Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

T.A. Louis, J.K. Bailey

Controlling error rates using prior information
and marginal totals to select tumor sites

Department of Mathematical Statistics Report MS-R8704 April

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Controlling Error Rates Using Prior Information and Marginal Totals to Select Tumor Sites

Thomas A. Louis, Julia K. Bailey
Department of Biostatistics, Harvard School of Public Health

1980 Mathematics Subject Classification: 62F15, 62H17, 62K99, 62P10.
Key Words & Phrases: Bayes methods, bioassay, conditional power, multiplicity.
Note: This work was supported by grant ES02709 from the National Institute of Environmental Health Sciences, grant INT8512148 from the National Science Foundation, and a cooperative agreement between the Envirionmental Protection Agency and Harvard University through SIMS. The NSF grant supported Dr. T.A. Louis's visit to the Centre for Mathematics and Computer Science, Amsterdam, The Netherlands. We thank John Carlin and Nan Laird for their advice.

Report MS-R8704
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

ABSTRACT

Carlin & Louis (1985) propose a selection procedure designed to control the problems of multiplicity associated with P-values reported from carcinogen bioassays. Instead of searching the data for statistically significant tumor site/type combinations, the procedure uses site/type specific prior information and conditioning statistics to select sites and types with potentially significant P-values. Any single P-value selected by this method retains its usual meaning, and the size of the test procedure is controlled. We apply the Carlin & Louis procedure to a random sample of bioassays using male and female mice and rats from the National Cancer Institute's data base. From these data we estimate priors for lifetime incidence and dose effect. Then, we compare the performance of the selection procedure to use of Bonferroni adjusted and unadjusted minimum observed P-values.

Estimated priors for the occurrance of any malignancy show that over all strata prevalence in controls is about 25%, and that generally there is a positive association between dose and malignancy. Mice are far more sensitive than rats, with male rats the least sensitive. The association between dose and survival to terminal sacrifice shows a negative association with dose, so control rodents tend to live longer than others, suggesting that proper analysis should adjust for time until tumor. For all tumor sites, generally the prior standard deviation on the dose is effect is high, allowing the bioassay to contribute important information on determining carcinogenicity. The liver, mammary gland, spleen, and skin show at least a moderate positive dose effect, while site lymphosarcomas show a consistently negative association between dose and lifetime tumor incidence.

Our investigation indicates that the conditional selection approach has a rejection rate for the 5% level test competitive with unadjusted minimum P-values, and generally greater than that for adjusted P-values. Certain sites are picked quite fre-

quently. These sites have generally high a priori power, but the conditional power frequently does change the a priori indications. Plots of the empirical cumulative distribution functions for the P-values selected by the various rules show that the conditional selection approach produces a distribution that is between that for unadjusted and the Bonferroni adjusted minimum observed P-values. Theoretical development supports these observations.

The present analysis is primarily demonstrative. Our data-base does not contain information on time to tumor, dose, or cause of death. Conditional selection procedures using this information depend on straightforward technical generalizations of the Carlin and Louis procedure. But, specifying priors for target and nuisance parameters and implementing the selection rule pose serious challenges. Even with the current data each step of the analysis could be improved. Estimated priors could incorporate the full effect of estimating parameters (Morris 1983, Laird and Louis 1986), or the non-parametric maximum likelihood estimate of priors (Laird 1982) could replace those based on the Gaussian assumption. In practice, priors should be developed using a combination of information on chemicals with similar structure and expert opinion. Models could combine evidence on the priors over gender and strain.

Though the regulatory setting is too complicated to permit selection of a single site for analysis, use of our procedure to prioritize sites will focus debate. If statistical significance appears in a low-priority site, the burden of proof will be to justify the finding. If a high-priority site is significant, the burden switches to arguing why the result should not be believed. As important, much of the debate would occur before the experiment during the construction of priors.

Finally, we should remember that statistical significance computed from P-values provides

only one type of input into the risk assessment process. Yet, P-values do carry a great deal of weight, and increasing their validity will improve the risk assessment process. The previously noted extensions and refinements will be required before the selection procedure becomes "on-line". This development depends on the efforts and expertise of Toxicologists, Pathologists, Statisticians, Computer scientists and other experts. However, the effort and expense should be justified by the expense of the bioassay and its importance in the risk assessment process.

1. INTRODUCTION

In the typical carcinogen bioassay numerous statistical tests are performed simultaneously. Within each of the four common species-sex combinations (male-female, mice-rats), there may be 30 or more site and tumor-type combinations examined for tumors. A test statistic is calculated for each using a trend test for monotone dose response or pair-wise dose comparisons (see Lagakos and Louis 1985).

In the situation of a single species-sex combination with only one test statistic per site-type pair, the chance of obtaining at least one significant result under the null hypothesis of no treatment effect can be as high as $1-(1-\alpha)^K$, where K is the number of sites and α is the nominal level. This classical multiple comparisons problem can be circumvented by requiring the minimum individual P-value to be below α/k (the Bonferroni method), but determination of k and the loss of power makes this approach unattractive (Mantel 1980).

Recent research (Haseman 1983,5) suggests that in practice the multiplicity problem is not really serious, because the effects of discrete data and biological judgement produce an overall false positive rate that is quite close to the nominal level. The basic argument originates with Fears, Tarone, and Chu (1977). Dempster and Meng (1985) propose a more formal approach to the multiplicity problem based on a random effects model, and Finkelstein and Schoenfeld develop a single test statistic based on total tumors.

