

# ON SEQUENTIAL CLOSED TESTING PROCEDURES FOR A COMPARISON OF DOSE GROUPS WITH A CONTROL

Jessica Y. Chang and Hongshik Ahn

Department of Applied Mathematics and Statistics  
State University of New York at Stony Brook  
Stony Brook, NY 11794-3600

James J. Chen

Division of Biometry and Risk Assessment  
National Center for Toxicological Research  
Food and Drug Administration  
Jefferson, AR 72079

*Key Words:* dose-response; familywise error rate; Cochran-Armitage Test; Peto Cause-of-Death Test; Poly-3 Test; trend.

## ABSTRACT

Methods for a sequential test of a dose-response effect in pre-clinical studies are investigated. The objective of the test procedure is to compare several dose groups with a zero-dose control. The sequential testing is conducted within a closed family of one-sided tests. The procedures investigated are based on a monotonicity assumption. These closed procedures strongly control the familywise error rate while providing information about the shape of the dose-response relationship. Performances of sequential testing procedures are compared via a Monte Carlo simulation study. We illustrate the procedures by application to a real data set.

## 1. INTRODUCTION

Dose-Response experiments are often conducted to evaluate treatment effects of a test compound in medical and toxicological studies. In such experiments, subjects are randomly allocated to groups receiving different dose levels

of the compound. Included among the groups is a zero-dose control group, and a primary objective is to determine whether the mean response level for any of the non-zero treatment groups is significantly different from that of the control group. However, in addition to determining whether the test compound produces an effect (toxic or therapeutic) on the response of interest, it is more informative to investigate the shape of the dose-response curve. For example, in clinical trials, further analysis of the dose-response relationship may lead to determination of the lowest dose level at which there is a statistically significant compound-related effect. In pre-clinical studies, we seek to find the highest dose level at which there is no significant compound effect. Tukey et al. called this the NOSTASOT (NO-STATistical-Significance-Of-Trend) dose group. It is intuitively reasonable to assume a priori that the dose-response relationship is monotonically non-decreasing (or non-increasing).

In order to handle multiple comparisons among dose groups, it may be necessary to test multiple statistical hypotheses. To avoid inflation of the overall Type I error rate by overtesting, the use of Bonferroni's inequality is one standard approach. However, this method is known to result in a loss of power due to overadjustment. For this reason, it is generally considered to be overly conservative.

A multiple test procedure possesses a desirable property if, for each possible configuration of the dose-response curve, the probability of rejecting at least one of the true hypotheses is no more than a given threshold such as 5%. A test procedure with this property is said to control the Familywise Error Rate (FWE) in a strong sense (Lehmacher et al., 1991). A sequential testing procedure controlling the FWE can be constructed when the family of hypotheses to be tested is closed under intersections. Such a procedure allows each individual test to be performed by any valid method at the nominal significance level.

A general sequential testing procedure was described in Marcus et al. (1976) for a closed family of one-sided null hypotheses using the closure method of Peritz (1970). In this paper, we present and compare three testing procedures based on Peritz's closure principle and the closed family of hypotheses of Marcus et al. We adapt the procedures to the situation where a monotone dose-response relationship is assumed a priori. The first multiple testing scheme is formulated according to the suggestion of Tukey et al. (1985). We also present an alternative testing procedure which makes slightly different use of the available data. Third, we consider the sequential testing procedure proposed by Kodell and Chen (1991).

An animal bioassay for tumorigenicity (e.g., NTP, 1984) is commonly designed as a dose-response experiment. In this work we will focus our attention on such experiments, although the general results and recommendations will be broadly applicable. Most statistical analyses of dose-response bioassays for tumorigenicity employ trend tests to assess the strength of any dose-response relationship that might exist. Among those trend tests for detecting a linear trend across dose groups in the overall proportions of animals with the tumor, the availability of cause-of-death information is a factor in choosing which trend tests to apply. For data without cause-of-death information, Bailer and Portier (1988) proposed the Poly-3 Trend Test, which modifies the Cochran-Armitage Trend Test (Cochran 1954; Armitage, 1955) to handle situations when the mortality rates differ across dose groups. For data with cause-of-death information, the Peto's Cause-Of-Death Test (Peto, 1974; Peto et al. 1980) is commonly used. In this work, we have chosen the Cochran-Armitage Test, Poly-3 Test, and Peto Cause-Of-Death Test to carry out the trend test using the different procedures.

