q_i values were then plotted on the ordinate and their ranks on the abscissa. The estimated slope of the linear portion of the curve, $\hat{\beta}$, was used to derive the estimated number of "true" null hypotheses, $\hat{N}_0$, as $\hat{N}_0 = (1/\hat{\beta}) - 1$. (A) For the contrast of the control group with the group of rats administered diazepam, $\hat{N}_0 = 8$; (B) Control versus ketamine, $\hat{N}_0 = 24$; (C) Control versus ketamine plus diazepam, $\hat{N}_0 = 17$.

origin, which reduces the bias that would result from the inclusion of P values, corresponding to "false" null hypotheses (Schweder and SpjØtvoll, 1982). Note that the issue of variance reduction is relevant in this context, i.e., an unbiased estimator with high variance would produce, with large probability, estimates $\hat{N}_0 \ll N_0$ and therefore alter the ability to control the FWE at the desired level. In fact, the consideration of variance reduction is the second problem and drives the need for a method for the calculation of the regression line that takes into account the statistical attributes of the ordered statistics.

Both problems are addressed here in a new approach that departs from previous practices that were based on "ad hoc" least squares methods (Hochberg and Benjamini, 1990; Brown and Russel, 1997).

The problem of selection of K identically distributed scores from the N observed $p_1, p_2, ..., p_N$, is a traditional change-point problem (for a review see, for example, Csörgo and Horváth, 1997). The typical solution consists of applying a test of uniformity to iteratively reduced sets of ordered P scores. If the test is rejected, the smallest P value is discarded and the test reapplied. If the test is accepted, the surviving set is selected for estimation of the slope of the P plot. For this type of application, tests for uniformity are usually based on the residuals $v_i = p_i - i/(K + 1)$, where p_i are the ordered P values and $i/(K - i + 1)$ their expected values. A suitable statistic is then $C^+ = \max(v_i)$ (see D'Agostino and Stephens, 1986, pp. 336–337, for background and tabulated percentiles). The scores p_1, $p_2, \ldots, p_K$ selected in the above step can then be used for estimation of the slope of the plot, from which the estimated number of true null hypotheses is then determined by Eq. (5). Each p_i is a beta random variable with parameters $(i, K - i + 1)$, mean $i/(K + 1)$ and variance $[i(K - i + 1)]/[(K + 1)^2(K + 2)]$ (Quesenberry and Hales, 1980). For the slope estimation we can adopt the normal approximation to the beta distribution and use a weighted least squares (WLS) procedure with variances $[i(K - i + 1)]/[(K + 1)^2(K + 2)]$ to weight the residuals.

MATERIALS AND METHODS

The following sections describe two experiments with simulated and measured data sets. In the first experiment we applied the P plot graphical method to simulated data sets in order to validate the numerical procedure and to investigate its properties. The second experiment illustrates a real-case application by considering the analysis of a set of ROI autoradiographic data.

In all instances, the P plot was computed according to the method described in Section 1.6. The significance value for the C^+ test for uniformity was set at $\alpha = 0.01$. All algorithms were implemented in MATLAB (The Mathworks Inc., Natick, MA) and run on a Ultra-Sparc 10 Sun Workstation (SunSystems, Mountain View, CA).

1. Monte-Carlo Simulations

Simulated data sets were designed to resemble experimental conditions usually met in neuroimaging studies, particularly autoradiography and PET, in terms of sample sizes, treatment effects, and number of variables (regions of interest) tested.

The first set of simulated data modeled a control and a treatment group, each consisting of 10 experimental subjects, with 50 ROIs. Data were normally distributed with standard deviation equal to unity; a treatment effect d equal to one standard deviation was added to the mean of the treatment group in 47, 45, 40, 35, 30, 20, 10, or 0 ROIs so that the number of true null hypotheses, N_0, was 3, 5, 10, 15, 20, 30, 40, or 50, respectively. In each ROI the difference in means between the control and treatment group was tested with the two-tailed Student's t test. The P plot method was applied to each set of 50 P values resulting from the Student's t tests. The process was repeated 5000 times, and the mean and standard deviation of the estimates $\hat{N}_0$ were calculated. For each simulated data-set, P values were then fed into a Hochberg step-up procedure that used either N or $\hat{N}_0$ as number of true null hypotheses; FWE was controlled at $\alpha = 0.05$. This enabled the evaluation of the performance of the combined use of the plot and a multiple comparisons procedure both in terms of control of FWE and power.

