Repeated Challenge Studies: A Comparison of Union-Intersection Testing with Linear Modeling

Richard A. Levine[2] Pamela A. Ohman[3]

Cornell University

BU-1279-MA March 1995
Revised March 1996
To appear in *Psychometrika*

[1]Research supported by the Office of Naval Research Graduate Fellowship and NIH Training Grant No. ES07261

[2]Research supported by NSF Training Grant No. DMS956682

Acknowledgments: We would like to thank Professor George Casella and Susan Alber for many helpful discussions and advice during the course of this research and for comments on drafts of this manuscript. We would also like to express gratitude to Lorna Bayer and Professor Barbara Strupp for providing the data on dopamine exposure in rats. Finally, we would like to thank three referees and the editor for very helpful and insightful remarks that lead to a much improved version of this paper.

ABSTRACT

Challenge studies are often implemented for assessing whether a subject is sensitive to a certain agent or allergen. In particular, researchers test groups of subjects to determine if there really exists a causal relationship between some agent of interest and a response. To answer such a question, we need to detect the presence of the phenomenon in just one individual. Typically, however, there are a large number of unknown risk factors associated with the response and a potentially small population prevalence. Hence, standard statistical techniques, by averaging the treatment effect across the group, may miss a significant response of a single individual and lead to inconclusive results. We develop an alternative approach based on union-intersection testing that will allow a practitioner to correctly examine observations on an individual apart from the other subjects and test the hypothesis of interest: does the phenomenon exist in the population. More specifically, we show how this technique adjusts for the multiple number of tests encountered when analyzing data for each individual subject separately. Furthermore, we demonstrate power calculations for the determination of sample size prior to performing the study. The performance of the union-intersection approach in comparison to linear models and semi-parametric techniques is considered through sample size calculations and simulations. The union-intersection testing methodology out performs the Kolmogorov tests. However, the nested linear model performs as well if not better than the union-intersection tests. To illustrate the ideas presented in the paper, we provide an application in which we analyze psychological data collected by way of a challenge study design.

Key Words: hypothesis testing, multiple tests, power analysis, population prevalence, Kolmogorov tests, application

AMS 1990 Subject Classification: Primary 62F03; Secondary 62K99

1 Introduction

What statistical procedures should a researcher use in order to test whether a particular agent can cause measurable responses when it is expected that only a small fraction of the subjects under study will respond to the agent? Usually, the majority of subjects in a study are expected to respond to the agent in question. For example, smoking is thought to result in decreased lung capacity. This response would be expected in all individuals regardless of other possible risk factors. On the other hand, one does not expect all persons who smoke to develop cancer. There are risk factors other than smoking that are involved in the development of cancer. Some of these risk factors are known, but there may also be some unknown factors. What if the risk factors for the response in question are mostly unknown? It would then be nearly impossible to select a sample in which all subjects would give the expected response. In fact, if the number of responders in the general population is very small and it is unknown what causes a person to be a "responder," the researcher may have difficulty even showing that the phenomenon exists.

This problem may arise in many studies directed at establishing a link between food additives and child behavior. Feingold (1975) suggested that food colors may be associated with hyperactivity in children. However, many subsequent attempts to study the hypothesized relationship have offered inconclusive results (see Prinz (1985) and Van-Dusseldorp (1989) for reviews of the literature). The large number of potential causes of hyperactive behavior (Pollock and Warner (1990), Prinz (1985)) and the conceivably small population prevalence may explain the lack of significant results in the literature. In this paper, we will present a methodology for directly testing whether a particular rare phenomena with unknown risk factors exists (i.e., prevalence is greater than zero) given that we can test any individual multiple times.

Tests of the hypothesis that the prevalence of a phenomenon is zero occur in the allergy, asthma, and food sensitivity literature. The repeated challenge study (Prinz (1985), Warner (1987), and Metcalfe and Sampson (1990)) is one study design that has been used to test such hypotheses. This design consists of repeatedly challenging subjects with the agent under consideration (e.g. food additives) and with a control substance (placebo). Following

each challenge, a response (e.g. behavior) is measured. At the conclusion of all the challenges, the scientist checks for a statistically significant difference between the placebo and non-placebo responses. If such a disparity occurs, we may infer that there is a significant relationship between the agent and the response (e.g. an association between food additives and hyperactivity). The challenge study design may be thought of as a generalization of a crossover design whereby each subject receives the placebo and agent on alternating sessions beginning from a random starting point. (Note that we should not confuse the crossover design with the analysis.) For examples of challenge studies see Behar et al. (1984), Rosén et al. (1988), Rowe (1988), Roshon and Hagen (1989), and Pollock and Warner (1990).

Typically, the observations from a challenge study have been analyzed using an analysis of variance (ANOVA) with repeated measures or crossover analysis techniques (related to the ANOVA). Such analyses, while taking into account that the observations for an individual subject are correlated over time, average the treatment effect across the group of subjects. The treatment effect here refers to the difference in response between the placebo (control) and non-placebo (agent) challenges. Hence, these approaches consider the average treatment effect for the group in making the decision of whether or not the phenomenon exists. If in fact the true prevalence of the phenomena under study is minuscule, only a small fraction of the subjects may be significantly effected by the agent or treatment. The ANOVA techniques will interpret the significant responses of the few true responders as merely part of the chance variation in the observations. Thus, these analyses may conclude no significant relationship between the agent under study and response even though some individuals in the group display a true significant treatment effect. Consequently, the typical analyses may mask a real effect and lead to the incorrect inference of no association between agent and response for all individuals.

The central problem with the ANOVA/crossover approach is that, by combining all of the data into one statistic, they are testing the wrong hypotheses. The ANOVA procedure tests against the alternative hypothesis that the average treatment effect is significantly different than zero. If the alternative hypothesis is true, then we can conclude that, in at least a few of the subjects (and probably in the majority), the phenomenon exists.

However, given the effect of the treatment varies from subject to subject, we gain no insight into the effect of the treatment on any one individual if the test does not conclude that the average treatment effect is different than zero. Yet, if there is at least one individual for whom the true treatment effect is not zero, then we would like to reject the null hypothesis and conclude that the prevalence of the phenomenon is positive. Thus, for testing the null hypothesis that the prevalence of a theorized phenomenon is zero, we should study each individual and establish if the observations for the particular individual indicates a treatment effect (specific to that subject) which is significantly different than zero. This can be accomplished through a technique called union-intersection testing. This scheme allows us to analyze observations on each individual subject separately rather than at the group level. Thus, we will be able to make inferences about the hypothesis of interest.

We should note that the above comments are applicable for statistical analyses of many studies outside the repeated challenge study design framework. For example, often in studies dealing with perception, learning, memory, and cognition, the psychological hypotheses of interest imply some law holds for a single individual, not just averages computed across individuals. Therefore, the methods developed in this paper are useful for behavioral research in general.

As an alternative to the ANOVA-type approach, we first introduce a technique based on union- intersection tests. In Section 2 we present the methodology and power calculations for the union- intersection testing approach. Note that we are considering *repeated* challenge studies here instead of an experiment in which the subjects are challenged with the placebo and agent once each. The reason for such a design is that, when testing each subject independently, a proper diagnosis for an individual can be reached only by repeated administrations of the placebo and agent (Pearson (1985), Milich et al. (1986), Warner (1987), and Metcalfe and Sampson (1990)). The difficulty is, how many challenges and how many subjects are necessary to obtain a desired power for the tests? This question is answered in Section 3. Section 4 compares the union-intersection methodology with alternative approaches through sample size calculations and simulations. Section 5 presents an analysis of a data set dealing with the effect of dopamine on behavior in rats to emphasize

and illustrate some results from the previous sections.