Carlin and Louis (1985) summarize these arguments and develop a method of selecting tumor sites for analysis using the conditional expected power (or the conditional expected P-value) of the test procedure for each site. Margins of two-way tables used in the analysis provide the conditioning and the procedure depends on prior distributions for site-specific tumor prevalence and dose effect. Sites with the highest conditional expected power (for a pre-selected significance level) or the smallest conditional expected P-value are selected for

the dose group comparison. Since site selection is based on prior information and table marginals (conditioning statistics for the usual frequentist analyses of these experiments), the selection process does not induce problems of multiplicity. In addition, standard reporting and interpretations are valid. The procedure has the potential to focus analysis and interpretation on a small number of sites and tumor types, but its statistical properties have not yet been investigated either theoretically or empirically.

In this report we apply the selection procedure to data from a random sample of carcinogen bioassays conducted by the National Cancer Institute (see Young 1986). We use these data to estimate site, gender, and species (mouse and rat) specific prior distributions for tumor prevalence and dose effect. Then, we apply the Carlin and Louis selection procedure and compare results to those using adjusted and unadjusted minimum observed P-values. We find that the power of the procedure compares favorably with that for unadjusted P-values, while controlling the overall size of the test procedure. Theoretical development supports this observation.

In section 2 we present the Carlin and Louis procedure; section 3 describes the data base, outlines the method of estimating of priors, and the method of comparing selection rules. Section 4 presents results, section 5 gives some theoretical background, and we close with a discussion in section 6.

2. THE CARLIN & LOUIS PROCEDURE

2.1 General setting

In this section we describe a general approach to picking one from a number of tests and then adapt the method to the carcinogen bioassay. Consider selecting a single test statistic from those available in such a way that the size of the selected test is unaffected by

multiplicity, and standard analysis and interpretations are valid. The procedure uses prior information about the likely magnitude of the non-centrality parameter for each test, and ancillary or partially ancillary information from the data, the latter based on a conditioning statistic. For conditional tests typically used in the bioassay (see Lagakos and Louis 1986) we use marginal tumor totals. P-values are conditional on such totals and so are free from multiple comparisons distortion, if only totals are used to select a single test statistic for the dose group comparison.

Suppose we are processing a number of results summarized by significance tests (P-values) for a scalar parameter θ. A single test uses data X to compare the null hypothesis $\theta = 0$ to the alternative $\theta > 0$. We consider frequentist P-values:

$$PVAL = pr[T(X) > T(x) \mid \theta = 0), \tag{2.1}$$

where T(·) is a test statistic. We judge the single test significant, if PVAL $\leq \alpha$, for some pre-determined cut-off.

Before seeing the P-values we want to choose the experiment that has the greatest statistical power. With a prior distribution (G) for θ we average over the parameter space Θ, producing the expected power for a single test:

$$EP(\alpha) = E[pr(PVAL \leq \alpha)] = \int_{\Theta} pr(PVAL \leq \alpha \mid \theta)\, dG(\theta). \tag{2.2}$$

In examining EP we are focusing on one feature of the distribution of the random variable "PVAL", namely the $100(1-\alpha)\%$ quantile. Other statistics may be of interest, especially ones that do not depend on α, such as the mean and variance of the distribution of "PVAL". For discrete data computing the complete distribution of "PVAL" is no more difficult than computing a single feature. Definition (2.2) can be extended to give the expected power conditional on a statistic by conditioning on the statistic in both "pr(·)" and G.

Using this conditional expected power, or any other feature of the conditional distribution of the P-value to select a test statistic induces no problems of multiplicity, even if the statistics are correlated.

2.2 Comparing Binomial Distributions

Although bioassays commonly are conducted with a control group and two or three dose groups, we have implemented the procedure only for pair-wise comparisons based on Fisher's exact test. Consider two groups of size n_0 (control) and n_1 (treated). Members of each group are examined for the presence or absence of tumors at several sites and of several types. For a single site and type, we denote by "T" the number of tumor-bearing animals in the treated group, and by "S" the total number of tumor-bearing animals (see Figure 1). Letting p_0 and p_1 denote the tumor rates, we test $H_0: p_1 = p_0$ versus $H_A: p_1 > p_0$.

Parameterizing in the logit scale, we write:

$$\mu = logit(p_0) = log[\frac{p_0}{1-p_0}],$$

and

$$\theta = logit(p_1) - logit(p_0).$$

The null and alternative hypotheses translate into $\theta = 0$ and $\theta > 0$, where θ should be considered the slope on the logistic dose-response curve with the treated animals receiving dose "1". With more than two groups, θ would multiply the actual dose or log dose.

The frequentist similar test results from conditioning on S, producing:

$$pr(T=t|S=s,\mu,\theta)=pr(T=t|S=s,\theta)=\frac{\binom{n_0}{s-t}\binom{n_1}{t}e^{\theta t}}{\sum_{j=0}^{s}\binom{n_0}{s-j}\binom{n_1}{j}e^{\theta j}}, \tag{2.3}$$

where a binomial coefficient is set to 0, if the bottom is bigger than the top. This non-central hypergeometric distribution is free of the nuisance parameter μ. At $\theta=0$ we have the standard Fisher's exact test.

For this testing problem the marginal distribution of S does depend on θ as well as on μ, so the distribution (2.3) fails to use all the information relevant to θ. This information loss is particularly acute when S is near 0 or n_0+n_1. A full Bayesian analysis would incorporate this information, but in this report we use the standard conditional analysis. Thus, in order to obtain the conditional distribution of PVAL given S for use in (2.2) we compute, (assuming that G has a density g):

$$pr_G[T=t|S=s]=\int\int pr(T=t|S=s,\theta)g(\mu,\theta|S=s)d\mu\,d\theta. \tag{2.4}$$

Substituting (2.3) into (2.4) and using Bayes' rule produces:

$$pr_G[T=t|S=s]\ \alpha\ \binom{n_0}{s-t}\binom{n_1}{t}\int\int\frac{e^{\mu s+\theta t}}{(1+e^{\mu})^{n_0}(1+e^{\mu+\theta})^{n_1}}g(\mu,\theta)d\mu\,d\theta. \tag{2.5}$$

From 2.5 we can produce the conditional expected power:

$$EPOW_G(S,\alpha)=pr_G(PVAL\leq\alpha|S), \tag{2.6}$$

and the conditional expectation ($EP_G[S]$) and variance ($VP_G[S]$) of the P-value.

