

An Age-Adjusted Bootstrap-Based Poly- k Test

Hojin Moon¹, Hongshik Ahn² and Ralph L. Kodell¹

¹Division of Biometry and Risk Assessment
National Center for Toxicological Research
Food and Drug Administration
3900 NCTR Drive, Jefferson, AR 72079
email: hmoon@nctr.fda.gov
phone: 870-543-7931 fax: 870-543-7662

²Department of Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY 11794-3600

Short title: An Age-Adjusted Bootstrap-Based Poly- k Test

Corresponding author: Hojin Moon

Hongshik Ahn's work was supported by NIH grant 1 R29 CA77289-06 and was partially supported by the Faculty Research Participation Program at the National Center for Toxicological Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between USDOE and USFDA.

SUMMARY

The assumption of an asymptotic normal distribution of some test statistics may be invalid in certain dose-response trend tests. For instance, the survival-adjusted Cochran-Armitage test, known as the Poly- k test, is asymptotically standard normal under the null hypothesis. However, the asymptotic normality is not valid if there is a deviation from the tumor onset distribution that is assumed in this test or if the competing risks survival rates differ across groups. We develop an age-adjusted bootstrap-based method to assess the significance of assumed asymptotic normal tests for animal carcinogenicity data. The proposed method differs from conventional bootstrap methods in the aspect of preserving the mortality rate in each dose group under the null hypothesis of equal tumor incidence rates among the groups. We investigate an empirical distribution of the Poly-3 trend test statistic using the proposed age-adjusted bootstrap-based method and compare it with the Poly-3 test statistic referenced to the assumed standard normal distribution. A simulation study is conducted to evaluate the robustness of these tests to various Weibull-family tumor onset distributions. The proposed method is applied to National Toxicology Program data sets to evaluate a dose-related trend of a test substance on the incidence of neoplasms.

Key Words: Bioassay; Competing risk; Single sacrifice; Survival-adjusted test; Trend test.

1 INTRODUCTION

It is required by law that sponsors of new drugs, biologics, food additives, and medical devices conduct animal studies to assess the oncogenic potential of chemicals encountered in food or drugs for the protection of public health. Standard long-term animal carcinogenicity studies of pharmaceuticals and food additives are usually conducted in both sexes of mice and rats for the majority of those animals' typical life spans, generally 24 months of rats and mice. The Center for Drug Evaluation and Research (CDER) in the U.S. Food and Drug Administration (FDA), recommends that drug sponsors conduct carcinogenicity studies at least 18 months in mice and 24 months in rats [1]. Kodell et al. [2] studied the effect of shortened duration on the statistical power of carcinogenicity studies, and the results support the CDER recommendation. The studies often involve a problem of testing the statistical significance of a dose-response relationship. Refer to Ahn and Kodell [3] for various statistical testing schemes for a dose-response relationship.

Peto [4] proposed an approach that requires cause-of-death (COD) for each animal or context of observation for each tumor to be determined by pathologists. The statistical analysis of animal carcinogenicity data and the Peto COD controversy are current issues in the government-regulated pharmaceutical industry [5–8]. An analysis of the COD assignment in selected rodent carcinogenicity studies at the National Center for Toxicological Research (NCTR) indicated that pathologists were inclined to assign a single lesion as the probable COD for most dead/moribund animals [9]. Unfortunately, the analysis could not assess the degree of accuracy of COD assignment in these studies. An appropriate alternative to the Peto-type tests is the Poly-3 test [10], which has been adopted by the National Toxicology Program (NTP) as its official test for carcinogenicity. The Poly-3 procedure does not require COD data but does require a critical assumption about the shape of the Weibull-family tumor onset distribution. The Poly- k test [10; see also 11] is a survival-adjusted quantal-response procedure that modifies the Cochran [12] - Armitage [13,14] trend test to take dose-group differences in intercurrent mortality into account. Intercurrent mortality refers to all deaths other than those resulting from a tumor of interest.