2 Union-Intersection Tests

When conducting a challenge study, the hypothesis of interest is

$$H_0 : \phi_g = 0 \text{ vs. } H_A : \phi_g > 0,$$

where ϕ_g is the prevalence of the phenomenon in the general population. In testing this hypothesis, we will take a sample of N subjects and for each one test whether or not they exhibit the phenomena. In other words we test whether the population from which the sample is drawn has a prevalence of zero or not. Therefore, we are testing the hypothesis

$$H_0 : \phi = 0 \text{ vs. } H_A : \phi > 0, \tag{1}$$

where ϕ denotes the prevalence in the sample. It is clear that

$$\phi > 0 \Rightarrow \phi_g > 0 \tag{2}$$

since the individuals in the sampled population are part of the general population.

The population from which one selects a sample is assumed to have a prevalence ϕ which is greater than or equal to the prevalence in the general population, i.e., $\phi \geq \phi_g$. This assumption is reasonable because, in general, a challenge study consists of subjects who believe they are sensitive to the treatment. For instance, in trying to select children who are hyperactive after ingesting a particular food additive, it is possible to start with a prevalence that is higher than that for the rest of the population by only selecting children whose parents believe them to be sensitive. The self-selected sample will ensure that the prevalence in the sample is greater than the prevalence in the general population. Consequently, if we fail to reject the hypothesis that the sample prevalence is zero, then we can conclude that the population prevalence is zero as well. In mathematical terms, if $\phi = 0$ and $\phi \geq \phi_g$, then

$$\phi = 0 \Rightarrow \phi_g = 0 \tag{3}$$

Equations (2) and (3) imply that the hypothesis concerning the prevalence of the general population and the hypothesis concerning the prevalence of the sample are interchangeable.

Hence, from this point on in our discussion, we will refer only to the hypothesis of the sampled prevalence (1).

Under the null hypothesis of (1), we expect none of the subjects to show a response to the challenges. If we conclude that at least one of the subjects does show a response, then the prevalence of the sampled population must be greater than zero. Therefore, at the experimental level, we actually test the hypothesis

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ vs. } H_A : \text{at least one } \theta_i \neq 0, \tag{4}$$

where the ith element of the $N-$ dimensional vector $\boldsymbol{\theta}$ is the treatment effect for the ith subject. Rejection of the null hypothesis of (4) implies rejection of the null hypothesis (1).

Notice that the null hypothesis in (4) can be rewritten as the intersection of the null hypotheses from the individual tests of the treatment effect. In other words, since individual hypotheses are of the form

$$H_0^i : \theta_i = 0 \text{ vs. } H_A^i : \theta_i \neq 0, \quad i = 1, \ldots, N, \tag{5}$$

we can write H_0 as $\{\boldsymbol{\theta} : \theta_i = 0 \text{ for all } i\} = \bigcap_{i=1}^{N} \{\theta_i = 0\}$. Consequently, the rejection region for testing H_0 is the union of the rejection regions for each of the individual tests (5). That is, H_0 is rejected if any one of the subjects shows a significant treatment effect, i.e., if we reject any of the N hypothesis tests in (5). This method of breaking down a complicated hypothesis into the intersection of simpler hypotheses and then constructing a rejection region to be the union of the rejection regions of the simpler hypothesis tests is identified as *union-intersection testing* in Casella and Berger (section 8.2.4, 1990). The union-intersection principle was first introduced by Roy (chapter 2, 1957).

The union-intersection setup allows us to test each individual separately and thereby more easily test the desired hypothesis. Keep in mind however, that the actual α-level and power calculations will still be based on (1) or equivalently (4). The overall α-level will be much greater than the significance level at which we test each individual hypothesis. In performing N separate tests, we have a multiple testing problem. Thus, the alpha levels that should be used for the individual tests will be a function of the number of subjects being tested and the desired overall alpha level of the test. Likewise, the power of the

overall hypothesis test will be a function of the number of subjects N, the powers of the individual hypotheses, and the prevalence ϕ. To find these associations between the powers and α-levels for the overall and individual tests, we will assume that there is independence between subjects, i.e., between the individual tests.

Let us first consider the α-level for testing (1). If prevalence equals zero, then none of the subjects chosen will ever have a true non-zero treatment effect. Therefore

$$\begin{aligned} \alpha_o &= P(\text{reject } H_0 \mid \phi = 0) \\ &= P(\text{at least one } H_0^i \text{ rejected } \mid \phi = 0) \\ &= 1 - P(\text{no } H_0^i \text{ is rejected } \mid \phi = 0) \\ &= 1 - \prod_{i=1}^{N}(1-\alpha) \end{aligned}$$

where α is the alpha level for each of the N individual tests (5). Hence, given the desired overall alpha level, α_0, the alpha levels for the individual tests can be determined by the formula

$$\alpha = 1 - (1 - \alpha_0)^{1/N}. \tag{6}$$

For example, for $N = 5$ subjects we need to set the α-level at 0.01 for the tests on each individual subject in order to attain the 0.05 level of significance for the overall test (4). For an overall alpha level of 0.05, the individual α-level drops off sharply between 1 and 5 subjects and then levels off at very small levels for larger N. In the remainder of the paper, we shall assume an overall alpha level of $\alpha_0 = 0.05$.

Next let us consider the power calculations for testing (1). If the population prevalence is greater than zero, then any number of the subjects may be sensitive to the treatment. In other words, between zero and N of the individual treatment effect parameters, θ_i (see 5), may be truly not equal to zero. Furthermore, the probability that any specific θ_i is not equal to zero is ϕ, the prevalence. Thus, j, the number of subjects who are truly sensitive has a Binomial(N, ϕ) distribution. Let $1 - \beta$ be the power of each individual test, i.e. the probability of accepting the alternative given that $\theta_i \neq 0$. Then the overall power, $1 - \beta_0$, can be determined as follows:

$$1 - \beta_0 = \text{P(reject } H_0 \mid \phi > 0)$$

$$
\begin{aligned}
&= \sum_{j=0}^{N} \mathrm{P}(\text{reject } H_0 \mid j\ \theta_i\text{'s} \neq 0, \phi > 0) \cdot P(j\ \theta_i\text{'s} \neq 0 \mid \phi > 0) \\
&\quad \text{(Bayes theorem)} \\
&= \sum_{j=0}^{N} \mathrm{P}(\text{reject at least one } H_0^i \mid j\ \theta_i\text{'s} \neq 0, \phi > 0) \cdot P(j\ \theta_i\text{'s} \neq 0 \mid \phi > 0) \\
&\quad \text{(Union-intersection test)} \\
&= \sum_{j=0}^{N} \Big(1 - P(\text{reject no } H_0^i s \mid j\ \theta_i\text{'s} \neq 0, \phi > 0)\Big) \cdot P(j\ \theta_i\text{'s} \neq 0 \mid \phi > 0) \\
&= \sum_{j=0}^{N} \Big(1 - [P(\text{do not reject } \mathrm{H}_0^i \mid \theta_i \neq 0)]^j \cdot [P(\text{do not reject } \mathrm{H}_0^i \mid \theta_i = 0)]^{N-j}\Big) \times \\
&\qquad P(j\ \theta_i\text{'s} \neq 0 \mid \phi > 0) \\
&= \sum_{j=0}^{N} \Big(1 - \beta^j (1-\alpha)^{N-j}\Big) \cdot \binom{N}{j} \phi^j (1-\phi)^{N-j}, \qquad (7)
\end{aligned}
$$

This series of equations leads to a general power formula for a union-intersection test given a prevalence ϕ, a number of subjects N, and an α-level and power $1-\beta$ for the individual tests. Unfortunately, a closed form for $1-\beta$ cannot be found as a function of the overall power due to the summation. In addition, we must specify a particular value for the prevalence ϕ, which will usually be a very rough estimate. Also, note that in these formulas we have assumed $1-\beta$ to be constant across individuals. Using derivatives, one can show that an increase of at least one individual's power will result in an increase in the overall power of the test.