Three other simulations were performed, each one preserving the parameters of the first, except changing in turn the sample size (from 10 to 20 subjects), the treatment effect (d from 1 to 1.5 units), or the total number of regions of interest (from 50 to 20 ROIs).

2. Autoradiographic Data Set

The P plot was applied to data from a study of the effects of ketamine and diazepam anaesthesia on local cerebral glucose utilization ($lCMR_{glc}$) in rats (Entrei *et al.*, 1999). $lCMR_{glc}$ was measured with the quantitative [^{14}C]deoxyglucose method (Sokoloff *et al.*, 1977) in 43 regions of interest in four groups of rats: a control group ($n = 6$); a group treated with ketamine (10 mg/kg) ($n = 5$); a group treated with diazepam (0.5 mg/kg) ($n = 6$); and a group treated with both ketamine (5 mg/kg) and diazepam (0.25 mg/kg) ($n = 6$). The data are shown in Table 1. For the purposes of the current analysis, it was assumed that the contrasts of the control group against each of the three treatment groups are of interest. Homogeneity of variances among the four groups was determined by use of Bartlett's test (Snedecor and Cochran, 1980, pp. 252–253), t statistics were computed, and P values were determined from a t distribution with the appropriate de-

TABLE 1

Local Rates of Cerebral Glucose Utilization ($lCMR_{glc}$) in Control Rats and Rats Treated with Diazepam, Ketamine, or Ketamine Plus Diazepam[a]

	$lCMR_{glc}$ (μmol/100 g/min)			
	Control (n = 6)	Diazepam (0.5 mg/kg) (n = 6)	Ketamine (10 mg/kg) (n = 5)	Ketamine (5 mg/kg) plus Diazepam (0.25 mg/kg) (n = 6)
Auditory system				
Auditory cortex	139 ± 12	105 ± 18	101 ± 11	96 ± 14
Medial geniculate	115 ± 11	92 ± 16	80 ± 6	75 ± 12
Inferior colliculus	173 ± 18	152 ± 25	112 ± 18	117 ± 7
Superior olivary nucleus	119 ± 9	134 ± 30	123 ± 18	151 ± 23
Ventral cochlear nucleus	136 ± 18	154 ± 31	131 ± 14	138 ± 13
Lateral lemniscus	67 ± 8	59 ± 15	65 ± 5	53 ± 10
Visual system				
Visual cortex	93 ± 10	85 ± 17	110 ± 14	77 ± 15
Lateral geniculate	82 ± 8	62 ± 9	76 ± 3	68 ± 12
Superior colliculus	75 ± 9	70 ± 6	81 ± 14	63 ± 12
Sensorimotor system				
Sensorimotor cortex	101 ± 7	72 ± 7	90 ± 11	72 ± 11
Ventrolateral thalamic nucleus	95 ± 12	70 ± 13	89 ± 5	76 ± 15
Red nucleus	74 ± 3	58 ± 8	79 ± 8	61 ± 11
Medial vestibular nucleus	110 ± 15	96 ± 18	102 ± 8	98 ± 17
Hypoglossal nucleus	63 ± 6	52 ± 11	56 ± 3	55 ± 12
Cerebellar grey matter	58 ± 3	45 ± 10	53 ± 5	47 ± 4
Olfactory system				
Olfactory tubercle	110 ± 12	81 ± 16	151 ± 30	100 ± 13
Limbic system				
Fornix	58 ± 14	52 ± 7	81 ± 7	48 ± 8
Presubiculum	95 ± 7	61 ± 14	172 ± 11	97 ± 21
Cingulum	42 ± 5	34 ± 3	54 ± 5	38 ± 6
Retrosplenial agranular cortex	106 ± 14	69 ± 7	140 ± 12	84 ± 13
Cingulate cortex	119 ± 11	83 ± 11	156 ± 29	98 ± 16
Anteroventral thalamic nucleus	126 ± 23	73 ± 10	183 ± 23	100 ± 17
Mamillary body	109 ± 6	44 ± 11	173 ± 16	54 ± 9
Hippocampus: CA3	87 ± 8	64 ± 9	137 ± 18	93 ± 19
Entorhinal cortex	61 ± 7	45 ± 7	78 ± 9	53 ± 10
Amygdala	57 ± 8	58 ± 7	75 ± 9	61 ± 8
Nucleus accumbens	72 ± 10	58 ± 16	74 ± 20	76 ± 22
Interpeduncular nucleus	96 ± 12	87 ± 15	117 ± 16	96 ± 10
Medial Habenula	66 ± 8	58 ± 10	63 ± 2	57 ± 5
Lateral Habenula	95 ± 10	66 ± 7	76 ± 6	65 ± 9
Basal ganglia				
Caudate putamen	111 ± 10	84 ± 11	125 ± 15	108 ± 24
Substantia nigra, pars compacta	85 ± 10	65 ± 15	102 ± 15	86 ± 14
Substantia nigra, pars reticulata	45 ± 12	34 ± 7	54 ± 9	40 ± 7
Globus pallidus	52 ± 5	51 ± 9	65 ± 9	49 ± 11
Association areas				
Prefrontal cortex	90 ± 9	69 ± 7	94 ± 7	78 ± 14
Frontal cortex	94 ± 7	69 ± 9	95 ± 9	76 ± 10
Hypothalamus				
Paraventricular nucleus	52 ± 8	45 ± 7	50 ± 7	42 ± 4
Supraoptic nucleus	64 ± 6	40 ± 8	57 ± 4	41 ± 9
Myelinated fiber tracts				
Corpus callosum	37 ± 2	30 ± 4	49 ± 3	30 ± 5
Anterior commissure	43 ± 4	35 ± 7	50 ± 10	36 ± 3
Hippocampal fimbria	27 ± 3	20 ± 2	26 ± 2	21 ± 4
Cerebellar white matter	36 ± 4	29 ± 5	37 ± 4	30 ± 4
Whole brain weighted average	69 ± 8	52 ± 6	73 ± 6	57 ± 7

