NeuroImage **10,** 1–5 (1999)
Article ID nimg.1999.0439, available online at http://www.idealibrary.com on IDEAL®

COMMENTS AND CONTROVERSIES

How Many Subjects Constitute a Study?

Karl J. Friston, Andrew P. Holmes, and Keith J. Worsley

The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, United Kingdom

Received January 26, 1999

In fMRI there are two classes of inference: one aims to make a comment about the "typical" characteristics of a population, and the other about "average" characteristics. The first pertains to studies of normal subjects that try to identify some qualitative aspect of normal functional anatomy. The second class necessarily applies to clinical neuroscience studies that want to make an inference about quantitative differences of a regionally specific nature. The first class of inferences is adequately serviced by conjunction analyses and fixed-effects models with relatively small numbers of subjects. The second requires random-effect analyses and larger cohorts.

Key Words: **fMRI; inference; fixed and random effects; conjunctions; typical.**

typical /*tIpIc(a)l*/ *adj.* 2. characteristic of or serving to distinguish a type.
average /*ævarIg$_5$*/ *n., adj.,* & *v., — adj.* 1. usual, ordinary. 2. estimated or calculated by average. (*The Concise Oxford Dictionary,* OUP 1991)

INTRODUCTION

With the increasing use of fMRI many units, including our own, must contend with practical constraints on the number of subjects that constitute a study. These constraints are imposed by data-handling restrictions and access to scanning time. The following comments are based on discussions about the scientific motivation for scanning large numbers of subjects.

THE NATURE OF INFERENCES WE MAKE

A critical distinction that determines the number of subjects included in a neuroimaging study is between inferences about the particular subjects studied and inferences that pertain to the population from which those subjects came. This is equivalent to the distinction between *fixed*- and *random*-effect analyses and speaks to the notion that, in making inferences at the population level, we must account for having only a *sample* of subjects from the population. The distinction arises only when one has several observations from each subject or session. This is clearly the case in fMRI and engenders the following problem: Imagine that one has studied six subjects. These subjects are analyzed using a conventional, fixed-effect, statistical model using subject-specific parameter estimates (i.e., a design matrix that is separable over subjects or sessions). In this instance one can specify contrasts testing for an activation in each subject separately or for an average activation over the group. Let us say that the subject-specific contrasts show that two of the six subjects activate very significantly, whereas the remainder do not. By virtue of these two subjects the contrast testing for an average activation over the six subjects also shows significant effects. What then can be said about the population from which these six subjects came? In short, nothing. All that can be said is that relative to the within-session (i.e., scan to scan) variability, the activation expressed by two subjects is large and consequently the average activation over all six is significant. This allows one to infer that these and only these six subjects show an average activation and, in this instance, this average effect is accounted for by two of the six subjects. To assess the probability that a subject sampled randomly from the population would show an activation one needs to know the variability of the activation itself. This is simply obtained by the variation in activation from subject to subject. Clearly, in this example, this variation would be large relative to the average activation, which, as a result, would be deemed insignificant. The latter analysis corresponds to a random-effect analysis where the "random" speaks to the fact that one has allowed for the expression of each subject's activation to be modeled as a random variable. In the subject-separable fixed-effect analysis described above these differential contributions were modeled implicitly, in terms of subject by condition interactions, but were assumed to represent "fixed" parameters of the statistical model employed. Put

1053-8119/99 $30.00

simply, if one wants to make an inference about a particular subject or group of subjects then one compares the average activation to the within-subject or session variability. To make an inference about the population from which these subjects came, then one compares the average activation to the variability of that activation over subjects. The implication is that to generalize one's inference to the population one must have a large number of subjects to reliably assess the between-subject variability. This is the critical point of contact between the nature of inference required and the number of subjects required.

In what follows we will review the status of fixed- and random-effect analyses in relation to fMRI data analysis, focusing specifically on inferences that can be made about the population from which the subjects were drawn. The main aim of this discussion is to demonstrate that fixed-effect models can be used to make inferences at a population level through the use of conjunction analyses and the notion of *typicality*.

TYPICAL AND AVERAGE CHARACTERISTICS

In fMRI the scan to scan variability is generally much lower than the session to session variability and the distinction between a random- and fixed-effect analysis is crucial. This distinction is made more acute by the fact that the ratio of scans to subjects is generally high, rendering a greater difference between fixed- and random-effect analyses (the emphasis on between-subject effects increases with this ratio in random-effect analyses). It is worthwhile considering the various approaches adopted in other disciplines faced with similar problems. One more useful example is the distinction between electrophysiology in awake behaving primates and in human event-related studies. In electrode recording it is quite common, and acceptable by peer review, to present a careful statistical characterization of a single monkey and supplement these observations with replications in a further one or two subjects. From the perspective of fMRI this corresponds to a fixed-effect analysis or case study of two to three subjects. In contradistinction, in most human studies, evoked responses are averaged within subject and then analyzed. The number of subjects here is generally much greater (e.g., 8–16) and the inference is based explicitly on subject to subject response variability. This corresponds to a random-effect analysis using 8 to 16 subjects. Which is the most appropriate for fMRI studies? Clearly the answer to this question depends upon the experimental question and leads us to think carefully about the nature of the inference that we want to make.