To get an idea of how the prevalence affects the overall power, Figure 1 shows for $N = 10$ subjects the change in overall power as ϕ increases from 0 to 0.6 with powers of 0.6, 0.7, and 0.8 for the individual tests (5). Note that the overall power "levels off" for prevalences of 0.5 or greater. Therefore, the experimenter planning to sample ten subjects should strive to obtain a sample such that the prevalence among the subjects is believed to be at least 0.5. He/she should not, however, worry too much about getting a prevalence very much higher than 0.5 since that will not significantly improve the overall power. Similarly for $N = 5$ subjects (Figure 2), the power "levels off" after a prevalence of 0.7. At $\phi = 0.6$, the overall power is greater than 0.9 in all three cases. Therefore, since overall power reaches the 0.9 level more quickly (in relation to prevalence) for ten than for five subjects, one would want

to use more subjects if the experimenter expects the prevalence to be low. Recall that in general we will have a self-selected sample, thus resulting in a larger sample prevalence as compared to the overall population prevalence. Hence, we hope to usually find ourselves near the asymptotes of the curves in Figures 1 and 2, i.e., the overall power will be high.

3 Sample Size Determination

In planning any study, an essential question is how big should the sample be? In the case of the challenge study, the "sample size" has two components: the number of subjects to include in the study and the number of visits to plan (or number of observations) for each subject. The first step in determining these two components is to choose an overall $\alpha-$level and power and estimate the prevalence for the pool of potential subjects. Having made these decisions, particular combinations of values of N, the number of subjects, and values for the individual powers will result in the appropriate $\alpha-$level, power, and estimated prevalence (see Table 1). For example, with $\alpha = 0.05$ and the prevalence $\phi = 0.50$, we find that testing three individuals, each at a power of 0.81, and testing four individuals, each at a power of 0.65, will both achieve the same desired overall power of 0.80.

After determining how many individuals to test at an individual power using Table 1, the experimenter can calculate the number of visits needed for each subject given the number of subjects, N, in the study and a desired overall power. These power calculations will depend on the kind of data that is collected and the analysis that is planned for each test. In the present context, the response variable is assumed to be a single continuous variable. The test to be conducted on the individual subject's data is an F-test.

More formally, consider the general linear model

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\begin{pmatrix} \mu \\ \theta_i \end{pmatrix} + \boldsymbol{\epsilon}, \\ &= \mathbf{X}\boldsymbol{\tau}_i + \boldsymbol{\epsilon} \end{aligned} \tag{8}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, θ_i is the treatment effect, $\mathbf{Y}$ is the response, and $\mathbf{X}$ is the design matrix for the experiment on the ith individual. The generalized least squares estimators

are (section 5.8, Searle (1971))

$$\hat{\tau}_i = \begin{pmatrix} \hat{\mu} \\ \hat{\theta}_i \end{pmatrix} = \left(\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y} \tag{9}$$

and the estimated variance-covariance matrix of these estimates is

$$\text{Var} \begin{pmatrix} \hat{\mu} \\ \hat{\theta}_i \end{pmatrix} = \left(\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \right)^{-1}. \tag{10}$$

So $\hat{\theta}_i$ will be the second element of the vector in (9) and $\text{Var}(\hat{\theta}_i)$ will be the element in the second row and second column of (10). Notice that since the dimension of $\boldsymbol{\Sigma}$ and $\mathbf{X}$ are dependent on visits, v, both of the estimators are functions of v.

In this terminology, our hypothesis test (5) can be written

$$H_0 : \mathbf{K}^T \boldsymbol{\tau}_i = 0 \quad \text{vs.} \quad H_A : \mathbf{K}^T \boldsymbol{\tau}_i \neq 0 \tag{11}$$

where $\mathbf{K}^T = (0, 1)$ and hence $\mathbf{K}^T \boldsymbol{\tau}_i = \theta_i$. From linear model theory, (11) can be tested using the F-statistic (section 3.6, Searle (1971))

$$\mathcal{F} = \frac{\hat{\theta}_i \left[\text{Var}(\hat{\theta}_i) \right]^{-1} \hat{\theta}_i}{\left(\mathbf{Y}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \hat{\tau}_i^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} \right) / (N - \text{rank}(\mathbf{X}))}.$$

Furthermore, $\mathcal{F} \sim F(1, N - \text{rank}(\mathbf{X}), \delta)$ where, for a desired detectable difference of Δ,

$$\delta = \frac{\Delta^2}{2 \cdot \text{Var}(\hat{\theta}_i)}, \tag{12}$$

is the noncentrality parameter. Under the null hypothesis in (11), $\mathcal{F}$ has a central F distribution with 1 and $N -$ rank($\mathbf{X}$) degrees of freedom so that

$$\alpha = \text{P}(\mathcal{F} > f_\alpha \mid \theta_i = \Delta = 0).$$

Therefore, we can easily find f_α from the F distribution tables. In the situation when some hypothesis $H_A : \mathbf{K}^T \boldsymbol{\tau}_i = \Delta$, not H_0 in (11), is assumed true, the power can be calculated using the noncentral F distribution,

$$1 - \beta = \text{P}(\mathcal{F} > f_\alpha \mid \theta_i = \Delta). \tag{13}$$

Notice that $\mathrm{Var}(\hat{\theta}_i)$, and hence δ, is a function of v. Therefore, given N, Δ, and α, we can compute the number of visits, v, required to obtain a specified power $1-\beta$ for the individual tests (5) from equation (13). Unfortunately no closed form expression can be given for v in terms of $1-\beta$, f_α, Δ, and N due to the nature of the noncentral F distribution and $\boldsymbol{\Sigma}$. However, if we assume a certain structure for $\boldsymbol{\Sigma}$, we can solve (13) for v.

Suppose, for example, that all of the observations from one subject are equicorrelated. In other words, $\boldsymbol{\epsilon}$ is distributed $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{J}],$$

where $\mathbf{I}$ is a $v \times v$ identity matrix and $\mathbf{J}$ is a $v \times v$ matrix with all its elements equal to one. The inverse of this matrix is

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{(1-\rho)\sigma^2}\left[\mathbf{I} - \frac{\rho\mathbf{J}}{1+(v-1)\rho}\right].$$

This inverse exists only if $\rho > -1/(v-1)$ (section 3.2, Press (1982)).

For simplicity we will assume that a non-placebo challenge is administered every two visits so that for every one non-placebo challenge there are three additional placebo challenges. Therefore, our estimate of the number of visits needed to obtain some particular power will be some multiple of two. (In practice, under the assumption of an equicorrelated error structure, estimates of $\hat{\theta}_i$ and $\hat{\mu}$ will not depend on where in the sequence the challenges appear). Hence our design matrix is

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & \ldots & 1 & 0 & 1 & 0 \end{pmatrix}.$$

Under this setup it can be shown that

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_i) &= \left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}_{[2,2]} \\ &= \frac{4(1-\rho)\sigma^2}{v} \end{aligned}$$

and the noncentrality parameter, (12), is then

$$\delta = \frac{\Delta^2 v}{8(1-\rho)\sigma^2}. \tag{14}$$

If the observations are assumed independent, just substitute $\rho = 0$ in all of the above equations.

Now that we have specified how to find v, the number of visits per subject, given the individual powers, we can study the relationship between v and N for testing our original hypothesis (1) at a desired α–level and power. Equations (13) and (14) allow us to easily evaluate this relationship. For example, using the powers of the individual tests given in Table 1, Tables 2 and 3 display the number of visits, v, required assuming independent (with $\sigma = 1$) and equicorrelated (with $\rho = 0.75$ and $\sigma = 1$) observations within a subject respectively. We assume here that the detectable difference is two standard deviations, $\Delta = 2\sigma = 2$. Notice that in general far fewer visits are needed when the observations are equicorrelated than when they are independent. Additionally, increasing v seems to have a greater proportional impact on power than increasing N. For example, in Table 2, with seven subjects, to increase $1 - \beta_0$ from 0.90 to 0.95 we can include eight more subjects keeping the number of visits at 12 (a proportional increase in N of 214 %) or ask the seven subjects to visit the site 14 times (a proportional increase in v of 117 %).