[a] Values are Means ± SD for the number of animals indicated in parentheses. From Eintrei *et al.* (1999).

TABLE 2

Number of True Null Hypotheses Estimated by Graphical *P*-Plot Method in Simulated Data Sets[a]

Graphical *p*-plot method: Simulation 1

Number of regions (*N*)	50 Regions of interest
Subjects per group	10 Subjects
Treatment effect (*d*)	1.0 Unit

Number of true null hypotheses	Estimated number of true null hypotheses ($\hat{N}_0$)		Hochberg step-up procedure ($N_0 = N$)		Hochberg step-up procedure ($N_0 = \hat{N}_0$)	
N_0	Mean	S.D.	FWE ($\alpha = 0.05$)	Power	FWE ($\alpha = 0.05$)	Power
3	13.2	3.79	0.005	0.089	0.017	0.189
5	15.3	3.80	0.007	0.090	0.021	0.175
10	20.0	3.79	0.009	0.089	0.029	0.148
15	24.6	3.83	0.019	0.088	0.035	0.130
20	29.0	3.69	0.019	0.088	0.033	0.119
30	37.3	3.46	0.030	0.086	0.040	0.102
40	44.7	2.83	0.040	0.086	0.046	0.092
50	50.5	2.20	0.049	N/A	0.048	N/A

Graphical *p*-plot method: Simulation 2

Number of regions (*N*)	50 Regions of interest
Subjects per group	10 Subjects
Treatment effect (*d*)	1.5 Unit

Number of true null hypothesis	Estimated number of true null hypotheses ($\hat{N}_0$)		Hochberg step-up procedure ($N_0 = N$)		Hochberg step-up procedure ($N_0 = N_0$)	
N_0	Mean	S.D.	FWE ($\alpha = 0.05$)	Power	FWE ($\alpha = 0.05$)	Power
3	5.50	1.76	0.005	0.427	0.038	0.668
5	7.70	1.98	0.008	0.423	0.038	0.617
10	13.1	2.29	0.012	0.411	0.039	0.540
15	18.2	2.45	0.022	0.401	0.047	0.492
20	23.2	2.53	0.028	0.398	0.047	0.463
30	33.1	2.67	0.033	0.380	0.044	0.412
40	42.6	2.68	0.039	0.371	0.042	0.382
50	50.4	2.17	0.053	N/A	0.054	N/A

Graphical *p*-plot method: Simulation 3

Number of regions (*N*)	50 Regions of interest
Subjects per group	20 Subjects
Treatment effect (*d*)	1.0 Unit