Generally, calculation of (2.5) requires a numerical method, and Carlin and Louis discuss possible approaches. In the subsequent analyses we use independent Gaussian priors for μ and θ and a monte-carlo evaluation method that samples directly from the joint distribution of (μ,θ). For fixed (μ,θ) the joint distribution of (S,T) is the product of binomial probabilities. To compute the conditional distribution, $pr_G(T=t|S=s)$, we accumulate the values of the joint probabilities over a large sample of (μ,θ) pairs drawn from g.

Conditional

probabilities result from normalizing the vector of joint probabilities. We control numerical problems in calculating the binomial probabilities by computing one reference binomial distribution using Pascal's triangle recursion and calculating the others by multiplying this reference by the appropriate factor. With 10,000 replications the procedure runs rapidly on an IBM/AT and produces numerically stable results.

3. DATA AND METHODS

3.1 The data base

Young (1986) and companion papers by Bickis and Krewski (1986); Louis (1986); and Sanathanan et al (1986) give details on the data set and analyses. Twenty-five chemicals were selected at random from the NCI data base. For each chemical, data were obtained on four associated experiments (males and females, mice and rats). Thus, the data set contains information from 100 bioassays. One pair of the mouse experiments was set aside, leaving data on 98 experiments.

Data from each experiment consist of dose and site-specific lifetime tumors (malignant and non-malignant), a summary of dose-specific lifetime rates of having any tumor (per rodent, any tumor generates an "event"; denoted ALL in our tables), and dose-specific numbers of rodents surviving to terminal sacrifice. The data base contains no direct information on time to tumor and we coded doses "0", ".5", and "1". We base the current investigation on data for malignant tumors and use unadjusted lifetime rates. We estimated priors only for those tumor sites and types with at least one tumor in at least 10 experiments within a sex/strain stratum; with a separate determination in each stratum. The data base only lists sites where at least one tumor occurs, but for estimating the control rates, we use all

experiments, imputing zero tumors for the sites not reported.

3.2 Estimating priors

To implement the selection rule we need gender and strain-specific prior distributions for μ and θ. We estimate these priors using a simple variance components analysis that separately estimates the prior on the control rate and dose effect. More sophisticated approaches to this empirical Bayes analysis can be taken (Stiratelli et al 1984), but this approach will serve our present purpose.

Consider estimating a prior for μ from data $(X_k, n_k), k=1, \ldots, K$, where X_k is the number of tumor-bearing control rodents and n_k is the number of rodents for the k^{th} experiment. The X's come from a compound model where for each X_k a μ is drawn from the prior and then, conditional on μ, X follows a binomial distribution with success probability $p = e^{\mu}/(1+e^{\mu})$. First we estimate the prior distribution for the logistic normal random variable "P", and then transform it to a distribution for μ.

Estimates are based on the method of moments, by equating the observed weighted sample mean and variance to their expectation under the assumption that the prior mean and variance are related as in a beta distribution. Following Louis and DerSimonian (1982), estimate the expectation (m) and the variance (τ^2) of P by:

$$\hat{m} = \frac{X_+}{n_+},$$

$$\hat{\tau}^2 = \left[\frac{D - \hat{m}(1-\hat{m})}{n_+ \frac{(1-\sum p_k^2)}{K-1} - 1} \right]^+,$$

where

$$D = \frac{1}{K-1} \sum_{k=1}^{K} n_k (r_k - \hat{m})^2,$$

and $\rho_k = n_k/n_+$, and $r_k = X_k/n_k$.

Also, assuming the beta distribution relation, we can compute the precision (C) of the beta prior, defined by:

$$C = \frac{m(1-m)}{\tau^2}.$$

The value (C-1) plays the role of an effective sample size for this conjugate prior to the binomial distribution. We use C in transforming to a prior on μ.

Failing to account for the uncertainty in this estimated prior can cause problems. These problems can be seen most dramatically when $\hat{\tau}^2 = 0$. The prior induces a constant underlying tumor rate from experiment to experiment, since the estimated posterior distribution is degenerate at $\hat{m}$. Morris (1983), and Laird and Louis (1986) discuss this issue. Morris uses a delta-theorem expansion, and Laird and Louis use the bootstrap to introduce estimation uncertainty into the prior.

For a simple fixup, we introduce the uncertainty in estimating m by reducing the precision so that the prior variance includes a summand for the variance of $\hat{m}$ in a manner similar to that discussed by Morris (1983) for posterior distributions. Specifically, let

$$C^* = \frac{\hat{m}(1-\hat{m})}{\hat{\tau}^2},$$

and

$$C = C^* \frac{n_+}{n_+(1+\sum \rho_k^2) + C_* - 1}.$$

If $C < 1.01$, we set it equal to 1.01.

Notice that $C < C^*$ and if $C^* = \infty (\hat{\tau}^2 = 0)$, then $C = n_+$, introducing some variance into the prior. More sophisticated adjustments can be used, but this method will serve the present purpose.

Now, we convert the beta prior to a Gaussian prior on μ by:

$$E(\mu) = \log\left[\frac{C\hat{m} + .5}{C(1-\hat{m}) + .5}\right]$$

$$V(\mu) = \left[\frac{C^2}{C-1}\right]\frac{1}{(C\hat{m} + .5)[C(1-\hat{m}) + .5]},$$

in analogy to empirical logits.