Finally, one could argue that the hypothesis of interest is better represented by one-sided alternatives. In these cases, the desired tests are based on Wald statistics for the treatment parameter (section 3.1, Neter, Wasserman, and Kutner (1990)). Calculations analogous to those in this section can then be performed to extend the results to the one-sided situation.

4 Comparisons with Alternative Methods

A natural question to ask is how the union-intersection methodology compares with procedures typically utilized when analyzing data from a repeated challenge study. We shall consider four alternative techniques including three general linear models (LM) and one semi-parametric method for comparison with the union-intersection approach. In Section 4.1 we will introduce these four procedures. In Section 4.2, we provide an initial comparison between the union-intersection method and the LM approaches via a power analysis analogous to the development in Section 3. Section 4.3 compares all the techniques based on powers estimated via Monte Carlo simulations. The simulations not only provide

a means of comparing the nonparametric approach with the others, but also allows for a fair comparison of all four procedures under a controlled, common framework.

4.1 Alternative Methods

One standard approach for analyzing repeated measures type data is to fit a general linear model (LM) to the observations. The LM allows us to account for effects due to subjects, treatments, and random error in the data under the assumption that these factors are additive; similar to the two-factor analysis of variance. The main issue in specifying such a model is the assignment of fixed and random effects (see Searle, Casella, and McCulloch (1994) chapter 1 for a discussion of this dilemma). The subjects are essentially blocks each receiving both placebo and challenge treatments. Classically, blocks are presumed to be random effects. Furthermore, the particular subjects used in a given experiment are not actually of interest. The treatment effect may be modeled in a number of different ways. We shall consider three.

To motivate the first model, recall the treatment factor represents the effect of the control and challenge on the response. Since we are interested specifically in these two levels, a fixed treatment effect would be appropriate. Therefore, we may consider the mixed model

$$Y_{ijk} = \mu + \alpha_i + s_j + \epsilon_{ijk} \tag{15}$$

where

$$s_j \ iid \ N(0, \sigma_s^2)$$

$$\epsilon_{ijk} \ iid \ N(0, \sigma_\epsilon^2)$$

$$i = 1, 2; \ j = 1, \ldots, n; \ k = 1, \ldots, v/2.$$

Here Y_{ijk} is the observation from the kth visit during the ith treatment on the jth subject, s_j represents the random subject effect, α_i represents the fixed treatment effect, ϵ_{ijk} represents the random error, and μ is the constant overall mean. This model assumes no interaction between the subject and treatment factors. Furthermore, the constraint $\alpha_1 = 0$ so that

α_2 corresponds to the average of treatment differences previously represented by the θ_i's in Section 3.

One could certainly argue for the presence of an interaction between subject and treatment. Therefore the second model considered is the same as the first with interaction. That is, let

$$Y_{ijk} = \mu + \alpha_i + s_j + (\alpha s)_{ij} + \epsilon_{ijk} \tag{16}$$

where

$$s_j \ iid \ N(0, \sigma_s^2)$$

$$(\alpha s)_{ij} \ iid \ N(0, \sigma_{\alpha s}^2)$$

$$\epsilon_{ijk} \ iid \ N(0, \sigma_e^2).$$

The parameters here are the same as before with the addition of an interaction effect represented by $(\alpha s)_{ij}$. Since s_j is a random factor, $(\alpha s)_{ij}$ must also be assumed random.

A third model assumes the treatment factor is nested within the subjects. This supposition allows for a model with random treatment effects, one for each individual. The nested model is written

$$Y_{ijk} = \mu + \alpha_{ij} + s_j + \epsilon_{ijk} \tag{17}$$

where

$$s_j \ iid \ N(0, \sigma_s^2)$$

$$\alpha_{ij} \ iid \ N(0, \sigma_\alpha^2)$$

$$\epsilon_{ijk} \ iid \ N(0, \sigma_e^2)$$

The notation here is the same as that in the previous models except α_{ij} represents the effect of treatment i within subject j. This model most closely approximates the situation for which we have proposed the union-intersection test, in which the treatment difference varies across individuals. The nested model differs in that here, the treatment effects are drawn from a normal distribution. For the situation described in this paper, there is a proportion of individuals $(1 - \phi)$ for which the treatment difference equals zero. For a more detailed description of these three LMs see Searle, Casella, and McCulloch (1994) chapter 4.

The general linear models (15), (16), and (17), however, are susceptible to the drawbacks discussed in Section 1. In particular, though the three models may be appropriate for modeling the repeated measurements within a subject inherent in our design, they do not directly test the hypothesis of interest.

Alternatively, we may consider a method based on semi-parametric statistics. The fourth approach, the Kolmogorov procedure, tests the hypothesis that for all individuals, the test statistic comes from a particular distribution (chapter 6, Conover (1980)). In other words, we are testing

$$H_0: \ F = F_0 \quad \text{vs.} \quad H_A: \ F \neq F_0. \tag{18}$$

In the problem examined in this paper, we are interested in testing whether the t-statistics, one for each subject, arise from a central T-distribution corresponding to a treatment difference of zero. Thus, to test that the treatment difference is zero for all individuals, set F_0 in (18) equal to a central T- distribution. When the null is rejected we conclude that at least one of the test statistics comes from a distribution other than a central T-distribution. Another interpretation is that rejection of the null indicates that the statistics come from a distribution which is either stochastically greater or smaller than the central T, suggesting that the average treatment difference is either smaller or greater than zero.

The test statistic for the Kolmogorov procedure is given by

$$D_N = \sup \left| \hat{F}_N(t) - F_0(t) \right| \tag{19}$$

where

$$\hat{F}_N(t) = [\text{number of } t_i \leq t]/n \ . \tag{20}$$

$\hat{F}_N(t)$ is commonly referred to as the empirical cdf of the test statistics t_i. The null hypothesis is rejected when D_n is too large. Critical values for this statistic are equal for all distributions F_0 with support on the real line and are given for selected α-levels and sample sizes, N, in Table IX of Conover (1980).

This test seems particularly promising for cases where the errors of the observations do not all necessarily come from a normal distribution with mean zero. In addition, this test, along with the union-intersection test, allows for the possibility that variances and

correlations differ between subjects. The previous three alternatives do not explicitly allow for this possibility.

4.2 Power Comparisons

Let us first compare the union-intersection methodology with the LM approach to analyzing challenge study data. Analogous to the work in Section 3, the hypothesis test of interest in each of the models (15), (16), and (17) may be tested by some statistic, $\mathcal{F}$, defined by the appropriate ratio of mean squares. Recall from linear model theory (e.g., Searle, Casella, and McCulloch (1994) chapter 4) that $\mathcal{F}$ is proportional to an F-statistic. Sample size determination proceeds by fixing the number of subjects N, $\alpha-$level, and detectable difference Δ, and determining the number of visits v required per subject to attain a specified power $1-\beta$. We first calculate f_α utilizing the equation

$$\alpha = P(\mathcal{F} > f_\alpha | \alpha_2 = \Delta = 0).$$

Then we plug f_α into the equation

$$1-\beta = P(\mathcal{F} > f_\alpha | \alpha_2 = \Delta \neq 0).$$

As we will show shortly, $\mathcal{F}$ is a function of v. Hence, upon fixing Δ, these two equations can be solved simultaneously and will give us the number of visits necessary to obtain a specified power. The only difficulty in these computations, then, is determining the detectable difference Δ and the precise distribution of $\mathcal{F}$.

The detectable difference, Δ, we desire under the general linear models is different than that used for the union-intersection power calculations. Recall from Section 3 we assumed an individual susceptible to the phenomena of interest will show a response of two standard deviations or $2\sigma_e$ greater than the baseline when challenged. In the population of subjects, then, we expect a percentage, ϕ, to display a response to the challenge where ϕ is the population prevalence. Therefore, a single individual from our sample will exhibit a treatment effect of $2\sigma_e$ with probability ϕ and zero with probability $1-\phi$.