These inferences can fall into one of two classes. First, we are trying to delineate, in a qualitative sense, some *typical* aspect of functional architecture in the human brain or, second, we may want to make a quantitative statement about some *average* trait that characterizes a particular cohort of subjects or patients. The distinction is subtle but crucial from a statistical perspective. For example, the inference that farmers *typically* own tractors is not refuted by the fact that some farmers do not have a tractor. It only implies that farmers are more likely than not to own one. On the other hand, the statement that farmers in Wales have, on *average*, more than 0.86 tractors has a different connotation and provides a more refined characterization in terms of the quantity of tractors owned. The distinction can be appreciated in terms of the corresponding null hypotheses. For *typical* characteristics the null hypothesis is that less than a specified proportion of the population evidences the trait. For average traits the null hypothesis is that the average expression of the trait is not significantly different from some specified level (e.g., 0 or, in the example above, 0.86). Note that the trait or characteristic is treated as a *qualitative* variable from a typicality perspective (i.e., one does or does not own a tractor) and as a *quantitative* metric in terms of average traits (i.e., one can own one or more tractors). The reason that fMRI lends itself to this dichotomy is that any neurophysiological effect can be inferred to be present or absent (in a statistical sense using a single-subject analysis) or characterized in terms of the effect itself (the parameter estimates of the effect's size).

A more relevant example, of the distinction between inferences about typical and average characteristics, would be the difference between asking "do normal subjects *typically* active their left prefrontal cortex during the encoding of semantic material?" and "do schizophrenic patients with psychomotor poverty show, on *average*, lower left prefrontal cortical activation during semantic encoding than those with reality distortion?" In the first case one would be satisfied with an inference that a neurophysiological response was typical of the population from which the subjects were drawn, accepting that the occasional subject may not show this effect. In the second case one wants to make an inference of a quantitative sort that is always demonstrably true, given enough subjects.

This difference between inferences about typical and average responses is important because fixed-effect analyses can, contrary to convention, be used to make inferences about typical characteristics at a population level. To make inferences about average characteristics one needs to use random-effect analyses. The next subsection describes how fixed-effect models can be used to make population inferences through the use of conjunction analyses.

FIXED-EFFECT MODELS: CONJUNCTIONS AND INFORMATION

Here we consider conjunction analyses in subject-separable fixed-effect statistical models. The term "conjunction" is used to denote the joint refutation of two or more null hypotheses. More simply, a conjunction of effects arises when this effect is significant in every subject-specific contrast. Note that, because of the subject-separable nature of the design matrix, the ensuing contrast of parameter estimates is orthogonal and can be construed as approximately independent. To make the role of conjunctions clear consider the following example:

Suppose that I told you that men from Caius College Cambridge have a small *g* tattooed discreetly in their inguinal region. If you selected a man at random from Caius, and we indeed confirmed that he had a small tattoo, would you believe me? You may or may not. Say that we then selected six men from Caius at random and they all had similar tattoos. You then might be more convinced. Clearly this does not constitute definitive evidence that all men from Caius are embellished in this way, but you have available to you evidence that conveys a substantial amount of information about how typical this attribute is, in relation to men from Caius. We can formalize this anecdotal example in the following way:

Let $p(n)$ be the probability of n subjects testing conjointly for some effect. The test has a *specificity* of α and *sensitivity* of β. Specificity is simply the probability of getting a positive result under the null hypothesis that the subject does not show the effect in question. The sensitivity is the probability of a positive result under the alternate hypothesis that he/she does. Generally we can specify α because the behavior of the observations, under the null hypothesis, is known, whereas we do not know β because the exact nature of the data under the alternate hypothesis is generally not. Now let γ represent the number of subjects in a population showing the effect. The probability of getting a conjunction of positive tests is the sum of the conditional probabilities of these positive results over all possible outcomes of the selection of n subjects. For one subject the possible outcomes are o_i where o_0 is "no effect" and o_1 means the effect is expressed, i.e., $p(o_0) = 1 - \gamma$ and $p(o_i) = \gamma$. The probability of a conjunction of n tests is given by

$$p(n) = [\Sigma p(1|o_i)p(o_i)]^n = [\alpha(1 - \gamma) + \beta \cdot \gamma]^n . \quad (1)$$

The information of this conjunction is simply $-\log(p(n))$. Clearly the information depends upon our prior expectations that any subject chosen at random would express this affect, namely, γ. In the example above the likelihood of any one having a small *g* tattoo is relatively small (say one in a million). Assume that the specificity and sensitivity of our ability to find the tattoo was 5 and 90%, respectively, then the probability of getting positive results from six men would be 0.0000000156 and the corresponding information would be 25.9 bits. This information would be substantially reduced if γ were equal to 0.5. This might be the case if we noted that six men from Caius were all taller than 182 cm, where the probability of being taller than 182 cm was 0.5. In this instance $p(6) = 0.0115$ and the information would only be 6.4 bits.