We estimate the prior for θ using the method in Louis (1986). He uses the Z-score for a trend test comparing dose groups to impute an estimate of θ and a variance of this estimate for each experiment and then processes these estimates assuming a two-stage Gaussian sampling process. We use the control and two dose groups in this analysis, with the dose effect parameterized by:

$$\text{logit}[p_{\mu,\theta}(d)] = \mu + d\theta.$$

The trend test Z-score can be represented as Z = U/B, where "U" is a score statistic and B^2 is the estimated variance of the score statistic (computed from the table marginal) under the null hypothesis that $\theta = 0$. For logistic dose-response curves and small θ, $E_\theta(Z) \approx \theta B$, so we estimate θ_k for the k^{th} experiment by $\hat{\theta}_k = Z_k/B_k$. We estimate the sampling variance of this estimate by, $V_k = B_k^{-2}$, then process these $(\hat{\theta}_k, V_k)$ pairs with an iterative maximum likelihood variance components program that produces estimates of the prior mean and variance of θ. Then, we adjust the prior variance as described below. This approach depends on similar dose spacings in all experiments within a gender/strain stratum. These bioassays were conducted according to standard NCI protocols with most using the MTD, 50%MTD, and control, so a degree of similarity can be expected.

We adjust the prior variance (τ^2)for uncertainty in the prior estimates by first estimating the variance via maximum likelihood, producing τ^{*2} and then adding Q^{-1}, where:

$$Q = \sum_k (V_k + \tau^{*2})^{-1}.$$

This adjustment can be motivated by considering a Bayes model where the prior mean and variance have an noninformative hyper-prior. Then, using iterated expectations and variances, the mean of θ is approximately the estimated prior mean, and its variance is approximately τ^{*2} plus the variability of the estimated prior mean. The value Q^{-1} estimates this variance.

3.3 Comparing selection rules

For the high dose versus control test we compared the empirical rejection/acceptance rates of Carlin and Louis selection rules to the minimum observed P-value with and without adjustment. We applied the selection rules to data from the four strata, each with and without deleting "ALL" from contention, and for the two mouse strata also without allowing LI934 (liver hepatocellular carcinoma) to be selected. The minimum observed P-values were left unadjusted, inflated by the $\sqrt{K}$ (Mantel 1980), and by K (Bonferroni), where K is the number of site/type combinations in the experiment that have at least one tumor. For example, though we allow up to 34 sites to produce the minimum P-value, if only six sites have tumors, we multiply by six. Therefore, the adjustment for multiplicity is also conditional on one aspect of the margins. Even though the selection criteria relate to the high-dose versus control comparison, in addition we compared rules where observed P-value was computed from the trend test based on controls and two doses, but with site-selection rules based on the high- dose versus control 2×2 tables.

We summarize results by summaries of sites selected by the various rules, a series of 2×2 tables classifying the P-values above or below α, the cumulative distribution of P-values for selected sites, test statistics comparing computed probability of rejecting the null with the actual rejections, and statistics assessing the fit of the conditional expected P-values to those observed. We have applied both score and chi- square tests. The score test

computes:

$$\frac{\sum_{sites}(YES? - EP_G)}{[\sum_{sites} EP_G(1-EP_G)]^{\frac{1}{2}}},$$

and the chi-square is:

$$\sum_{sites} \frac{(YES? - EP_G)^2}{EP_G[1-EP_G]},$$

where YES? indicates if the P-value is below α. Similar computations apply to the observed and expected P-values.

4. RESULTS

4.1 Estimated Priors

Our use of tumors where at least four experiments provide information on θ (had at least one tumor) produced estimates for 34 sites. Tables 1a-d display estimated priors for sites where at least 10 experiments provided information on θ, and Table 2 gives the tumor types. The "ALL" rows show that over all strata prevalence in controls is about 25%, and that generally there is a positive association between dose and malignancy. Mice are far more sensitive than rats (look at the E(θ) columns), with male rats the least sensitive. The "SURV" rows indicate the association between dose and survival to terminal sacrifice. They show a negative association with dose, so control rodents tend to live longer than others, suggesting that proper analysis should adjust for time until tumor. For all tumor sites, generally the prior standard deviation on θ is high, allowing the bioassay to contribute important information on determining carcinogenicity. Sites ALL, LI934, MG902, SP932, SP939, SK972 show at least a moderate positive dose effect, while site MU939 shows a consistently negative association between dose and lifetime tumor incidence. The negative association between $\hat{m}$ and C suggests a departure from the beta distribution's

relation between mean and variance. The logistic dose-response relation probably doesn't hold at low doses.

4.2 Comparing Selection rules

Table 3 displays a selected set of comparisons between the conditional expected power and minimum P-value selection rules for nominal 5% level tests. They indicate that the conditional approach has a rejection rate competitive with unadjusted minimum P-values, and generally greater than that for adjusted P-values. Of course, we have no gold standard to determine a correct decision. Table 4 shows that certain sites (eg. LI934 for mice), are picked quite frequently. These sites have generally high a priori power, but the conditional power frequently does change the a priori indications. For example, when ALL is removed from contention, in mice LI934 is picked in 36 of the 48 experiments. When both ALL and LI 934 are removed, site selections are quite variable. Interestingly, though $E(\theta)$ for MU939 is negative and has a relatively small prior standard deviation, it is picked. Conditioning on the margins can overcome prior expectations. Results for the .025 level and use of minimum expected P-values give similar results.

In interpreting these results bear in mind that although 34 sites are potential contenders for selection, in a single experiment the number with at least one tumor is far smaller. Again, conditioning has dramatically changed the a priori situation. Table 4 displays the median and range for these numbers when ALL is a contender. This dramatic reduction in actual contenders supports Haseman's and Mantel's view that one should not adjust for multiplicity by multiplying the minimum observed P-value by the number of potential contenders. Of course, the smaller number of contenders does help our conditional approach do well, but this reduction is inherent in the bioassay.