We need to relate this concept to the components of the LMs considered in Section 4.1. The fixed treatment effects, α_i, in the mixed models (15) and (16) measure the average

treatment difference across all N subjects. Therefore, we expect α_i to take on a value of $(2\sigma_e \cdot N\phi + 0 \cdot N\phi)/N = 2\sigma_e\phi$ from the above definition. The random "interaction" effects $(\alpha s)_{ij}$ and α_{ij} in the mixed model (16) and nested model (17) respectively are the variances of the treatment effect for a subject. Thus we expect these variances to take on the value $4\sigma_e^2 \cdot \phi(1-\phi)$. Therefore, when testing the fixed effects α_i, the detectable difference Δ is assigned $2\sigma_e\phi$ under the alternative. For the tests of the random effects and $(\alpha s)_{ij}$ and α_{ij}, $\Delta = 4\sigma_e^2 \cdot \phi(1-\phi)$. Notice if $\sigma_e = 1$ and $\phi = 0.5$, then $\Delta = 1$ in both cases. Also, note that due to the manner in which the LMs approximate the situation assumed in this paper, these calculations are at best approximations.

We now must determine how v and N affects the F-statistics for treatment effect in each of the three LMs. Let $\mathcal{F}$ be the generic test statistic in each testing scenario defined by the appropriate ratio of mean squares. In the mixed model with no interaction (15) we are interested in testing the null hypothesis of no treatment effect, $\alpha_i = 0$ for all i. From mixed model theory (see Searle, Casella, and McCulloch (1994) section 4.3 c, 4.5 a, and 4.7 b for details of the calculations to follow), $\mathcal{F}$ has a non-central F distribution with 1 and $2Nv - N - 1$ degrees of freedom and noncentrality parameter

$$\delta = N\Delta^2 v/8\sigma_e^2,$$

denoted F(1, $2Nv - N - 1$; δ). Under the null hypothesis, $\Delta = 0$ and hence $\mathcal{F}$ has a central F distribution.

Similarly, for testing $\alpha_i = 0$ in the mixed model with interaction (16), $\mathcal{F}$ has distribution F(1, $N - 1$; δ) where the noncentrality parameter is

$$\delta = \frac{N\Delta^2}{2(\sigma_{\alpha s}^2 + 2\sigma_\epsilon^2/v)}.$$

Using this model, we would also interested in testing the hypothesis that $\sigma_{\alpha s}^2 = 0$ for all i and j. Tests of the random parameters are simpler since $\mathcal{F}$ is proportional to a central F distribution over the whole parameter space. In particular, for testing $\sigma_{\alpha s}^2 = 0$ in the mixed model with interaction, $\mathcal{F}$ is proportional to the F($N - 1$, $(v-2) \cdot N$) distribution. The constant of proportionality is $(1 + v\sigma_{\alpha s}^2/2\sigma_e^2)$.

Likewise, for testing $\sigma_\alpha^2 = 0$ in the nested model (17), $\mathcal{F}$ is proportional to the F(N, $(v-2)\cdot N$) distribution, with constant of proportionality $(1 + v\sigma_\alpha^2/2\sigma_e^2)$.

Therefore, given $\alpha = 0.05$, $\phi = 0.5$, $\sigma_e = 1$ so that $\Delta = 1$ under the respective alternative hypotheses, we may calculate the combination of subjects and visits necessary to attain an overall power $1 - \beta$. Tables 4, 5, 6 display the sample size determinations for various overall powers.

The test of the fixed effect in the mixed model with interaction (16) is not included because it requires more than 100 visits per subject to attain a power of 0.50 for thirty subjects. In comparison to Table 3, note that all three LMs require more visits than the union-intersection method for a small number of subjects. For a large number of subjects, the LMs perform well with four visits per subject in contrast to the union-intersection approach which needs at least six to obtain an overall power greater than 0.80. Furthermore, the no interaction mixed model in Table 4 attains the best subject to visit ratios amongst the three LMs.

As described in Section 2, we assume some individuals may display a positive treatment effect whereas others show a zero treatment effect. The alternative linear models do not represent this situation. Therefore a comparison of sample sizes via the calculations above is misleading. A better comparison of the four models requires a common sequence of data sets over which sample size and power determinations may be made. The simulations in the next section provides the perfect framework for such a comparison.

4.3 Simulations

In order to further examine the power of the union-intersection approach compared with the four alternative procedures under the exact same testing conditions, we conducted a number of Monte Carlo simulations. We consider the powers of these five tests under a number of different sample sizes, visits, and correlation structures. We also allow the variances to differ from subject to subject. All of the simulations presented here are based on 50,000 replications.

These simulations were completed using GAUSS (Aptech Systems, 1992). In partic-

ular, random normal variables were generated by the fast acceptance-rejection algorithm proposed by Kinderman and Ramage (1976). The initial seed was chosen from a random number table. Each subsequent seed is a function of the previous seed:

$$newseed = (a \times oldseed) \bmod m,$$

where $a = 397204094$ and $m = 2^{31} - 1$.

First, suppose that eight visits are conducted for each subject, four with the placebo and four with the challenge. Furthermore, suppose that for each individual, the visits have a correlation of 0.75 with variance of one and each individual mean placebo response is drawn from a standard normal distribution. Tables 7-9 show the powers for each of the five tests with prevalences of 0.25, 0.50, and 0.75 respectively. Within each prevalence, powers are determined for treatment differences, δ, of 1, 2, and 3, and sample sizes, N, of 2, 4, 6, 8, and 10. In each cell of the table, the powers are presented for the following order of tests: union-intersection methodology, Kolmogorov procedure, LM with only fixed treatment effect, LM with nested treatment effect, and LM with both fixed treatment effect and interaction with subject. In the last category, the power for the test of no treatment effect is listed first and the test of no interaction second.

Notice that in all cases, the nested model performs at least as well as the other techniques, and often times much better. The union-intersection approach is comparable when the prevalence is low, .25, or when the treatment effect is moderate to large.

The tests using Kolmogorov procedure has a consistently low power when compared to all of the other models. The nonparametric aspect of this procedure seems to be a disadvantage. Being able to make assumptions about a set of data allows us to employ more powerful tests which are customized to those assumptions.

Further simulations demonstrate that as the correlation increases, the power of the union-intersection tests look more and more like the power of the tests under the nested model. As the correlation decreases however, the nested model is more powerful.

Now suppose that the variance of responses is different for each individual, and that the treatment effect varies from individual to individual as well. In the Monte Carlo simulations, we assume that the variance comes from an exponential distribution with mean 2.

Furthermore, the treatment effect is assumed to be proportional to the individual standard deviation such that the constant of proportionality is drawn from a uniform distribution. Mathematically speaking,

$$\sigma_i^2 \; iid \; Exp(2) \tag{21}$$

and

$$\delta_i = c_i \cdot \sigma_i \tag{22}$$

where

$$c_i \; iid \; U(0,3). \tag{23}$$

These simulations were done for prevalences of $\phi = 0.25$, 0.50, and 0.75 and sample sizes N as before. The resulting powers are given in Table 10.

Note that for smaller sample sizes, the union-intersection test actually does better than its major competitor, the tests under the nested model, when variances vary across individuals. As the sample size increases however, the tests under the nested model again perform at least as well as the union-intersection tests.

In summary, the simulations demonstrate that the tests of parameters in the nested general linear model are frequently the most powerful. The Kolmogorov tests perform poorly in general. The union- intersection tests perform similarly to the tests under the nested model, although sometimes a bit less powerfully, and is even more powerful when the sample size is small or the prevalence is small.

Despite the fact that the nested model performs quite well in these simulations, we wish to extend a few words caution for those who would like to conclude that the nested model should be used over of the union-intersection approach. First note that occasionally, anecdotal information motivates an analysis of individual subjects following a challenge study. In these cases, there is no formal way of studying individual subjects once the nested model has been fit to the data. Second, the interpretation of the statement "the null hypothesis was rejected" is quite different under these two methodologies. Perhaps increasing the sample size and then applying the union-intersection procedure may be preferable to making the wrong interpretations using the LM.