The Poly- k method uses a fractional weighting scheme for animals not at full risk of tumor development (i.e., animals dying from a natural cause without the tumor of interest). The fractional weight w_{ij} is defined as 1 for animals dying with the tumor, and $(t_{ij}/t_{\max})^k$ for animals dying without the tumor, where t_{ij} is the death time of the j th animal in the i th treatment group, and t_{\max} is the time of terminal sacrifice. The weight w_{ij} can be related to the value of a shape parameter k of the Weibull survival function [15]. The value $k = 3$ was recommended by Bailer and Portier [10] following an evaluation of neoplasm onset time distributions in control F344 rats and B6C3F1 mice [16]. They suggested that this test can be improved by using a general k reflecting the shape of the tumor onset distribution. Bieler and Williams [17] further modified the Poly-3 test by an adjustment of the variance estimation of the test statistic using the delta method [18], and showed that the Bailer-Portier Poly-3 test is anticonservative for low tumor incidence rates and for high treatment toxicity. Recently, Chen et al. [19] explained well the characteristics of the Bailer-Portier Poly-3 test and the Bieler-Williams Poly-3 test through a simulation study.

The performance of the Poly-3 test depends on how closely it represents the correct specification

of the time-at-risk weight in the data. As Bailer and Portier mentioned, if the shape of the tumor incidence function is expected to follow time to some power k , which is different from 3, then the Poly- k test with $k \neq 3$ should have superior operating characteristics to the Poly-3 test.

Moon et al. [15] proposed a method for estimating k for data with interval sacrifices. Interval sacrifices include interim sacrifices and a single terminal sacrifice. Estimation of k for data with a single terminal sacrifice is more difficult than that for data with interval sacrifices due to the lack of information on tumor development among live animals before the termination of the experiment. Since most of the animal carcinogenicity studies are designed with a single terminal sacrifice, as an alternative approach to estimating k for data with a single terminal sacrifice, we propose the method of age-adjusted bootstrap-based resampling to improve the Poly-3 test for data collected in a two-year carcinogen bioassay with a single terminal sacrifice via a modification of the permutation method of Farrar and Crump [20] which was used for exact statistical tests.

Our goal is to develop a statistical testing methodology for a dose-related trend in tumor incidence rates of non-palpable tumors. The main idea is to replace z_α from the normal 5% significance level with a bootstrap critical value for the Poly-3 test of Bailer and Portier via our age-adjusted bootstrap method instead of directly estimating k in the Poly- k test. In the proposed test, we not only preserve the tumor incidence rate under H_0 but also do not alter the competing risks survival rate. We use an age-adjusted resampling scheme and assess the significance of the Poly-3 test while taking into account the presence of competing risks that are possible COD. We propose the methods in Section 2. We evaluate the performance of our age-adjusted bootstrap-based methods over the Poly-3 test via Monte Carlo simulation studies described in Section 3.

The proposed procedures are applied to NTP data sets in the 2-year gavage study of furan (C_4H_4O) in F344/N rats and B6C3F₁ mice [22]. To illustrate our procedures, we concentrate on carcinogenic activity of furan in female F344/N rats based on increased incidences of cholangiocarcinoma or hepatocellular neoplasms of the liver and in male B6C3F₁ mice on incidences of adenocarcinoma or alveolar/bronchiolar adenoma of the lung. Table 1 shows the two data sets to be illustrated in the Examples described in Section 4.

2 AGE-ADJUSTED BOOTSTRAP-BASED METHOD

Estimation of the tumor incidence rate is not an easy task for a non-palpable tumor because of various confounding factors such as the presence of treatment-induced mortality unrelated to the tumor of interest. The Poly- k statistic is asymptotically standard normal under the null hypothesis of equal tumor incidence rates among the dose groups [17]. This assumption is valid only if the correct value of k is used in the Poly- k test. In order to find a suitable k for the Poly- k test, Moon et al. [15] recently proposed a method to estimate k for data with interval sacrifices.

In this study, we develop the method of bootstrap resampling with an age-adjusted scheme as an alternative approach to estimate k for the Poly- k test. We estimate the empirical distribution of the test statistic and the corresponding critical value of the Poly-3 test while taking into account the presence of competing risks. It is accomplished via a modification of the permutation method of Farrar and Crump [20] used in exact statistical tests. An approximately valid permutation procedure on the same general footing as the bootstrap method could also be developed. The goal of this study is to make the Poly-3 test robust to the various Weibull-family tumor onset distributions and various competing risks survival rates in rodent bioassays with a single terminal sacrifice.