Figure 2 plots the empirical cumulative distribution functions for the P-values selected by

the various rules using the Blom adjustment. As can be seen, the $E_{.05}$ approach produces a distribution that is between that for unadjusted (PMIN) and the number of sites adjusted (KPMIN) minimum observed P-values. In fact, it corresponds very closely to the curve for square root of sites adjusted minimum P-values (not plotted). At the 5% P-value, the $E_{.05}$ and the KMIN curves superimpose, consistent with the findings in Table 3. These stochastic relations suggest that for small P-values the conditional approach is picking up the same signals as those based on minimum observed P-values, while protecting the inference from multiplicity.

We computed score tests and chi-square type statistics for each site-type combination based on selected sites and on all sites. If our model is correct, the score tests should approximately follow a standard normal distribution (though correlations among sites make this assumption less tenable when more than one site from each experiment is used). Summary statistics indicate that our conditional expected power model under-predicts the power (the score statistics are positive), but the scores generally fall below the 2.5% Gaussian cut point. The chi-square statistics support the hypothesis of binomial variation. The expected P-value model produces negative scores (indicating that the observed P-values tend to be smaller than expected), a finding consistent with under-predicting the power.

5. THEORETICAL BACKGROUND

5.1 General setting

In this section we analyze a simplified model to investigate the potential power of selection strategies. We compare a selection strategy to the exact Bonferroni adjustment of the minimum P-value from independent tests. Of course, comparison with the unadjusted or $\sqrt{K}$ adjusted minimum P-value would produce results less favorable to a selection

strategy. Our simplified notation disguises the feature that the selection strategy can depend on coordinate-specific priors and conditioning information, but the analysis applies to these cases. The tests in our application are most likely dependent, so these results do not directly apply, but do provide support for the efficacy of the selection strategy.

Consider two strategies for interpreting the results of K independent hypothesis tests: One (R), selects a test according to a probability distribution $\mathbf{w} = (w_1, \ldots, w_K)$, and the other (M) adjusts the smallest observed P-value by the exact Bonferroni method. The probability vector $\mathbf{w}$ can depend on prior distributions and conditioning statistics, but not on the observed P-values. We want to compare the statistical power of the two procedures under various alternatives and values of K. Let $PVAL_{\mathbf{w}}$ be the observed P-value for the random selection strategy and $PVAL_K^*$ be the unadjusted observed P-value for the minimum P-value strategy. Then, the exact Bonferroni adjustment method converts $PVAL_K^*$ to:

$$PVAL_K = 1 - (1 - PVAL_K^*)^K .$$

Assume continuous distributions for all P-values and denote a cdf for a P-value by F_θ, with $\theta = 0$ the null (F_0 is uniform [0,1]), and $\theta > 0$ the alternative. Then, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is the true model, the distributions of $PVAL_{\mathbf{w}}$ and $PVAL_K$ are:

$$pr(PVAL_{\mathbf{w}} > \alpha \,|\, \boldsymbol{\theta}) = 1 - \sum_k w_k F_{\theta_k}(\alpha), \tag{5.1}$$

and, with $\beta = 1 - \alpha$,

$$pr(PVAL_K > \alpha \,|\, \boldsymbol{\theta}) = \prod_k \left[1 - F_{\theta_k}(1 - \beta^{\frac{1}{K}}) \right]. \tag{5.2}$$

Under $H_0: \boldsymbol{\theta} = \mathbf{0}$, $PVAL_{\mathbf{w}}$ and $PVAL_K$ have uniform distributions. Under alternatives we have the following:

Theorem 1: If $\boldsymbol{\theta} > 0$ (component-wise), then both test procedures are unbiased.

proof: In this case, since every component of θ is in the alternative space, $F_{\theta_k}(u) \geq u$ for all $u \in [0,1]$ and all k. Therefore, both (5.1) and (5.2) evaluate to $\leq 1-\alpha$, proving the result.

If some coordinates of θ are negative, then the tests may be biased. Bias can be prevented in procedure R by putting large weight on the components with positive $\theta's$. Procedure M can be protected only by eliminating coordinates with negative $\theta's$. These ideas underly Mantel's (1980) discussion of controlling the size of the overall test procedure.

At a given significance level, α, power comparisons between R and M depend on (5.1) and (5.2). If (5.1) is *smaller than (5.2),* then procedure R is more powerful.

5.2 Analysis as $K \to \infty$.

Consider first the case where $\theta_1 = \theta_2 = \cdots = \theta_K = \theta$.

Theorem 5.2: For all K and all w:

$$pr(PVAL_w > \alpha \mid \theta) = 1 - F_\theta(\alpha), \tag{5.3}$$

and

$$\lim_{K \to \infty} pr(PVAL_K > \alpha \mid \theta) = (1-\alpha)^{f_\theta}, \tag{5.4}$$

where $f_\theta = f_\theta(0)$; the density of the P-value distribution evaluated at 0.

proof: Equation (5.3) is simply (5.1) under the condition of the theorem. Under this condition, the right-hand side of (5.2) becomes:

$$\left[1 - F_\theta\left(1-\beta^{\frac{1}{K}}\right)\right]^K.$$

To simplify taking limits write $\beta^{\frac{1}{K}} = u$, so $K = \log(\beta)/\log(u)$, and the log of the right-hand side becomes:

$$log(\beta)\frac{log[1-F_\theta(1-u)]}{log(u)},$$

and we want the limit as $u \to 1$. l'Hôpital's rule produces the result.

Note that f_θ can be 0 or ∞ and depends on the characteristics of the tail of the distribution of the underlying random variable producing the P-values.