5 Application

We will consider the practical implications of the theory developed in the previous two sections. To accomplish this task we will analyze data from a small portion of an experiment conducted by Bayer, Snow, and Strupp (1994). In this part of the study, eighteen rats were measured for performance on a task after being challenged with either dopamine or a control substance (purified, deionized water). The treatment administrations consisted of six injections (three dopamine, three control) randomly assigned to six experimental sessions for each rat. The injections were at least 48 hours apart to allow for an appropriate washout period.

The animals were tested fifteen minutes after the injections and received approximately 100 test trials on the task in a given session. The task required the rat to poke her nose in one of three funnels. A correct response corresponded to a poke in the funnel with a light cue. The data, therefore, consist of the percent of correct responses for each rat on a session of the experiment. This response is presumed to be a measure of attentional function or distractibility. The question of interest, then, is: does dopamine significantly affect attention in rats.

Hence, we are confronted with a challenge study design where the hypothesis under consideration is whether there exists a rat affected by dopamine; i.e.,

$$H_0 : \phi = 0 \text{ vs. } H_A : \phi > 0, \tag{24}$$

where ϕ is the "prevalence" of the effect of dopamine on rats as in (1). By way of union-intersection testing methodology, we can test (24) by checking for the existence of an effect in each rat individually and combine the results as described in Section 2. In other words, analogous to (5), we need to test each of

$$H_0^i : \theta_i = 0 \text{ vs. } H_A^i : \theta_i \neq 0, \quad i = 1, \ldots, 18, \tag{25}$$

where θ_i is the treatment effect for the ith rat. Under the assumption of independent observations over a given rat, we can test (25) by way of an F-test from the analysis of variance (ANOVA). The treatment effect here is measured by comparing responses on

the dopamine trials to those on the control trials. Independence is not a valid assumption, however, since the set of six observations comes from a single animal. One might expect any particular rat to respond in a similar manner from session to session. Thus an assumption of equicorrelated observations may be more appropriate. Using matrix manipulations, the distribution of our test statistic under the null hypotheses in (25) can be shown to be exactly the same using the independence or equicorrelation assumption. Hence the ANOVA p-values are identical under either correlation structure.

Table 11 displays the p-values for testing (25) for each of the eighteen rats. According to (6), the α-levels for each of the individual tests (25) needs to be 0.0028 for an α-level in the overall test (24) of 0.05. Therefore, even though two p-values are less than 0.10, none of the rats are significantly effected by dopamine at the desired level of 0.0028.

The lack of significance may be due to an insufficient number of challenges or rats. A power analysis may help us reach such a conclusion. For example, let us assume a constant variance, σ^2, across rats and use the average mean square error from the data, $\hat{\sigma}^2 = 0.00292$, as an estimate of this variance. In the notation of Section 2, there are $N = 18$ rats studied during $v = 6$ sessions. Suppose the desired detectable difference, Δ, is $2 \cdot \hat{\sigma}$. Then, utilizing (13) and (14) we can calculate the power for testing (25). As a function of ρ, the correlation between observations on a given rat (assuming observations on each rat are equally correlated), Figure (3) indicates the probability of a Type II error (false acceptance) is less than 0.0025 for all ρ. Therefore, it seems unlikely the large number of nonsignificant results is due to Type II errors.

Note that the power for testing the hypothesis of zero prevalence is at least 0.9975 (Figure 3). However, the researchers may perhaps be satisfied with an overall power of 0.90. It is interesting to know how much we need to change N or v to obtain a power of, say, 0.90. For sake of illustration, suppose $\rho = 0$. Then in each of the $N = 18$ tests, by (13) and (14), we would need to test the rats on two sessions. On the other hand, if we have the facilities to challenge and test the rats on six sessions, we need eight rats to obtain the same power for the individual tests (25). Therefore, fewer challenges or rats seem necessary in this experiment to attain a desired power of 0.90 for testing (24).

6 Conclusion

In planning a repeated challenge study, the investigator needs to carefully consider what the true hypothesis of interest is. When the question of interest centers around whether a phenomenon exists, the union-intersection analysis directly tests the hypothesis of whether the prevalence equals zero. The Kolmogorov test is a nonparametric method which also correctly tests the hypothesis of interest. However, the union-intersection test almost always has significantly greater power than the Kolmogorov test. On the other hand, the simulation studies indicate that the test from a nested linear model is often more powerful than the nonparametric and union-intersection methods. The union-intersection test, though, may be preferred because of its interpretation.

The union-intersection approach presented in this paper is not limited to the repeated challenge study. The concepts introduced by this methodology may provide insights into experiments in many other fields. Also, the union-intersection technique may be used to combine results of other tests besides F-tests. Furthermore, the analysis is straightforward and easy to implement in a variety of studies from a broad range of disciplines.

References

Aptech Systems. (1992). GAUSS: The GAUSS System Version 3.0. Aptech Systems, Inc., Washington.

Bayer, L. E., Snow, K., and Strupp, B. J. (1994). Effect of SKF 81297 in Control and Lead Exposed Rats: Evidence that D1 Dopaminergic Activity Modulates Attention. *Society for Neuroscience Abstracts* 20, 153.

Behar, D., Rapoport, J. L., Adams, A. J., Berg, C. J., and Cornblath, M. (1984). Sugar Challenge Testing with Children Considered "Sugar Reactive." *Nutrition and Behavior* 1, 277-288.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Wadsworth, California.

Conover, W. J (1980). *Practical Nonparametric Statistics*, 2nd edition. Wiley, New York.

Feingold, B. F. (1975). Hyperkinesis and Learning Disabilities Linked to Artificial Food Flavors and Colors. *American Journal of Nursing* 75, 797-803.

Kinderman, A. J. and Ramage, J. G. (1976). Computer Generation of Normal Random Numbers. *Journal of the American Statistical Association* 71, 893-896.

Metcalfe, D. D. and Sampson, H. A. (1990). Workshop on Experimental Methodology for Clinical Studies of Adverse Reactions to Foods and Food Additives. *Journal of Allergy and Clinical Immunology* 86 (3), 421-442.

Milich, R., Wolraich, M., and Lindgren, S. (1986). Sugar and Hyperactivity: A Critical Review of Empirical Findings. *Clinical Psychology Review* 6, 493-513.

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Statistical Models*, 3rd edition. Irwin, Illinois.

Pearson, D. J. (1985). Food Allergy, Hypersensitivity and Intolerance. *Journal of the Royal College of Physicians of London* 19 (3), 154-162.

Pollock, I. and Warner, J. O. (1990). Effect of Artificial Food Colours on Childhood Behaviour. *Archives of Disease in Childhood* 65, 74-77.

Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd edition. Krieger Publishing Company, Florida.

Prinz, R. J. (1985). Diet-Behavior Research with Children: Methodological and Substantive Issues. *Advances in Learning and Behavioral Disabilities* 4, 181-199.

Rosén, L. A., Booth, S. R., Bender, M. E., McGrath, M. L., Sorrell, S., and Drabman, R. S. (1988). Effects of Sugar (Sucrose) on Children's Behavior. *Journal of Consulting and Clinical Psychology* 56 (4), 583-589.

Roshon, M. S. and Hagen, R. L. (1989). Sugar Consumption, Locomotion, Task Orientation, and Learning in Preschool Children. *Journal of Abnormal Child Psychology* 17 (3), 349-357.

Rowe, K. S. (1988). Synthetic Food Colourings and 'Hyperactivity': a Double-blind Crossover Study. *Australian Paediatric Journal* 24, 143-147.

Roy, S. N. (1957). *Some Aspects of Multivariate Analysis.* Wiley, NY.

Sampson, H. A. (1992). Food Hypersensitivity: Manifestations, Diagnosis, and Natural History. *Food Technology* 46, 141-144.

Schardt, D. (1994). Food Sensitivity Nothing to Sneeze at. *Nutrition Action Healthletter* 21, 12- 14.

Searle, S. R. (1971). *Linear Models.* Wiley, New York.