It can be seen that the information that obtains from a conjunction, based upon a fixed-effect model, depends upon the prior expectations γ. Unfortunately in neuroimaging we have no principled way, other than measuring large numbers of subjects, of determining γ for a particular regionally specific effect. However, by formulating a null hypothesis at the population level, in terms of γ, there is a lower bound on the information that obtains from any conjunction. If we define a typical characteristic of human brain functional architecture as that which is expressed in more than a certain proportion (γ_c) of normal subjects, for example, 50% then, under the null hypothesis, $\gamma < \gamma_c$. The lower limit on the information corresponds to the maximum value of $p(n|\gamma < \gamma_c)$ that, in turn, obtains when γ reaches its upper limit of γ_c.

$$p(n|\gamma < \gamma_c) < [\alpha \cdot (1 - \gamma_c) + \beta \cdot \gamma_c]^n. \quad (2)$$

$p(n|\gamma < \gamma_c)$ is effectively p value for a population inference about how typical the effect is. In other words, under the null hypothesis, the probability of getting a conjunction over n subjects is less than, the value of the term of the right-hand side of Eq. (2) (see Fig. 1 for an example with $n = 3$ subjects). Figure 1 suggests that even three subjects are sufficient to make a population inference, if we accept a suitably low value for γ_c and the sensitivity of our test is small enough (see below). Conversely for any specified γ_c there a lower limit on n that renders the upper limit on the p value in Eq. (2) suitably small, say 0.05, where the critical value of n is given by

$$n_c = \log(0.05)/\log(\alpha \cdot (1 - \gamma_c) + \beta \cdot \gamma_c). \quad (3)$$

Note that the critical number of subjects n_c is a function of specificity, sensitivity and the criteria defining what is a typical characteristic γ_c. A plot of this function for various values of β is given in Fig. 2. If β is not known it is assumed to be 1. This analysis suggests that conjunctions, in the context of fixed-effect analyses, can be used to verify the alternate hypothesis that a particular regionally specific effect is typical of normal functional anatomy. Typical is operationally defined as being expressed in γ_c of the population or more. For example,

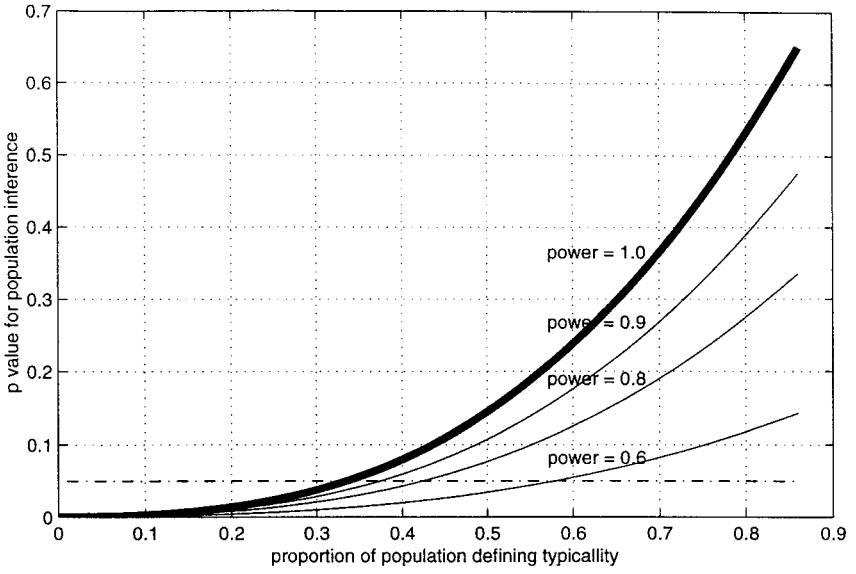

FIG. 1. Plot of the upper limit on $p(n|\gamma < \gamma_c)$, the p value for a population inference using a conjunction and fixed-effect analysis for just three subjects with a test that has 5% specificity and a range of sensitivities ($\beta = 1$, 0.9, 0.8, and 0.6). The ordinate γ_c is the critical proportion of the population that defines "typicality" of the effect tested for. The broken line represents a threshold of $p = 0.05$.

according to Fig. 2, a conjunction over six subjects enables one to say that over 60% of the population are likely to show this effect (irrespective of the test's sensitivity). It should be noted, from Fig. 1, that reducing the power of the test enables smaller numbers of subjects to justify an inference of typicality. For example, with a power or sensitivity of 0.6, three subjects are sufficient to say that a particular activation is typical of a population with a γ_c of about 60%. This somewhat paradoxical result follows from the fact that the actual likelihood of obtaining this conjunction, with a less sensitive test, is exceedingly small.