## 2. SEQUENTIAL TESTING PROCEDURES

Consider a dose-response study consisting of  $m$  dose groups and a zero-dose

control group. The treatment groups are indicated by the integers 0 to  $m$  in order of increasing dose level with 0 denoting the control group. Let  $\mu_i$  be the mean compound-related response of interest for dose group  $i$ . Assuming monotonicity of the dose-response relationship (i.e.,  $\mu_0 \leq \mu_1 \leq \dots \leq \mu_m$ ), our objective is to simultaneously compare  $\mu_i$  with  $\mu_0$  for  $i = 1, 2, \dots, m$ . Furthermore, if there is sufficient evidence to conclude that  $\mu_0 < \mu_m$ , we seek to find the highest dose level at which there is no significant compound effect. With this objective as well as the monotonicity assumption, it is natural to consider the family  $\mathcal{H} = \{H_1, H_2, \dots, H_m\}$  of (one-sided) hypotheses, where  $H_i$  is the hypothesis of homogeneity  $\mu_0 = \mu_1 = \dots = \mu_i$  which is also denoted by  $(0, 1, \dots, i)$  as in Kodell and Chen (1991).

## 2.1 Tukey's Procedure

One testing strategy suggested by Tukey et al. (1985) is described as follows. At a pre-determined significance level  $\alpha$ , we sequentially test each of the hypotheses  $(0, 1, \dots, m)$ ,  $(0, 1, \dots, m - 1)$ ,  $(0, 1, \dots, m - 2)$ ,  $\dots$ , so long as the hypothesis at the previous step is rejected. If  $(0, 1, \dots, i)$  is not rejected, the procedure terminates and  $\mu_0 = \mu_1 = \dots = \mu_i$  is concluded.

## 2.2 An Alternative Procedure

In conducting an experiment comparing several dose groups with a zero-dose control, one may think of comparing each of the dose groups alone with the control in a sequential approach. With the notation  $(0, i)$  representing the (one-sided) hypothesis of no difference between dose group  $i$  and the control (i.e.,  $\mu_0 = \mu_i$ ), we propose another sequential testing scheme here.

We test  $(0, m)$ ,  $(0, m - 1)$ ,  $(0, m - 2)$ ,  $\dots$ , sequentially, so long as the hypothesis at the previous stage is rejected. Every test is performed at a designated, fixed significance level  $\alpha$ . The procedure terminates when a failure of rejecting hypothesis  $(0, i)$  is reached and one can conclude that the level of

group  $i$  is the NOSTASOT dose.

This procedure is similar to Tukey's procedure with respect to the hypotheses tested and the order in which hypotheses are considered. Both procedures begin at the high end of the dose levels and step towards the low end. In fact, due to the monotonicity assumption on the dose-response relationship, both  $(0, 1, \dots, i)$  and  $(0, i)$  represent the same hypothesis. The main difference between these two procedures is that the former procedure tests  $(0, 1, \dots, i)$  using all data available from those groups while the latter one tests  $(0, i)$  using only the data from the control and dose group  $i$ .

### 2.3 Kodell-Chen Procedure

Kodell and Chen (1991) proposed a strategy for sequential testing in a dose-response study. For expository purposes, the description will be confined to experimental data consisting of three dose groups plus a control group, although it could be easily adapted to other situations. The proposed strategy proceeds as follows. At stage 1, a trend test is conducted to test  $(0, 1, 2, 3)$ . If  $(0, 1, 2, 3)$  is not rejected, then testing stops; otherwise, one proceeds to test the set of hypotheses in stage 2:  $(0, 1, 2)$ ,  $(0, 1) \cap (2, 3)$ ,  $(1, 2, 3)$ . If any of the three hypotheses in this stage is not rejected, then testing stops. If all three hypotheses are rejected, then one proceeds to test the hypotheses at stage 3:  $(0, 1)$ ,  $(1, 2)$ , and  $(2, 3)$ . Note that the notation  $(0, 1) \cap (2, 3)$  indicates the null hypothesis that groups 0 and 1 are not statistically different from one another, and that groups 2 and 3 are not different from each other. The first two groups may or may not be different from the last two (i.e., the null hypothesis:  $\mu_0 = \mu_1, \mu_2 = \mu_3$ ).