Number of true null hypotheses	Estimated number of true null hypotheses ($\hat{N}_0$)		Hochberg step-up procedure ($N_0 = N$)		Hochberg step-up procedure ($N_0 = \hat{N}_0$)	
N_0	Mean	S.D.	FWE ($\alpha = 0.05$)	Power	FWE ($\alpha = 0.05$)	Power
3	5.06	1.53	0.004	0.398	0.039	0.684
5	7.22	1.85	0.007	0.395	0.043	0.629
10	12.6	2.18	0.015	0.386	0.043	0.539
15	17.8	2.37	0.021	0.377	0.046	0.482
20	22.9	2.40	0.028	0.369	0.046	0.445
30	32.8	2.61	0.031	0.356	0.040	0.394
40	42.5	2.62	0.038	0.341	0.042	0.341
50	50.3	2.18	0.051	N/A	0.051	N/A

TABLE 2—*Continued*

Graphical *p*-plot method: Simulation 4

Number of regions (N)	20 Regions of interest
Subjects per group	10 Subjects
Treatment effect (d)	1.0 Unit

Number of true null hypotheses	Estimated number of true null hypotheses ($\hat{N}_0$)		Hochberg step-up procedure ($N_0 = N$)		Hochberg step-up procedure ($N_0 = \hat{N}_0$)	
N_0	Mean	S.D.	FWE ($\alpha = 0.05$)	Power	FWE ($\alpha = 0.05$)	Power
3	7.43	2.3	0.010	0.157	0.029	0.255
5	9.27	2.4	0.012	0.152	0.031	0.224
8	11.9	2.3	0.023	0.155	0.039	0.199
10	13.6	2.3	0.027	0.154	0.043	0.183
12	15.2	2.1	0.037	0.150	0.045	0.171
15	17.3	1.9	0.043	0.147	0.051	0.159
17	18.6	1.6	0.037	0.143	0.043	0.150
20	20.5	1.9	0.050	N/A	0.052	N/A

[a] Values are Means ± SD of 5000 simulated data sets. N/A, Not applicable (there were no false null hypotheses in these simulations).

grees of freedom. FWE was controlled at $\alpha = 0.05$ for each of the 3 sets of 43 comparisons.

RESULTS

1. Simulation Studies

The means and standard deviations of the number of true null hypotheses estimated by the *P* plot method in the simulation studies are displayed in Table 2. Included also are the experimental FWE (frequency of at least one true null hypothesis rejected) and power (fraction of non-null distributions rejected) of the Hochberg step-up procedure that used either the total number of hypotheses tested ($N_0 = N$) or the number of true null hypotheses estimated by the graphical *P* plot method ($N_0 = \hat{N}_0$). Consistent with previous work (Schweder and SpjØtvoll, 1982) the estimator was found to be biased. Although bias is usually an undesirable feature of an estimation procedure, in this case it is not problematic because the shift is in the conservative direction ($\hat{N}_0 > N_0$). Note that bias is reduced when the statistical analysis gains power with a greater number of subjects ($n = 20$) or with an increased treatment effect ($d = 1.5$ units). Remarkably, the procedure is effective also with a smaller number of *P* values ($N = 20$ ROIs). The resulting increase of power of the Hochberg procedure is evident with gain up to 100% for $N_0 \ll N$.

In all simulations the FWE was controlled at 0.05 as the experimentally observed frequencies of rejection of at least one true null hypothesis fell below the upper limit of the 90% confidence interval surrounding the rejection frequency of $\alpha = 0.05$, in this case the interval 0.05 ± 0.005. This interval can be computed by noting that the rejection of the null hypothesis is a binomial event. The number of events (5000 in each simulation) allows the use of the Normal approximation to the binomial distribution and therefore the approximate 90% confidence interval surrounding the rejection frequency $\alpha = 0.05$ is given by the formula $\alpha \pm 1.645[(\alpha)(1 - \alpha)/5000]^{1/2}$ (Noreen, 1989, pp. 34–35).