Theorem 5.2 can be generalized to the case where the $\theta's$ come from a distribution G. Then, we obtain:

$$pr(PVAL_{w_G} > \alpha \mid G) = 1 - \int w_G(\theta) F_\theta(\alpha)\, dG(\theta), \tag{5.3'}$$

with $\int w_G(\theta)\, dG(\theta) = 1$, and

$$\lim_{K\to\infty} pr(PVAL_K > \alpha \mid G) = (1-\alpha)^{\int f_\theta \, dG(\theta)}. \tag{5.4'}$$

We conjecture that (5.3) is always greater than (5.4). If so, then for K large and all components having equal power, the adjusted minimum P-value strategy will have greater power than any selection strategy. Formulae (5.3') and (5.4') show, however, that when the $\theta's$ vary sufficiently, intelligent selection rules (w_G) can be more powerful. Recall that procedure R can use conditioning information and coordinate-specific priors. Dependency of the weight function $w_G(\theta)$ on θ should be interpreted as allowing these dependencies.

Specializing the above to the case where the test statistics producing the P-values come from a common location family with location parameters θ, we find:

Theorem 5.3: If the location family has distribution H with density h, then:

$$f_\theta = \lim_{x\to\infty} \frac{h(x-\theta)}{h(x)}.$$

proof: A straightforward adaptation of the proof of Theorem 5.2.

When the tail of h damps exponentially such as for the logistic and bilateral exponential distributions, $f_\theta = exp(\theta)$. For $\theta > 0$ (5.3) is greater than (5.4) and procedure M is more

powerful than R. However, as mentioned previously, when the $\theta's$ vary, intelligent selection rules can make R better than M.

For the Gaussian distribution, f_θ equals ∞ 1 or 0 according as θ is greater than, equal to, or less than 0. This result implies that if G puts any mass on the positive real line, procedure M has asymptotic power equal to 1. If G has finite, positive real line support, then M is asymptotically better than R. But, if G has unbounded positive support, then R can compete by placing large weight on $\theta_{[K]}$, the largest θ in the sample. Analysis of this race to asymptosis requires comparison of the tails of H and G.

5.3 The case K=2

If the $\theta_1 = \theta_2 > 0$, then for the logistic and bilateral exponential examples, procedure M has higher power than R. If the $\theta's$ differ sufficiently, then placing large weight on the larger gives R higher power than M. In the Gaussian example relations depend on the size of the test, α.

6. DISCUSSION

The present analysis is primarily demonstrative. Our data-base does not contain information on time to tumor, dose, or cause of death. Conditional selection procedures using this information depend on straightforward technical generalizations of the Carlin and Louis procedure. But, specifying priors for target and nuisance parameters and implementing the selection rule pose serious challenges. Even with the current data each step of the analysis could be improved. Estimated priors could incorporate the full effect of estimating parameters (Morris 1983, Laird and Louis 1986), or the non-parametric maximum likelihood estimate of priors (Laird 1982) could replace those based on the Gaussian assumption. In practice, priors should be developed using a combination of information on chemi-

cals with similar structure and expert opinion. Models could combine evidence on the priors over gender and strain.

Results indicate that the use of empirical priors to compute summaries of the conditional distribution of P-values has an empirical rejection rate competitive with the common approach of selecting the smallest observed P-value, and controls the statistical size of the overall procedure. Though we have no gold standard to assess the statistical power, theoretical calculations show that such selection procedures have potentially high power. The results should be interpreted as an upper bound on the performance of the conditional selection procedures, for we use the same data to estimate the priors and evaluate the procedure. There is some distance between the data used for prior estimation and assessment of performance, since we use all doses to estimate priors, but apply the method to high dose versus control comparisons. We could employ the bootstrap or jackknife (Efron 1982) for a more valid assessment, and expect our qualitative conclusions would hold up.

Though the regulatory setting is too complicated to permit selection of a single site for analysis, use of our procedure to prioritize sites will focus debate. If statistical significance appears in a low-priority site, the burden of proof will be to justify the finding. If a high-priority site is significant, the burden switches to arguing why the result should not be believed. As important, much of the debate would occur before the experiment during the construction of priors. As section 5 shows, even if the priors are misspecified, the size of the procedure is controlled, though the power will be reduced. This protection results from using the Bayesian structure to select site, but a frequentist analysis that conditions on table marginals. This conditioning reduces information, and this loss can be considerable for high or low prevalence tumors. A full Bayesian analysis could be performed, but would require even greater consensus on priors and generate its own controversy in place of the multiplicity problem.

Finally, we should remember that statistical significance computed from P-values provides only one type of input into the risk assessment process. Yet, P-values do carry a great deal of weight, and increasing their validity will improve the risk assessment process. The previously noted extensions and refinements will be required before the selection procedure becomes "on-line". This development depends on the efforts and expertise of Toxicologists, Pathologists, Statisticians, Computer scientists and other experts. However, the effort and expense should be justified by the expense of the bioassay and its importance in the risk assessment process.

REFERENCES

Bickis & Krewski (1986). A blind reanalysis of a Random Subset of NCI Bioassay studies: An empirical evaluation of statistical decision rules. *Fund. and Appl. Toxicology* (to appear).

Carlin J, Louis TA (1985). Controlling error rates by using conditional expected power to select tumor sites. *Proc. Biopharmaceutical Section of the Am. Statist. Assoc.:* 11-18.

Dempster AP, Meng C (1985). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Proc. Long term animal carcinogenicity studies: A Statistical Perspective.* Am. Statistical Assoc, Washington DC.: 66-72.

Efron B (1982). *The Jackknife, Bootstrap, and other Resampling Plans.* SIAM, Philadelphia, PA.

Fears TR, Tarone RE, Chu KC (1977). False-positive and False-negative rates for carcinogenicity screens. *Cancer Research 37:* 1941-1945.

Finkelstein DM, Schoenfeld DA (1986). Analysis of multiple tumor data from a rodent carcinogenicity experiment. *Dana Farber Cancer Institute Research Report, #*4827.

Haseman JK (1983). A reexamination of false-positive rates for carcinogenesis studies. *Fund. Appl. Toxicol. 3:* 334-339.