For the test of the hypothesis  $(0, 1) \cap (2, 3)$  in the Kodell-Chen procedure, there is more than one way to construct a valid test. The suggestion of Kodell and Chen (1991) is to calculate two separate statistics for each of  $(0, 1)$  and

(2, 3), and combine them in the manner of Peto et al. (1980) for tests of fatal and incidental tumors. In this study, we combine numerators and denominators of the separate Cochran-Armitage, Poly-3 and Peto  $z$ -statistics in order to have a pooled one-sided  $z$ -statistic.

All the possible outcomes of this procedure are enumerated and given symbolic representations in Tables 1 and 2 of Kodell and Chen (1991).

### 3. TWO ESSENTIAL FEATURES OF THE TESTING PROCEDURES

#### 3.1 Closed Test Property

A family of hypotheses is called closed under intersections if the intersection of any two hypotheses in the family is also in the family. Symbolically, let  $W = \{w_\beta, \forall \beta\}$  be a family of hypotheses. We say that  $W$  is closed under intersection if  $w_i, w_j \in W$  implies  $w_i \cap w_j \in W$  (Marcus et al., 1976).

Recalling the monotonicity of dose-response, both Tukey's procedure and the Alternative procedure have the same associated family of hypotheses  $\mathcal{H} = \{H_1, H_2, \dots, H_m\}$ , in which  $H_i$  is the hypothesis of homogeneity of the first  $i + 1$  dose groups. One can see that  $\mathcal{H}$  is closed under intersections since  $H_i \cap H_j = H_{\max\{i,j\}} \in \mathcal{H}$  for any integer  $i, j$  between 1 and  $m$ .

The set of hypotheses in the Kodell-Chen procedure forms a closed set since the null hypothesis in Stage 1 is the intersection of pairs of the null hypotheses in Stage 2, while the null hypotheses in Stage 2 are pairwise intersections of the ones in Stage 3.

#### 3.2 Controlling the Familywise Error Rate

A multiple test procedure is said to control the FWE in a weak sense if the probability of one or more false rejection of a hypothesis is bounded by  $\alpha$  under the global null hypothesis. A procedure controls the FWE in a strong

sense if the probability of rejecting at least one of the true hypotheses is no more than  $\alpha$ , no matter which of the hypotheses are actually true (Lehmacher et al., 1991). All three closed testing procedures presented control the FWE in a strong sense. A proof for the Kodell-Chen procedure is outlined in their paper (Kodell and Chen, 1991), and the proofs for the other two procedures are analogous.

#### 4. SIMULATION STUDY

For comparison purposes, a Monte-Carlo Simulation was conducted using the C++ programming language to evaluate the performance of various closed testing procedures. Among many methods that have been proposed for analyzing tumor incidence data from animal bioassays, Cochran-Armitage Test, Poly-3 Test, and Peto Cause-of-Death (COD) Trend Test were chosen to implement the individual trend tests comprising the different testing schemes.

##### 4.1 Model Design and Data Settings

A typical bioassay design with four dose groups of 50 animals each, and an experimental duration of 104 weeks, which is a normal term for a chronic study in rodents, was used in the study. The design was simulated to have a single terminal sacrifice at the end of the experiment as in the customary lifetime rodent bioassay. The four dose levels used are 0, 1, 2, and 4 for the no, low, intermediate and high doses, respectively.

For each animal, the observed outcome was assumed to be completely determined by the following three independent random variables (Portier et al., 1986):

1.  $T_1$  (Time to Tumor Onset):

The survival function for  $T_1$  was modeled as  $S(t) = \exp[-\theta\delta_1(t/104)^{\delta_2}]$ , where  $\theta = e^d \geq 1$ ,  $d$  is the dose level;  $\delta_1 \geq 0$  and  $\delta_2 \geq 0$ . The value

of  $\delta_2$  was set to 3. The parameter  $\theta$  was set equal to  $e^0 = 1$  and  $\delta_1 = -\ln[S(104)]$  was chosen such that the probability of tumor onset by 104 weeks was either .05, .20, .35, or .50.