Searle, S. R., Casella, G. , and McCulloch, C. E. (1994). *Variance Components.* Wiley, New York.

Van-Dusseldorp, M. (1989). Diet and Hyperactivity: A Review. *Voeding (The Hague)* 50 (1), 2-8.

Warner, J. O. (1987). Artificial Food Additive Intolerance: Fact or Fiction? In *Food Intolerance*, 133-147. Bailliere Tindall.

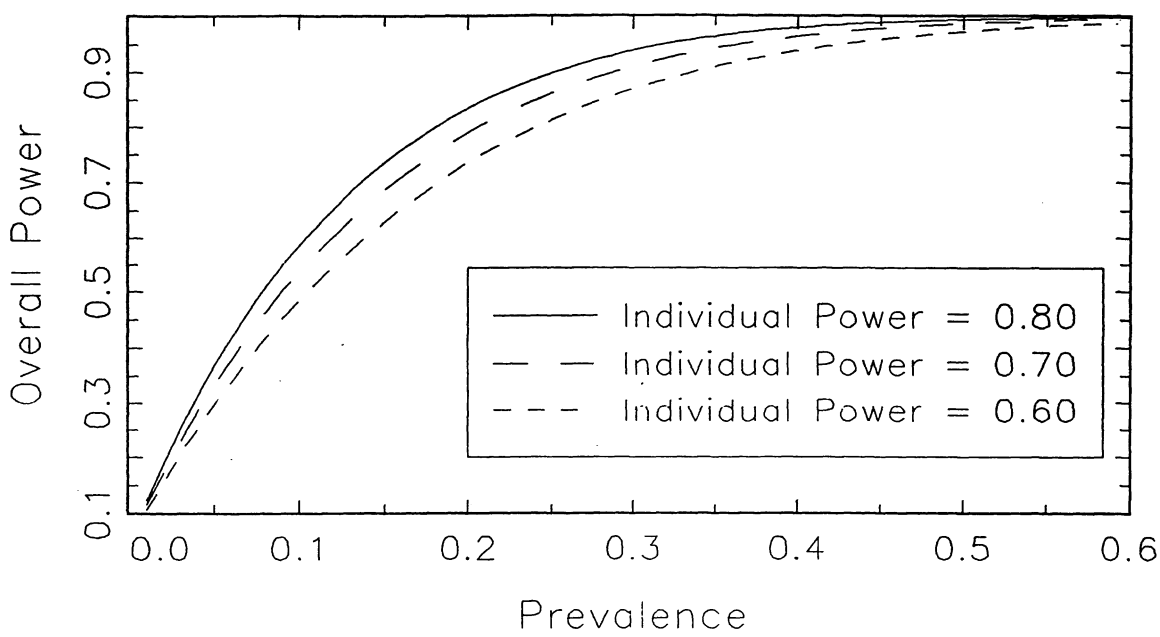

Figure 1: Overall power against prevalence of the phenomena of interest for three power levels of the individual tests. $N = 10$, $\alpha_0 = 0.05$.

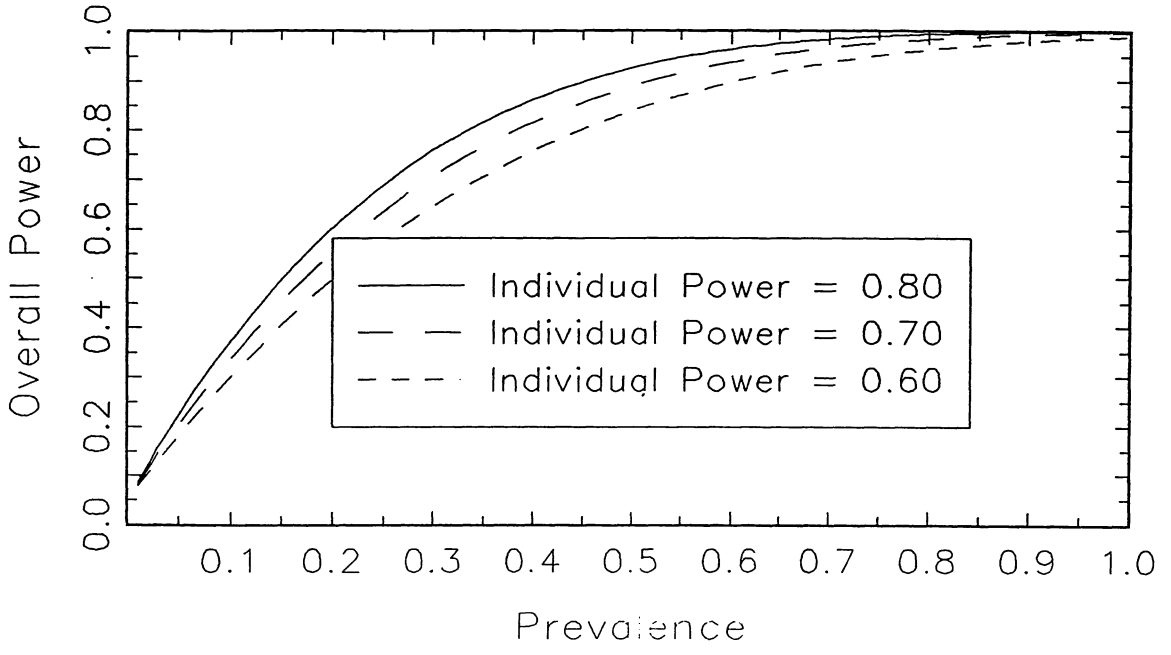

Figure 2: Overall power against prevalence of the phenomena of interest for three power levels of the individual tests. $N = 5$, $\alpha_0 = 0.05$.

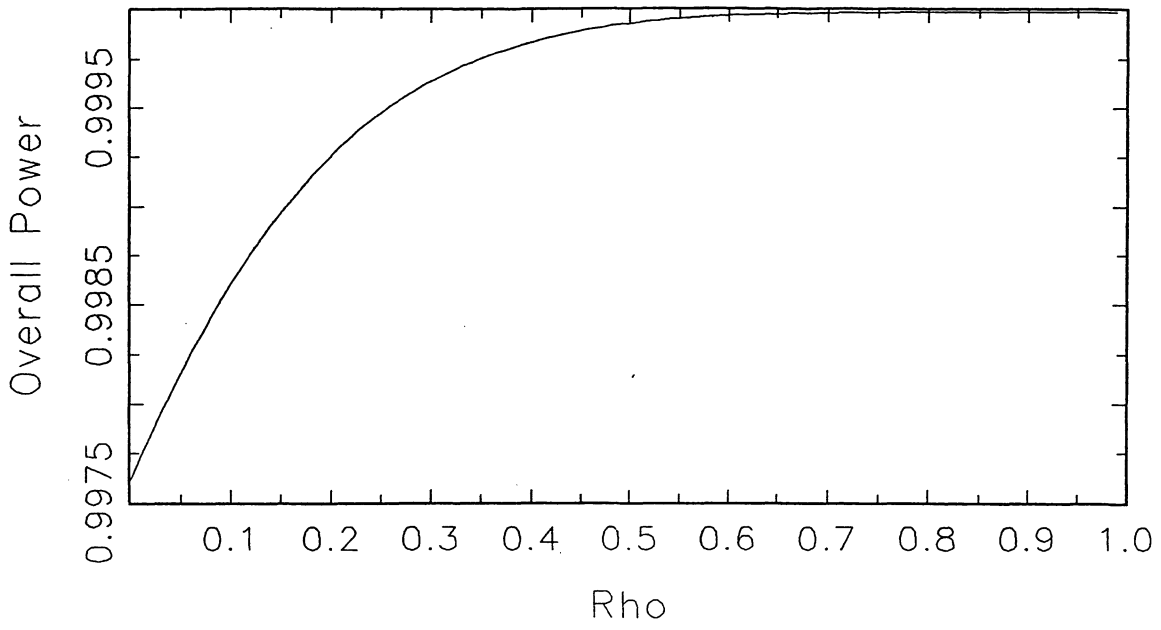

Figure 3: Overall powers against the correlation between observations, ρ, for the dopamine data. $N = 18$, $v = 6$, $\alpha = 0.05$, and $\Delta = 0.1081$.