Haseman JK (1985). Use of multiple control groups to estimate false positive rates in carcinogenicity studies. *Proc. Biopharmaceutical Section of the Am. Statist. Assoc.:* 1-10.

Lagakos SW, Louis TA (1985). The Statistical analysis of rodent tumorigenicity experiments. Chapter 7 in *Tozicological Risk Assessment Vol I: Biological and Statistical Criteria.* DB Clayson, D Krewski, IC Munro eds. CRC Press.

Laird NM (1982). Empirical Bayes estimates using the Nonparametric Maximum

Likelihood estimate for the Prior. *J. Statist. Comput. Simul. 15:* 211-220.

Laird NM, Louis TA (1986). Empirical Bayes confidence intervals based on bootstrap samples. (submitted to *J. Am. Statist. Assoc.)*

Louis TA (1986). Analysis of a random sample of two-year carcinogen bioassays from the NCI data base. *Fund. and Appl. Toxicology* (to appear).

Louis TA, DerSimonian R (1982). Health Statistics based on discrete population groups. In. *Regional variations in Hospital Use: Geographic and Temporal Patterns of care in the United States,* DL Rothberg, ed. Lexington Books, Lexington MA.

Mantel N (1980). Assessing laboratory evidence for neoplastic activity. *Biometrics 36:* 381-399.

Morris CN (1983). Parametric empirical Bayes Inference: Theory and Applications (with discussion). *J. Am. Statist. Assoc. 78:* 47-65.

Sanathanan LP, Lin CT, Carr RN (1986). A blind reanalysis of a random subset of NCI bioassay studies: Do rats predict mice? *Fund. and Appl. Toxicology* (to appear).

Stiratelli R, Laird NM, Ware JH (1984). Random effects models for serial observations with binary responses. *Biometrics 40:* 961-971.

Young S (1986). Introduction to the symposium: A blind reanalysis of a random subset of NCI bioassay studies. *Fund. and Appl. toxicology* (to appear).

TABLE 1A

ESTIMATED PRIORS FOR MALE MICE for sites with information on θ in at least 10 expts.

SITE	K	100 $\hat{m}$	C	E(θ)	s.d.(θ)
ALL	23	29	10	.88	1.76
SURV	23	68	3	-.34	1.84
LI 932	11	1	1112	.58	2.54
LI 934	24	19	13	.79	1.95
LI 939					
LN 939	11	3	81	-.04	.26
LU 904	13	3	21	.33	.18
MG 902					
MU 939	23	5	15	-.44	.09
PI 902					
SK 924	12	1	15	.06	.69
SK 972					
SP 932	10	1	755	.67	.36
SP 939					
TH 908					
TH 926					
UT 902					

TABLE 1B

ESTIMATED PRIORS FOR FEMALE MICE for sites with information on θ in at least 10 expts.

SITE	K	100 $\hat{m}$	C	E(θ)	s.d.(θ)
ALL	22	24	15	.76	1.88
SURV	23	80	6	-.58	.49
LI 932					
LI 934	22	4	84	1.61	2.41
LI 939	13	1	84	.06	.95
LN 939	13	2	65	.43	3.53
LU 904	12	2	59	-.11	.24
MG 902					
MU 939	24	12	38	-.44	.43
PI 902					
SK 924					
SK 972					
SP 932					
SP 939	10	1	38	.46	2.00
TH 908					
TH 926					
UT 902					

TABLE 1C

ESTIMATED PRIORS FOR MALE RATS for sites with information on θ in at least 10 expts.

SITE	K	$100\hat{m}$	C	E(θ)	s.d.(θ)
ALL	24	27	17	.00	2.35
SURV	24	62	5	-.26	.82
LI 932					
LI 934	11	1	233	1.26	1.84
LI 939					
LN 939					
LU 904	10	2	233	-.23	3.18
MG 902					
MU 939	24	13	16	-.50	1.02
PI 902					
SK 924	11	2	63	-.42	1.58
SK 972	10	1	75	1.00	3.39
SP 932					
SP 939					
TH 908	12	2	1007	-.09	.28
TH 926	16	2	1007	-.37	.65
UT 902					

TABLE 1D

ESTIMATED PRIORS FOR FEMALE RATS for sites with information on θ in at least 10 expts.

SITE	K	$100\hat{m}$	C	E(θ)	s.d.(θ)
LL	24	21	24	.41	1.22
SURV	24	67	9	-.36	2.60
LI 932					
LI 934					
LI 939					
LN 939					
LU 904					
MG 902	18	2	906	.89	.54
MU 939	19	10	77	-.58	1.88
PI 902	13	4	57	-1.05	.23
SK 924					
SK 972	10	1	103	-.45	4.26
SP 932					
SP 939					
TH 908	12	4	880	-.33	.52
TH 926					
UT 902	16	1	49	-.38	.34

Table 2: Tumor types associated with the Table 1 codes

SURV = survival until terminal sacrifice
ALL = any malignancy
LI932 = liver hemangiosarcoma
LI934 = liver hepatocellular carcinoma
LI939 = liver lymphosarcoma
LN939 = lymph node lymphosarcoma
LU904 = lung alveolar/bronchiolar carcinoma
MG902 = mammary gland adenocarcinoma
MU939 = multiple organs lymphosarcoma
PI902 = pituitary adenocarcinoma
SK924 = skin fibrosarcoma
SK972 = skin squamous cell carcinoma
SP932 = spleen hemangiosarcoma
SP939 = spleen lymphosarcoma
TH908 = thyroid c-cell carcinoma
TH926 = thyroid follicular cell carcinoma
UT902 = uterus adenocarcinoma

Table 3: Joint frequencies of statistically significant results for the EP.05 selection criterion. The first row of each table indicates significance at the .05 level for minimum (adjusted) observed P-values and the first column indicates significance for the EP.05 criterion. Since table totals for all but the combined tables equal 24 or 25, probabilities (times 100) can be obtained by multiplying by 4. The combined frequencies are close to the percents.