2.  $T_3$  (Time to Death from Competing Risks):

The survival function for  $T_3$  was taken to be  $Q(t) = \exp[-\phi(\gamma_1 t + \gamma_2 t^{\gamma_3})]$ , where  $\phi \geq 1$ ,  $\gamma_1 \geq 0$ ,  $\gamma_2 \geq 0$ , and  $\gamma_3 \geq 0$ . With  $\phi = 1$ ,  $\gamma_1 = 10^{-4}$  and  $\gamma_2 = 10^{-16}$ ,  $\gamma_3$  was chosen to be 7.425531 such that the probability of survival with respect to competing risks at 104 weeks becomes .90. The value of  $\phi$  was taken as  $\ln p / \ln .9$  so that the survival rate becomes  $p$ .

3.  $T_2$  (Time after Onset until Death from the Tumor):

The survival function for  $T_2$  has the same form as that for  $T_3$ , and the values of  $\gamma_1$ ,  $\gamma_2$  and  $\gamma_3$  remained the same. An intermediate tumor lethality rate of .35 (i.e., approximately 35% of the observed tumors are cause of death) was chosen.

While fixing the competing risks survival rate ( $\text{crsr} = .5$ ) and the tumor lethality rate ( $\text{tlr} = .35$ ), 10 underlying models are considered with different tumor onset rates across dose groups. Using the notation in previous sections, let  $\mu_i$  be the actual tumor onset rate for group  $i$  ( $i = 0, 1, 2, 3$ ). TABLE I illustrates these 10 models with the corresponding tumor rates. For each of the 10 models, 10,000 simulated data sets were generated and tested using various testing schemes.

## 4.2 Results and Comparison

For comparison purposes, we have also included the traditional Bonferroni method and the Bonferroni-Holm Procedure (Holm, 1979).

For these two procedures, the set of (one-sided) hypotheses  $\mathcal{B} = \{(0, 1), (1, 2), (2, 3)\}$  are considered in our simulation settings. The Bonferroni method

Table I: Cumulative tumor onset probability at 104 weeks in the absence of competing risks.

Model		N*	Tumor Rates			
			0	1	2	3
(1)	$\mu_0 = \mu_1 = \mu_2 = \mu_3$	3	.05	.05	.05	.05
(2)	$\mu_0 = \mu_1 = \mu_2 = \mu_3$	3	.20	.20	.20	.20
(3)	$\mu_0 = \mu_1 = \mu_2 = \mu_3$	3	.35	.35	.35	.35
(4)	$\mu_0 < \mu_1 = \mu_2 = \mu_3$	0	.05	.20	.20	.20
(5)	$\mu_0 < \mu_1 = \mu_2 < \mu_3$	0	.05	.20	.20	.35
(6)	$\mu_0 < \mu_1 < \mu_2 = \mu_3$	0	.05	.20	.35	.35
(7)	$\mu_0 < \mu_1 < \mu_2 < \mu_3$	0	.05	.20	.35	.50
(8)	$\mu_0 = \mu_1 < \mu_2 = \mu_3$	1	.05	.05	.20	.20
(9)	$\mu_0 = \mu_1 < \mu_2 < \mu_3$	1	.05	.05	.20	.35
(10)	$\mu_0 = \mu_1 = \mu_2 < \mu_3$	2	.05	.05	.05	.20

\*NOSTASOT dose group

tests each hypothesis in  $\mathcal{B}$  at  $\alpha/3$  significance level. In the Bonferroni-Holm procedure, the three  $P$ -values  $P_1, P_2, P_3$  for  $(0, 1), (1, 2), (2, 3)$  are calculated and ordered so that  $P_{(1)} \leq P_{(2)} \leq P_{(3)}$ . The hypothesis belonging to  $P_{(k)}$  is denoted by  $B_k$  ( $k = 1, 2, 3$ ). The following test decisions are made:  $B_1$  is retained if and only if  $P_{(1)} > \alpha/3$ . The procedure stops and neither  $B_1$  nor any further single hypothesis can be rejected. If  $P_{(1)} \leq \alpha/3$ , then  $B_1$  is rejected;  $P_{(2)}$  is compared with  $\alpha/2$  and considered in an analogous manner. In general, assuming  $m$  is the number of hypotheses in the set  $\mathcal{B}$ , if  $P_{(k)} > \alpha/(m - k + 1)$ , then  $B_k$  cannot be rejected, the procedure stops, and no more single hypotheses can be rejected. Hypothesis  $B_k$  is rejected if and only if  $P_{(k)} \leq \alpha/(m - k + 1)$ ; then  $P_{(k+1)}$  is considered. In this way, the ordered  $P$ -values are compared stepwise with  $\alpha/m, \alpha/(m - 1), \dots, \alpha/2, \alpha$ . Both procedures strongly control the FWE (Holm 1979).