For a data set, B bootstrap samples will be generated as in Figure 1. With the notation X and X^* denoting the original sample (a data set) and the bootstrap sample respectively, the bootstrap resampling can be carried out as described in Algorithm 1. X and X^* are vectors for each animal containing death times of animals and tumor status across the G dose groups.

Algorithm 1 *Bootstrap Method*

1. A data set X is used to calculate the Poly-3 statistic $T(X)$.
2. B bootstrap samples $X^{*1}, X^{*2}, \dots, X^{*B}$ are generated from the pooled original sample X . Each bootstrap sample contains n elements, uniformly generated by sampling with replacement from the original data set X . The chosen animals are randomly assigned to each of the G groups for conducting the test. Then, $T(X^{*1}), T(X^{*2}), \dots, T(X^{*B})$ are obtained by calculating the value of the test statistic on each bootstrap sample.

3. The critical value $CR(X)$, a threshold for rejecting the null hypothesis of equal tumor incidence rate at the significance level α , is estimated by the $100(1 - \alpha)$ th percentile of the values $T(X^{*1}), T(X^{*2}), \dots, T(X^{*B})$.
4. If $T(X) \geq CR(X)$, then the null hypothesis of equal tumor incidence rate across dose groups is rejected.

The above method is suitable for data with the same competing risks survival rate (CRSR). However, we need to note that if the CRSR is different across dose groups in the original data, the bootstrap samples from the pooled data may not reflect the CRSR of each group, while satisfying the null distribution of equal tumor incidence rates across dose groups. In order to preserve the survival rates in each dose group, we need to modify the above bootstrap method and develop an age-adjusted bootstrap-based scheme. The proposed method is illustrated in Algorithm 2.

Algorithm 2 *Age-adjusted Bootstrap Scheme*

1. A data set X , consisting of information on animals about death times and presence or absence of tumor across the G dose groups, is used to calculate the test statistic $T(X)$ as illustrated in Algorithm 1.1.
2. For the i th group, partition the total days of observation into $I(i, m), i = 1, \dots, G; m = 1, \dots, M_i$, consecutive intervals according to death times of animals in that group. These intervals need not correspond across groups, either in the number of intervals or in the number of days assigned to particular intervals. We denote by $A(I(i, m))$ the set of animals of the i th group whose death time is in the interval $I(i, m)$. We pool the animals that died or were sacrificed within the given interval, say $I(i, m)$, across dose groups. If this interval or interval $I(i, m + 1)$ does not contain any animal that died in other groups, then these two intervals are merged. This merging process is continued until the interval contains at least one animal from another group.
3. Intervalization for bootstrap is followed by pseudo-permutation of animals in each bootstrap-interval. Suppose the m th interval of the i th group contains r animals, and the m th interval

contains R animals in total. For each bootstrap re-sample, we first pool the animals within this interval and shuffle them to randomly select r animals without replacement for the i th group. Suppose there are x tumor-death animals among those r animals. We then bootstrap r animals among the R animals in the m th interval. If the number of tumor deaths exceeds x in the bootstrap of r animals, we randomly select animals that died without tumor in that interval to replace the excess tumor-death animals. This step is required for bootstrap to preserve the mortality pattern in each group, while satisfying the null distribution of equal tumor incidence rate across dose groups. The B bootstrap samples, based on these strata, are generated as illustrated in Algorithm 1, Steps 2 through 4.

Table 2 illustrates the method for generating the first few animals in each group. In Group 1, the first two intervals in days are $(0,185]$ and $(185, 345]$. In this group, animals (A–D) in the first interval are shuffled at each time of bootstrap to determine the maximum allowable number of tumor-death animals for the b th bootstrap sample, where $b = 1, \dots, B$. Typically, the number of bootstrap samples should be at least 1000 for bootstrap confidence interval construction [21]. We take $B = 5000$ bootstrap samples in this study. Then, the first bootstrap animal is generated randomly from the four animals (A–D) that died within interval $(0,185]$. If the maximum allowable tumor deaths for the b th bootstrap sample is zero, but the bootstrap animal died with tumor, then the animal is discarded, and an animal died without tumor among animals (A–D) is re-sampled. Otherwise, the bootstrap animal is kept. On the other hand, if the maximum allowable tumor death for the b th bootstrap sample is one, then the first bootstrap animal generated randomly from animals (A–D) is kept regardless of the tumor status of the first bootstrap animal. The second animal is generated randomly from the eight animals (E–L) in $(185, 345]$ in the same fashion.