2. Autoradiographic Study

The *P* plots for the three contrasts and the best-fitting lines through the linear part of the data are shown in Fig. 1. The estimated numbers of true hypotheses, $\hat{N}_0$, were 8, 24, and 17 for the contrasts of control vs diazepam, control vs ketamine, and control vs ketamine-diazepam, respectively. The corresponding estimates of the number of non-null distributions, $N - \hat{N}_0$, were, therefore, 35, 19, and 26, respectively, for the 43 regions of interest. These values indicate that the effects of diazepam are found throughout the brain whereas effects of ketamine, and ketamine and diazepam given in combination, are more regionally selective. The *P* values obtained by Student's *t* tests, before correction for multiple comparisons, are shown in Table 3. By application of the Hochberg procedure, $lCMR_{glc}$ was found to be statistically significantly different (FWE set at $\alpha = 0.05$) from control in 13 brain regions in the diazepam-treated animals, 13 regions in the ketamine-treated animals, and seven regions in the animals treated with both drugs in combination. Use of $\hat{N}_0$ instead of N in the Hochberg procedure resulted in an additional 7 regions in which $lCMR_{glc}$ was statistically significantly altered by diazepam and only one more by ketamine and the combination of ketamine and diazepam. These values reinforce the finding of a generalized effect of diazepam to decrease $lCMR_{glc}$ throughout the brain (Kelly *et al.*, 1986). Ket-

TABLE 3

Significance Level[a] for Comparison of Local Rates of Cerebral Glucose Utilization in Control Rats versus Rats Treated with Diazepam, Ketamine, or Ketamine Plus Diazepam

	Control versus Diazepam (0.5 mg/kg)	Control versus Ketamine (10 mg/kg)	Control versus Ketamine (5 mg/kg) plus Diazepam (0.25 mg/kg)
Auditory system			
Auditory cortex	0.0004[b]	0.0002[b]	<0.0001[b]
Medial geniculate	0.0029[c]	0.0001[b]	<0.0001[b]
Inferior colliculus	0.0547	<0.0001[b]	<0.0001[b]
Superior olivary nucleus	0.2491	0.7735	0.0182
Ventral cochlear nucleus	0.1428	0.7290	0.8534
Lateral lemniscus	0.1766	0.6783	0.0264
Visual system			
Visual cortex	0.3679	0.0544	0.0725
Lateral geniculate	0.0010[b]	0.2734	0.0122
Superior colliculus	0.4176	0.3984	0.0552
Sensorimotor system			
Sensorimotor cortex	<0.0001[b]	0.0428	<0.0001[b]
Ventrolateral thalamic nucleus	0.0021[c]	0.4454	0.0126
Red nucleus	0.0021[c]	0.3199	0.0088
Medial vestibular nucleus	0.1397	0.4311	0.2178
Hypoglossal nucleus	0.0380	0.1837	0.1084
Cerebellar grey matter	0.0013[b]	0.2082	0.0074
Olfactory system			
Olfactory tubercle	0.0140	0.0015[b]	0.3830
Limbic system			
Fornix	0.3016	0.0008[b]	0.0913
Presubiculum	0.0004[b]	<0.0001[b]	0.8002
Cingulum	0.0063	0.0010[b]	0.1036
Retrosplenial agranular cortex	<0.0001[b]	0.0002[b]	0.0059
Cingulate cortex	0.0024[c]	0.0024	0.0544
Anteroventral thalamic nucleus	0.0001[b]	0.0001[b]	0.0300
Mamillary body	<0.0001[b]	<0.0001[b]	<0.0001[b]
Hippocampus: CA3	0.0103	<0.0001[b]	0.4645
Entorhinal cortex	0.0031[c]	0.0028	0.1336
Amygdala	0.7901	0.0020[c]	0.3727
Nucleus accumbens	0.1734	0.8655	0.7009
Interpeduncular nucleus	0.2808	0.0172	0.9784
Medial Habenula	0.0828	0.6182	0.0503
Lateral Habenula	<0.0001[b]	0.0009[b]	<0.0001[b]
Basal ganglia			
Caudate putamen	0.0089	0.1440	0.7792
Substantia nigra, pars compacta	0.0231	0.0472	0.8507
Substantia nigra, pars reticulata	0.0527	0.1297	0.3464
Globus pallidus	0.8311	0.0291	0.5119
Association areas			
Prefrontal cortex	0.0018[c]	0.4943	0.0526
Frontal cortex	0.0001[b]	0.8544	0.0021[c]
Hypothalamus			
Paraventricular nucleus	0.0960	0.7197	0.0190
Supraoptic nucleus	<0.0001[b]	0.1393	<0.0001[b]
Myelinated fiber tracts			
Corpus callosum	0.0045[c]	0.0001[b]	0.0084
Anterior commissure	0.0398	0.0965	0.0640
Hippocampal fimbria	0.0007[b]	0.7157	0.0037
Cerebellar white matter	0.0149	0.7525	0.0313
Whole brain weighted average	0.0006[b]	0.2777	0.0105

[a] Determined by Student's *t* test for difference of means. *P* values are not corrected for multiple comparisons. (Values reported in Entrei *et al.* (1999) were corrected for the number of groups compared.)