In Tables II - IV, we report simulation results and comparisons made among various testing schemes. We use the abbreviation T to stand for Tukey's procedure, A for the alternative procedure, K-C for the Kodell-Chen procedure (1991), while B represents the Bonferroni method and B-H for the Bonferroni-

Holm procedure.

Three criteria used for assessing the performance of the procedures are:

1. Global Power: the probability of rejecting the global null hypothesis ( $\mu_0 = \mu_1 = \mu_2 = \mu_3$ ).
2. All Comparison Power: the probability of rejecting ALL false null hypotheses (Bauer and Budde, 1994).
3. Class Identification Rate: the proportion (out of 10,000) that the NOS-TASOT dose group is correctly identified by the procedure.

The Global Power is the Type I Error Rate for the null model (i.e.,  $\mu_0 = \mu_1 = \mu_2 = \mu_3$ ). Moreover, since the Global Power and the All Comparison Power are equivalent in the null cases, we omitted the All Comparison Power for Models (1)-(3) in TABLE III. Note that the K-C procedure has the same Global Power as the T procedure, and the B and B-H procedures have the same Global Power. In the K-C procedure, there are 7 equivocal experimental outcomes which would yield no specific conclusion about the classification in our problem. Here, the reported Class Identification Rate for the K-C procedure is the proportion of correct class identifications among all instances where information about class identification is available.

All three trend tests give similar results for these 5 procedures. When the means are equal (null case), all procedures control the Type I error rate. In particular, B and B-H, which are known to be conservative, obtained the lowest Type I error rate. Notice in this case, the K-C procedure has the highest Class Identification Rate, followed by the B and B-H procedures.

In the cases that the NOSTASOT dose group is 0 (Models (4)-(7)), the A procedure tends to be the most powerful as well as obtaining the highest Class Identification Rate for the Cochran-Armitage and Poly-3 Tests. In this case, the T procedure has slightly lower Class Identification Rate than the A

Table II: Simulated Global Power for Cochran-Armitage, Poly-3 and Peto COD (Combined) Trend Tests. Tumor lethality:  $\simeq .35$ ; Competing risks survival rate:  $.5$ .

M <sup>a</sup>	N <sup>b</sup>	Cochran-Armitage					Poly-3					Peto COD				
		T	A	K-C	B	B-H	T	A	K-C	B	B-H	T	A	K-C	B	B-H
(1)	3	.054	.057	.054	.023	.023	.055	.052	.055	.024	.024	.053	.048	.053	.025	.025
(2)	3	.053	.051	.053	.051	.051	.050	.050	.050	.047	.047	.052	.053	.052	.048	.048
(3)	3	.051	.054	.051	.046	.046	.047	.048	.047	.040	.040	.052	.052	.052	.046	.046
(4)	0	.441	.688	.441	.494	.494	.449	.683	.449	.488	.488	.450	.640	.450	.442	.442
(5)	0	.945	.975	.945	.626	.626	.947	.977	.947	.621	.621	.944	.935	.944	.576	.576
(6)	0	.955	.978	.955	.691	.691	.960	.980	.960	.686	.686	.956	.941	.956	.631	.631
(7)	0	.999	$\approx 1$	.999	.780	.780	$\approx 1$	$\approx 1$	$\approx 1$	.778	.778	.998	.981	.998	.726	.726
(8)	1	.789	.692	.789	.500	.500	.796	.688	.796	.489	.489	.793	.642	.793	.448	.448
(9)	1	.994	.978	.994	.685	.685	.995	.980	.995	.681	.681	.994	.942	.994	.631	.631
(10)	2	.790	.687	.790	.485	.485	.799	.684	.799	.475	.475	.800	.636	.800	.432	.432

<sup>a</sup>Model                      <sup>b</sup>NOSTASOT dose group

Table III: Simulated All Comparison Power for Cochran-Armitage, Poly-3 and Peto COD (Combined) Trend Tests. Tumor lethality:  $\simeq .35$ ; Competing risks survival rate:  $.5$ .