CONCLUSION

All neuroimaging studies, and ensuing inferences, aim to make some comment about the population from which the subjects studied were sampled. There are two classes of inference. The first aims to establish the observed effect as a typical characteristic of the popula-

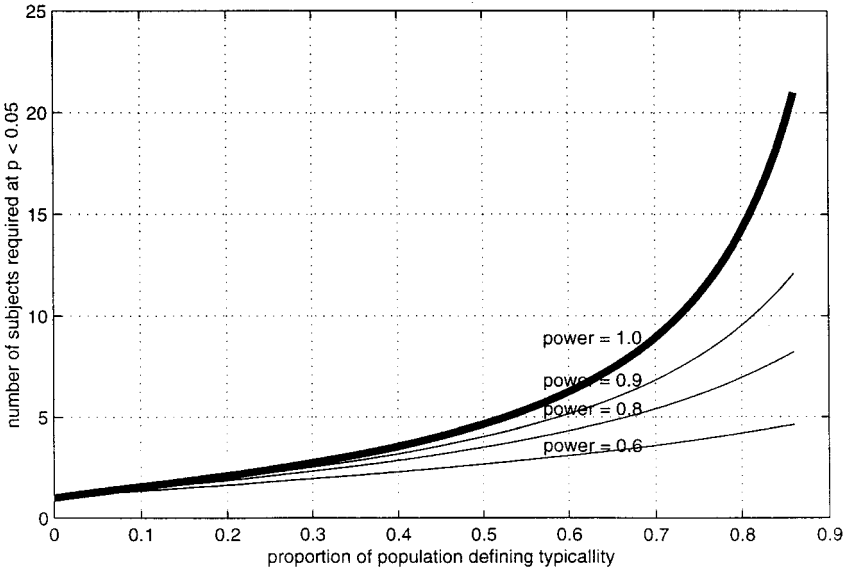

FIG. 2. Plot of the critical number of subjects required in a conjunction analysis using a fixed-effect model to ensure $p(n|\gamma < \gamma_c) < 0.05$ for a test with 5% specificity and a range of sensitivities ($\beta = 1$, 0.9, 0.8, and 0.6). The ordinate γ_c is the critical proportion of the population that defines "typicality" of the effect tested for.

tion, while allowing for the fact that some subjects may not show this effect. This sort of inference may be entirely sufficient when trying to characterize generic aspects of human functional brain architectures, sufficient in the sense that knowing a particular characteristic is typical is more useful than not knowing this fact. These sorts of inferences are obtained from a conjunction analysis using fixed-effect models. The second class of inferences is more stringent in the sense that the alternate hypothesis requires that the mean effect over the population is significantly greater than under the null hypothesis. This sort of inference requires a random-effect analysis and would be necessary in many examples from clinical neuroscience, where the effect should have some diagnostic or predictive validity. The "effect" is treated as qualitatively present or absent using conjunction analyses. The effect in random-effect analyses of average responses enters as a quantitative variable. By virtue of this a conjunction approach to establishing typicality can be seen as an inference (at a population level) about an inference (at a subject level). In other words, it is a metaanalysis.

In short, in basic imaging neuroscience, it may be the case that a conjunction analysis with a fixed-effect model is sufficient to infer something about characteristics that are typical of a population, whereas in clinical neuroscience it may well be necessary to use a random-effect analysis. By allowing for a more relaxed but operationally specified definition of typicality one can motivate the use of conjunctions and fixed-effect models and harness their greater sensitivity (that is obtained by predicating the inference on greater degrees of freedom).

Note that there are experimental designs that can only be analyzed using random-effect analyses. These include fMRI designs where there is no true replication of treatments within a subject. The more important examples of these involve learning experiments and psychopharmacological studies. Learning experiments, over protracted periods of time (as opposed to within-session adaptation), require random-effect inference because the treatment only exists on a subject- or session-specific level. Similarly, psychopharmacological studies, especially those employing antagonists whose receptor binding kinetics have long time constants, require random-effect analysis because the differences of interest only exist among sessions and not within them.

ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust. We thank Theresa Calvert for help preparing the manuscript.