[b] Hypothesis rejected by the Hochberg procedure with FWE set at $\alpha = 0.05$.

[c] Additional hypotheses rejected by the modified Hochberg procedure with the number of true null distributions estimated by the graphical *P* plot method. FWE set at $\alpha = 0.05$.

amine effects were most dramatic in the limbic system in which $lCMR_{glc}$ was increased in 9 of 14 regions examined and in the auditory system in which $lCMR_{glc}$ was decreased. These results agree with the proposal that limbic seizures occur in response to ketamine (Winters *et al.,* 1972) and the reported depressant effects of ketamine on auditory function (Crowther *et al.*, 1990; Dodd and Capranica, 1992). When the two drugs were administered together the dramatic increases in $lCMR_{glc}$ found with ketamine alone were reversed, and in mamillary bodies and the lateral habenula rates were even statistically significantly below those found in controls. These results are consistent with the clinical observation that diazepam given in conjunction with ketamine reduces the hallucinations and vivid dreams seen in emergence from ketamine anesthesia while maintaining the anesthetic effect.

DISCUSSION

Functional imaging methods have become widespread investigative tools because they can examine activity simultaneously in all areas of the brain, but the large numbers of regions or voxels in a functional imaging data set present a continuing challenge to develop statistical procedures that are sufficiently powerful yet control the number of false positive findings. In the present study we focused on the problem of analysis of multiple independent statistics obtained by neuroimaging sampling techniques; these techniques may include, for example, ROI analysis or voxel-by-voxel methods in wavelet space. A method from the statistical literature for analysis of a collection of univariate statistical tests, the *P* plot, was introduced. The *P* plot is a graphical method that allows the estimation of the number of sites of the image that are likely to be generated by a null distribution; this estimate can be used for general inference or for sharpening Bonferroni-type corrections.

In this report the application of the method was refined by use of a more rigorous statistical framework than the one previously available. The technique was tested through simulations that showed its ability to control the overall error rate while enhancing the power of the analysis especially when the number of "true" null distribution is small. Clearly this will be the case in those experimental settings where the measured treatment effect is expressed widely over the brain, encompassing many voxels or anatomical regions.

An autoradiographic data set was selected to illustrate the effect of such increase in power in the limiting conditions of a small sample size and a large number of testing sites. Application of the *P* plot resulted in finding statistically significant changes in brain metabolism due to anesthetic agents in a wider range of structures, a finding that is more congruent with published literature.

Finally, we remark that although the algorithms developed here are limited to the case of multiple independent statistics, the *P* plot is not and its extension to the dependent case represents an interesting topic for future work.

ACKNOWLEDGMENTS

The authors thank Dr. Louis Sokoloff for fruitful discussion and constructive comments. This project was developed in part when F.E.T. was postdoctoral fellow at the Laboratory of Cerebral Metabolism supported by a grant from the Wallerstein Foundation for Geriatric Life Improvement (West Orange, NJ); this support is gratefully acknowledged.

REFERENCES

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

Brown, B. W., and Russell, K. 1997. Methods correcting for multiple testing: Operating characteristics. *Statist. Med.* **16:** 2511–2528.

Crowther, J. A., Miller, J. M., and Kilney, P. R. 1990. Effect of anesthesia on acoustically evoked middle latency response in guinea pigs. *Hearing Res.* **43:** 115–120.

Csörgo, M., and Horváth, L. 1997. *Limit Theorems in Change-Point Analysis.* Wiley, New York.

D'Agostino, R. B., and Stephens, M. A. 1986. *Goodness-of-Fit Techniques.* Marcel Dekker, New York.

Dodd, F., and Capranica, R. R. 1992. A comparison of anesthetic agents and their effects on the response properties of the peripheral auditory system. *Hearing Res.* **62:** 173–180.

Duncan, D. B. 1965. A Bayesian approach to multiple comparisons. *Technometrics* **7:** 171–222.