M <sup>a</sup>	N <sup>b</sup>	Cochran-Armitage					Poly-3					Peto COD				
		T	A	K-C	B	B-H	T	A	K-C	B	B-H	T	A	K-C	B	B-H
(4)	0	.235	.455	.053	.470	.471	.232	.433	.053	.465	.467	.216	.367	.173	.419	.420
(5)	0	.453	.540	.126	.120	.170	.448	.527	.121	.119	.168	.413	.455	.098	.101	.143
(6)	0	.649	.675	.007	.056	.092	.651	.673	.007	.053	.093	.602	.595	.042	.047	.080
(7)	0	.683	.688	.037	.001	.020	.679	.683	.035	.001	.017	.630	.614	.024	.002	.016
(8)	1	.622	.552	.326	.479	.479	.623	.538	.326	.469	.469	.617	.479	.331	.428	.428
(9)	1	.737	.685	.164	.060	.099	.732	.681	.155	.056	.097	.725	.617	.094	.050	.084
(10)	2	.790	.687	.540	.475	.475	.799	.684	.543	.465	.465	.800	.636	.451	.421	.422

<sup>a</sup>Model                      <sup>b</sup>NOSTASOT dose group

Table IV: Simulated Class Identification Rate for Cochran-Armitage, Poly-3 and Peto COD (Combined) Trend Tests. Tumor lethality:  $\simeq .35$ ; Competing risks survival rate:  $.5$ .

M <sup>a</sup>	N <sup>b</sup>	Cochran-Armitage					Poly-3					Peto COD				
		T	A	K-C	B	B-H	T	A	K-C	B	B-H	T	A	K-C	B	B-H
(1)	3	.95	.94	.99	.98	.98	.95	.95	.99	.98	.98	.95	.95	.99	.98	.98
(2)	3	.95	.95	.99	.95	.95	.95	.95	.99	.95	.95	.95	.95	.99	.95	.95
(3)	3	.95	.95	.99	.95	.95	.95	.95	.99	.96	.96	.95	.95	.99	.95	.95
(4)	0	.24	.46	.08	.47	.47	.23	.43	.08	.47	.47	.22	.37	.23	.42	.42
(5)	0	.45	.54	.48	.47	.49	.45	.53	.48	.47	.49	.41	.46	.51	.42	.44
(6)	0	.65	.68	.30	.48	.50	.65	.67	.30	.47	.49	.60	.60	.65	.43	.45
(7)	0	.68	.69	.55	.48	.51	.68	.68	.55	.47	.51	.63	.61	.59	.43	.46
(8)	1	.58	.50	.57	.48	.48	.58	.49	.58	.47	.47	.58	.44	.58	.43	.43
(9)	1	.68	.63	.75	.48	.49	.68	.63	.75	.47	.49	.68	.58	.78	.43	.44
(10)	2	.74	.63	.69	.47	.46	.75	.63	.70	.46	.46	.75	.59	.65	.42	.41

<sup>a</sup>Model                      <sup>b</sup>NOSTASOT dose group

procedure in both Cochran-Armitage and Poly-3 Tests, but they are comparative in the Peto Cause-of-Death Combined Trend Test. The T procedure has substantially lower power than the A procedure if the tumor rate goes up at dose group 1 and dose not increase further. The K-C, B, and B-H procedures have low All Comparison Power in many circumstances.

When the NOSTASOT dose group is 1 or 2 (Models (8)-(10)), the T procedure is superior among all procedures. The K-C procedure performs comparatively well in terms of the Class Identification Rate and the Global Power, but it has low All Comparison Power. Although both Dose Identification Rate and the Global Power are slightly lower than the T and K-C procedures, the A procedure performs quite well and maintains reasonable All Comparison Power, while the B, B-H and K-C procedures have low All Comparison Power.