	Male Mouse	Female Mouse	Male Rat	Female Rat	Combined
with AN900	8 1	8 1	3 1	4 0	23 3
unadj	0 15	0 15	0 21	0 21	0 72
adj	8 0	7 0	3 0	3 0	21 0
	0 16	1 16	0 22	1 21	2 75
without AN900	7 1	8 0	2 0	3 2	20 1
unadj	0 16	0 16	0 23	0 22	0 77
adj	6 0	6 0	1 0	2 0	15 0
	1 17	2 16	1 23	1 22	5 78
without AN900 and LI934 unadj	2 0	2 0			
	0 22	0 22			
unadj	1 0	1 0			
	1 22	1 22			

Table 4: Median and range of the number of sites with a least one tumor (including ALL), and most frequently selected sites. A blank indicates a small (but not necessarily zero) count.

	Male Mouse	Female Mouse	Male Rat	Female Rat
Median	6	7	8	8
Range	3-9	3-13	6-13	3-12
Selected Sites				
with ALL				
ALL	19	17	19	19
LI934	5	6		
without ALL				
LI934	21	15		
MU939		4	8	9
MG902				6
AN902			5	
without ALL and LI934				
LI932	7	1		
MU939	6	9	not investigated	
LU904	5	2		

Figure 1: A typical 2x2 table comparing a dose group and control

	Control	Treated	Total
Tumor	s-t	t	s
No Tumor	n_0-s-t	n_1-t	n_+ - s
Total	n_0	n_1	n_+

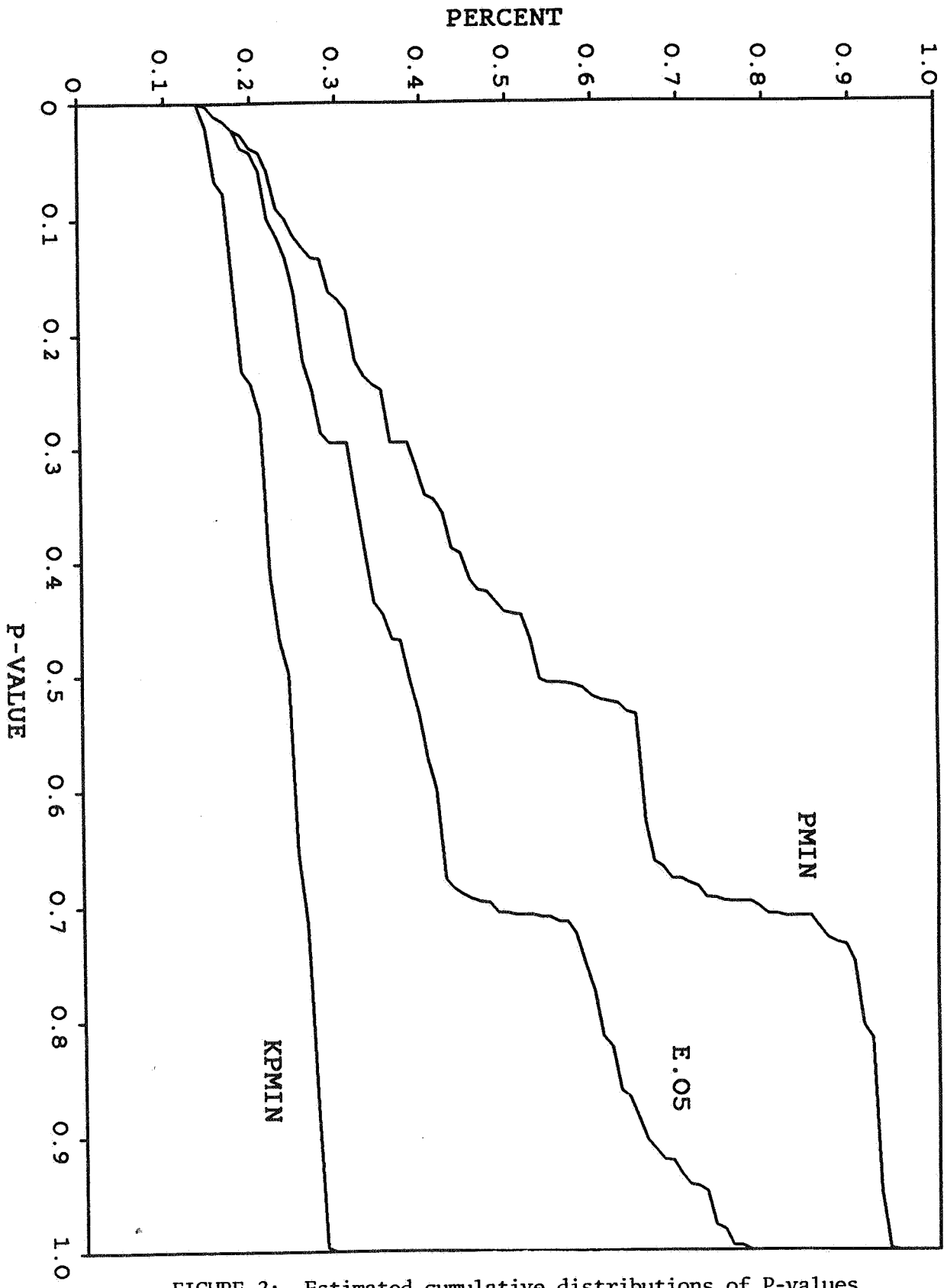

FIGURE 2: Estimated cumulative distributions of P-values for the minimum observed (PMIN), conditional expected power (E.05), and Bonferroni adjusted (KPMIN) decision rules.