Dunnet, C. W., and Tamhane, A. C. 1991. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statist. Med.* **10:** 939–947.

Dunnet, C. W., and Tamhane, A. C. 1992. A step up multiple test procedure. *J. Am. Statist. Assoc.* **87:** 162–170.

Dunnet, C. W., and Tamhane, A. C. 1995. Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics* **51:** 217–227.

Entrei, C., Sokoloff, L., and Smith, C. B. 1999. Effects of diazepam and ketamine administered individually or in combination on regional rates of glucose utilization in the rat brain. *Br. J. Anaesthesia* **82**(4): 596–602.

Ford, I., McColl, J. H., McCormack, A. G., and McCrory, S. J. 1991. Statistical issues in the analysis of neuroimages. *J. Cereb. Blood Flow Metab.* **11:** A89–A95.

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J. 1991. Comparing functional (PET) images: The assessment of significant change. *J. Cereb. Blood Flow Metab.* **10:** 458–466.

Hegazy, Y. A. S., and Green, J. R. 1975. Some new goodess-of-fit tests using order statistics. *Appl. Statist.* **24:** 299–308.

Hochberg, Y., and Tamhane, A. C. 1987. *Multiple Comparison Procedures.* Wiley, New York.

Hochberg, Y. 1988. A sharper Bonferoni procedure for multiple tests of significance. *Biometrika* **75:** 800–803.

Hochberg, Y., and Benjamini, Y. 1990. More powerful procedures for multiple significance testing. *Statist. Med.* **9:** 811–818.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6:** 54–70.

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 Metab.* **16:** 7–22.

Kelly, P. A. T., Ford, I., and McCulloch, J. 1986. The effect of diazepam upon local cerebral glucose use in the conscious rat. *Neuroscience* **19**(1): 257–265.

Marcus, R., Peritz, E., and Gabriel, K. R. 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63:** 655–660.

Noreen, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction.* Wiley, New York.

Poline, J.-P., and Mazoyer, M. 1994. Enhanced detection in brain activation maps using a multifiltering approach. *J. Cereb. Blood Flow Metab.* **14:** 639–642.

Quesenberry, C. P., and Hales, C. 1980. Concentration bands for uniformity plots. *J. Statist. Simul. Comput.* **11:** 41–53.

Rothman, K. J. 1990. No adjustments are needed for multiple comparisons. *Epidemiology* **1:** 43–46.

Saville, D. J. 1990. Multiple comparison procedures: The practical solution. *Am. Statist.* **44:** 174–180.

Schweder, T., and SpjØtvoll, E. 1982. Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69:** 493–502.

Snedecor, G. W., and Cochran, W. G. 1980. *Statistical Methods,* 7th ed. Iowa State Univ. Press, Ames, Iowa.

Sokoloff, L., Reivich, M., Kennedy, C., DesRosiers, M. H., Patlak, C. S., Pettigrew, K. D., Sakurada, O., and Shinohara, M. 1977. The [^{14}C]deoxyglucose method for the measurement of local cerebral glucose utilization: Theory, procedure, and normal values in the conscious and anesthetized albino rat. *J. Neurochem.* **28:** 897–916.

Soric, B. 1989. Statistical "discoveries" and effect size. *J. Am. Statist. Assoc.* **84:** 608–611.

SpiØtvoll, E. 1972. On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43:** 398–411.

Troendle, J. F. 1996. A permutational step-up method of testing multiple outcomes. *Biometrics* **52:** 846–859.

Turkheimer, F. E., Brett, M., Aston, J. A. D., Leff, A. P., Sargent, P. A., Wise, R. J. S., Grasby, P. M., and Cunningham, V. J. 2000. Statistical modeling of PET images in wavelet space. *J. Cereb. Blood Flow Metab.,* **20:** 1610–1618.

Westfall, P. H., and Young, S. S. 1993. *Resampling-Based Multiple Testing.* Wiley, New York.

Winters, W. D., Ferrar-Allado, T., Guzman-Flores, C., and Alcaraz, M. 1972. The cataleptic state induced by ketamine: A review of the neuropharmacology of anesthesia. *Neuropharmacology* **11:** 303–315.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. 1992. A three-dimensional analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12:** 900–918.

Worsley, K. J., Marret, S., Neelin, P., and Evans, A. C. 1996. Searching scale space for activation in PET images. *Hum. Brain Mapp.* **4:** 74–90.