Under many circumstances, the T and A procedures are superior and comparative in terms of all three criteria. The K-C procedure obtained good Class Identification Rate and the Global Power in some cases, but as the B and B-H procedures, it does not possess good All Comparison Power. However, notice that the All Comparison Power is a much heavier requirement for the K-C procedure than for other procedures since the K-C procedure has the largest set of hypotheses. The B and B-H Procedures have quite consistent Class Identification Rates in all cases. As an improved Bonferroni method, the B-H procedure has slightly higher All Comparison Power than the B procedure.

## 5. EXAMPLE

Stallard and Whitehead (1999) presented the results of an experiment with male mice in which the effect of tumors of any type was investigated. The experiment included a control and three dose groups of the test substance. The dose levels were 0, 25, 125 and 500. Each of the control, low and intermediate dose groups contained 60 animals and the highest dose group contained 59

Table V: Tumor data for experimental animals with male mice from Stallard and Whitehead (1999).

Dose	Deaths without tumors (frequency in parentheses)	Deaths with tumors (frequency in parentheses)
Control	15, 62, 90, 92, 96, 97, 101, 105(22)	56, 65, 66, 76, 77, 80, 81, 86*, 87, 89, 93, 95, 97, 98(2), 103, 104, 105*(14)
1	24, 27, 53, 64, 68, 74, 82, 83, 94, 96, 97, 99, 100(2), 103, 104, 105(27)	63, 75, 78, 84, 85, 95, 96, 97, 98, 101, 102, 105*(6)
2	5, 7, 39, 65, 70, 75, 76, 80, 82, 83, 87, 91(2), 92, 96(2), 97, 98(2), 99, 100(2), 102, 105(23)	47, 52, 65, 69, 70, 88, 91, 95, 99, 100, 104, 105*(3)
3	16, 18, 49, 55, 59, 77, 85(2), 105	57*, 60, 66, 70(2), 74(2), 76, 78 83(2), 84(3), 85, 88, 89, 92, 93(2), 94, 95*, 95, 96, 97, 98*, 98(2), 99, 100, 101, 102*, 102, 103, 104, 105*(15)
	all deaths at week 105 are terminal sacrifice	all tumors except those marked (* ) were judged to be fatal

animals. The study duration was 105 weeks. A terminal sacrifice was performed during the last week of the study. The data introduced in Stallard and Whitehead are given in TABLE V.

The Peto Cause-of-Death Combined Trend Test, Poly-3 Test, and Cochran-Armitage Test were conducted for the five sequential testing procedures examined in Section 4. The test results are shown in TABLE VI. The three trend tests resulted in the same conclusion for the five procedures. For all the tests, only the global test in Tukey's procedure and the first test in the Alternative procedure were rejected at  $\alpha = .05$ . Thus, from the above two procedures, the second highest dose in this experiment is the NOSTASOT dose. The identified model from these procedures is Model (10) in TABLE I ( $\mu_0 = \mu_1 = \mu_2 < \mu_3$ ). In the K-C procedure, the trend test in the first stage is rejected, and in the second stage, the former test is accepted and the latter two tests are rejected. Thus, we obtain the same model as Model (10). In Bonferroni procedure, only the hypothesis (2, 3) is rejected at  $\alpha/3 = .0167$ . In B-H procedure, (2,3) is

Table VI: Results of the sequential testing procedures for the data in TABLE V.

Method	$H_i$	Peto, trend		Poly-3		C-A	
		Z value	p-value	Z value	p-value	Z value	p-value
T	(0123)	4.28	$\approx 0$	3.75	.00009	3.21	.00066
	(012)	-2.42	.99	-2.81	.998	-3.26	.999
	(01)	-2.14	.98	-2.33	.99	-2.60	.995
A	(03)	4.82	$\approx 0$	4.75	$\approx 0$	3.85	.00006
	(02)	-2.31	.99	-2.73	.997	-3.19	.999
	(01)	-2.14	.98	-2.33	0.99	-2.60	.995
K-C	(0123)	4.28	$\approx 0$	3.75	.00009	3.21	.00066
	(012)	-2.42	.992	-2.81	.998	-3.26	.999
	(01) $\cap$ (23)	3.24	.0006	3.05	.0011	2.94	.0016
	(123)	6.56	$\approx 0$	6.39	$\approx 0$	6.14	$\approx 0$
	(01)	-2.14	.98	-2.33	.99	-2.60	.995
	(12)	-0.18	.57	-.45	.67	-.62	.73
	(23)	6.25	$\approx 0$	6.90	$\approx 0$	6.69	$\approx 0$

rejected at  $\alpha/3$ , but the next test (1, 2) is not rejected at  $\alpha/2 = .025$ . Based on the five procedures, we conclude that group 2 is the NOSTASOT dose group.