TABLE 1

Powers needed in individual tests for specified overall power and N. $\phi = 0.50$, $\alpha_0 = 0.05$

$1-\beta$ / N	2	3	4	5	6	7	8	9	10	15	20	30
0.80	–	0.81	0.65	0.54	0.46	0.40	0.36	0.32	0.29	0.20	0.15	0.10
0.85	–	0.92	0.74	0.62	0.53	0.47	0.42	0.37	0.34	0.23	0.18	0.12
0.90	–	–	0.86	0.73	0.63	0.55	0.49	0.45	0.41	0.28	0.21	0.15
0.95	–	–	–	0.89	0.78	0.69	0.62	0.56	0.51	0.36	0.28	0.19
0.99	–	–	–	–	–	0.96	0.87	0.80	0.73	0.53	0.41	0.28

TABLE 2

Number of visits needed in individual tests for specified overall power and N.
Correlation structure of independence

Beta / N	2	3	4	5	6	7	8	9	10	15	20	30
0.80	–	14	12	12	12	10	10	10	10	10	10	8
0.85	–	18	14	12	12	12	12	10	10	10	10	10
0.90	–	–	18	14	14	12	12	12	12	10	10	10
0.95	–	–	–	18	16	14	14	14	14	12	12	10
0.99	–	–	–	–	–	24	20	18	16	14	14	12

TABLE 3

Number of visits needed in individual tests for specified overall power and N.
Correlation structure of equicorrelation with $\rho = 0.75$

Beta / N	2	3	4	5	6	7	8	9	10	15	20	30
0.80	–	8	6	6	6	6	6	6	6	6	6	6
0.85	–	8	8	6	6	6	6	6	6	6	6	6
0.90	–	–	8	8	8	6	6	6	6	6	6	6
0.95	–	–	–	8	8	8	8	8	8	6	6	6
0.99	–	–	–	–	–	10	8	8	8	8	8	8

TABLE 4

Number of visits needed in tests of the fixed effect in the mixed model with no interaction

Beta / N	2	3	4	5	6	7	8	9	10	15	20	30
0.80	–	12	10	8	6	6	6	4	4	4	4	4
0.85	–	14	10	8	8	6	6	6	4	4	4	4
0.90	–	16	12	10	8	8	6	6	6	4	4	4
0.95	–	18	14	12	10	8	8	6	6	4	4	4
0.99	–	24	18	16	12	12	10	8	8	6	4	4

TABLE 5

Number of visits needed in tests of the interaction in the mixed model with interaction

Beta / N	2	3	4	5	6	7	8	9	10	15	20	30
0.80	–	26	16	12	10	8	8	6	6	6	4	4
0.85	–	36	20	14	10	10	8	8	8	6	4	4
0.90	–	54	26	18	14	12	10	8	8	6	8	4
0.95	–	104	42	26	18	14	12	10	10	8	8	4
0.99	–	384	100	50	32	24	20	16	14	10	10	6

TABLE 6

Number of visits needed in tests of the treatment effect in the nested model

Beta / N	2	3	4	5	6	7	8	9	10	15	20	30
0.80	–	16	12	10	8	8	6	6	6	6	4	4
0.85	–	20	14	12	10	8	8	8	6	6	4	4
0.90	–	26	18	14	12	10	8	8	8	6	6	4
0.95	–	42	26	18	14	12	12	10	10	6	6	4
0.99	–	104	52	34	24	20	16	14	12	8	8	6

TABLE 7

Powers of test when treatment effect fixed given person has the sensitivity where **Prevalence=.25** and Total visits=8, correlation=.75

δ / N	2	4	6	8	10
	.270[a]	.338	.388	.417	.436
	.142[b]	.152	.178	.220	.250
1	.219[c]	.297	.375	.441	.503
	.311[d]	.455	.568	.651	.717
	.050[e] .204[f]	.055 .346	.094 .458	.141 .550	.189 .624
	.459	.678	.797	.866	.907
	.149	.166	.185	.262	.274
2	.376	.493	.604	.695	.760
	.463	.699	.826	.901	.941
	.058 .389	.037 .687	.071 .821	.136 .898	.213 .939
	.466	.697	.830	.905	.944
	.150	.165	.183	.259	.275
3	.446	.591	.693	.778	.838
	.466	.696	.831	.906	.945
	.068 .408	.033 .693	.069 .831	.125 .906	.212 .945

In this and all following tables in this section, the rows in each box correspond to tests using the a: union intersection methodology, b: Kolmogorov procedure, c: fixed effects model, d: nested model, e and f: mixed model testing the fixed effect and interaction respectively.

TABLE 8

Powers of test when treatment effect fixed given person has the sensitivity where **Prevalence=.50** and Total visits=8, correlation=.75

δ / N	2	4	6	8	10
	.457	.561	.616	.648	.675
	.315	.407	.508	.642	.721
1	.426	.618	.757	.844	.900
	.532	.748	.866	.928	.961
	.081 .252	.169 .447	.323 .593	.480 .706	.614 .791
	.756	.927	.973	.990	.995
	.340	.451	.537	.738	.779
2	.646	.818	.917	.964	.982
	.762	.940	.983	.996	.999
	.130 .506	.153 .869	.362 .964	.580 .990	.744 .998
	.765	.939	.985	.997	.999
	.341	.451	.535	.737	.778
3	.738	.877	.949	.980	.993
	.764	.940	.985	.997	.999
	.175 .527	.145 .880	.367 .970	.603 .993	.776 .998

TABLE 9

Powers of test when treatment effect fixed given person has the sensitivity where **Prevalence=.75** and Total visits=8, correlation=.75

δ / N	2	4	6	8	10
	.624	.721	.766	.797	.818
	.572	.747	.862	.946	.978
1	.677	.883	.960	.988	.996
	.731	.922	.978	.994	.998
	.148 .202	.445 .349	.699 .459	.862 .549	.943 .624
	.933	.993	.999	1.000	1.000
	.628	.804	.895	.983	.990
2	.852	.971	.995	.999	1.000
	.938	.996	1.000	1.000	1.000
	.267 .388	.457 .684	.798 .821	.944 .898	.986 .938
	.940	.996	1.000	1.000	1.000
	.622	.806	.897	.982	.990
3	.919	.985	.998	1.000	1.000
	.940	.997	1.000	1.000	1.000
	.369 .407	.450 .696	.829 .828	.957 .905	.991 .947

TABLE 10

Powers of test when treatment effect and variance for an individual are random where Total visits=8, correlation=.75

ϕ / N	2	4	6	8	10
	.336	.502	.616	.698	.757
	.127	.133	.152	.196	.211
.25	.269	.349	.429	.505	.563
	.333	.495	.610	.695	.757
	.035 .283	.025 .464	.036 .585	.059 .672	.086 .735
	.569	.773	.869	.922	.951
	.271	.350	.424	.570	.627
.50	.480	.645	.762	.843	.898
	.558	.761	.865	.924	.957
	.042 .436	.048 .697	.122 .822	.234 .895	.367 .937
	.753	.912	.965	.984	.993
	.480	.645	.764	.898	.939
.75	.685	.862	.943	.978	.991
	.744	.909	.969	.988	.995
	.060 .518	.126 .809	.340 .917	.581 .962	.768 .983

TABLE 11

F-test p-values of tests for treatment effect

Subject	1	2	3	4	5	6	7	8	9
p-value	0.011	0.661	0.735	0.842	0.187	0.775	0.922	0.084	0.873

Subject	10	11	12	13	14	15	16	17	18
p-values	0.999	0.953	0.249	0.529	0.487	0.326	0.869	0.996	0.456

For overall alpha level of 0.05, need to reject the individual tests at alpha levels of 0.0028