## 6. DISCUSSION

A sequential closed testing procedure is a reasonable approach in a multiple test problem because it does not lose power due to an overadjustment as some traditional approaches do. Using the closure principle in a sequential approach allows researchers to test each single hypothesis at a designated  $\alpha$  level, instead of testing at a much smaller significance level. Moreover, the closed sequential test procedures presented in this paper guarantee that the FWE for statistical decisions is controlled strongly.

In addition to determining the NOSTASOT dose group, which is the main objective of this study, the procedures considered actually provide information about the shape of the dose-response. For example, the K-C procedure implicitly involves pairwise inferences among the dose groups. As long as a

significant trend is detected at stage 1, some pairwise comparisons among dose groups are made. Furthermore, it is a suitable closed sequential approach if one needs to draw conclusions about where each dose group stands in relation to all other dose groups with respect to mean response level. It could also be applied to situations in which no monotonicity of dose-response is assumed.

Simulating over a broad range of dose-response situations according to the monotonicity assumption, our results indicate that Tukey's procedure and the Alternative procedure are overall more favorable than others. However, there are circumstances where each procedure appears superior.

### ACKNOWLEDGEMENTS

Jessica Chang and Hongshik Ahn's work was supported by NIH grant 1 R29 CA77289-02. The authors would like to thank Dr. James Mancuso for thoughtful discussions and helpful comments.

### BIBLIOGRAPHY

- Armitage, P. (1955). "Tests for linear trends in proportions and frequencies," *Biometrics*, **11**, 417-431. Bailer, A. J. and Portier, C. J.] (1988). "Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples," *Biometrics*, **44**, 417-431.
- Cochran, W. G. (1954). "Some methods for strengthening the common  $\chi^2$  tests," *Biometrics*, **10**, 417-451.
- Holm, S. (1979). "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, **6**, 65-70.
- Kodell, R. L. and Chen, J. J. (1991). "Characterization of dose-response relationships inferred by statistically significant trend test," *Biometrics*, **47**, 139-146.
- Lehmacher, W., Wassmer, G. and Reitmeir, P. (1991). "Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate," *Biometrics*. **47**, 522-521.

- Marcus, R., Peritz, E. and Gabriel, K. R. (1976). "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, **63**, 655-660.
- NTP Board of Scientific Counselors (1984). "*Report of the NTP Ad Hoc Panel on Chemical Carcinogenesis Testing and Evaluation*," Research Triangular Park, North Carolina: National Toxicology Program, NIEHS.
- Peritz, E. (1970). "A note on multiple comparison," Unpublished manuscript, Hebrew University, Israel.
- Peto, R. (1974). "Guidelines on the the analysis of tumour rates and death rates in experimental animals," *British Journal of Cancer*, **29**, 101-105.
- Peto, R., Pike, M. C., Day, N. E., Gray, R. G., Lee, P. N., Parish, S., Peto, J., Richards, S. and Wahrendorf, J. (1980). "Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments," Annex to: *Long-term and Short-term Screening Assays for Carcinogens: a Critical Appraisal*. IARC monographs, Supplement 2. pp. 311-426. International Agency for Research on Cancer: Lyon, France.
- Portier, C., Hedges, J. and Hoel, D. G. (1986). "Age-specific models of mortality and tumor onset for historical control animals in the National Toxicology Program's carcinogenicity experiments," *Cancer Research*, **46**, 4372-4378.
- Stallard, N. and Whitehead, A. (1999). "Modified Weibull multi-state models for the analysis of animal carcinogenicity data," *Environmental and Ecological Statistics*, **6**. To appear.
- Tukey, J. W., Ciminera, J. L. and Heyse, J. F. (1985). "Testing the statistical certainty of a response to increasing doses of a drug," *Biometrics*, **41**, 295-301.

TABLE

TABLE TABLE TABLE TABLE TABLE