In Group 2, the first interval is chosen to be $(0, 176]$ because there is no animal that died in other groups in $(150, 176]$. In this group, the first two animals are randomly sampled with replacement from the three animals (A–C) in $(0, 176]$ up to but not more than maximum allowable number of tumor animals determined in the shuffle, and the third animal is randomly generated from the eight animals (D–K) in $(176, 343]$ in the same way. In Group 3, the first two animals are generated randomly from the seven animals (A–G) in $(0, 316]$ with replacement, and the third

animal is generated randomly from the seven animals (H–N) in (316, 385] in the same fashion. In Group 4, the first two animals are randomly generated from the five animals (A–E) in (0, 243] with replacement, and the next three animals are randomly generated from the five animals (F–J) in (243, 341] with replacement as explained above. The sixth animal is randomly generated from the four animals (K–M) in the same way, and so forth.

3 SIMULATION STUDY

A comprehensive Monte Carlo simulation study is conducted to evaluate the improvement of the proposed test over the Poly-3 test in terms of the robustness to a variety of Weibull-family tumor onset distributions. A typical bioassay design with a control and three dose groups of 50 animals each and an experimental duration of 2 years is used in this study according to standard designs of NTP. Data are generated to have only a single terminal sacrifice at the end of an experiment. Hence, all the remaining live animals are sacrificed at the end of the experiment. The dose levels for control, low, intermediate and high dose groups used in the simulation are 0, 1, 2 and 4, respectively.

It is assumed that three independent random variables T_1 (time to tumor onset), T_2 (time from onset until death from the tumor), and T_3 (time until death from a competing risk) completely determine the observed outcome for each animal. The survival function of T_1 for the i th group is modeled as the one used in Moon et al. [15] such that

$$S_i(t) = \exp \left\{ -\delta(l_i) (t/t_{\max})^k \right\}, \quad (1)$$

where t_{\max} represents the duration of the study or the time for a terminal sacrifice and l_i is a dose level for the i th group. The value of k is set to be 1.5, 3 or 6 for the Weibull tumor onset distribution with a shape parameter of 1.5, 3 or 6, respectively. If the onset distribution differs from Weibull-family distribution, it violates the assumption of the Poly- k test, which assumes the shape of the Weibull-family tumor onset distributions. The value of $\delta(l_i)$ is chosen such that the probability of tumor onset by the end of the experiment attains the desired rate. The tumor rates are chosen to be either 0.05, 0.15 or 0.30 for the control group. The tumor rates are set to be the

same across dose groups for size evaluations. For power comparisons, the tumor rates at the highest dose group by 104 weeks are set to be 5, 3 and 2 times the background tumor rates of 0.05, 0.15 and 0.30, respectively. The low and intermediate doses are chosen to be a quarter and a half of the highest dose, respectively. Our method is focused on common tumors, tumor rate of 5% or higher, in our simulations. For data with rare tumors, we expect that age-adjusted exact trend tests [23] would perform better.

The survival function for T_3 is modeled as $Q(t) = \exp\{-\phi(\gamma_1 t + \gamma_2 t^{\gamma_3})\}$, where $\phi \geq 1$, and γ_1 , γ_2 and γ_3 are nonnegative. With $\phi = 1$, $\gamma_1 = 10^{-4}$ and $\gamma_2 = 10^{-16}$, γ_3 is calculated as $\log[-\{\log Q_C(t_{\max}) + \gamma_1 t_{\max}\}/\gamma_2] / \log t_{\max}$ under the constraint that $Q_C(t_{\max}) < \exp(-\gamma_1 t_{\max})$, where $Q_C(t_{\max})$ is the probability of survival with respect to competing risks in the control group at the end of the study. The constraint $Q_C(t_{\max}) < \exp(-\gamma_1 t_{\max})$ is satisfied in our simulations because $\exp(-\gamma_1 t_{\max}) = 0.9896$ and $Q_C(t_{\max}) \leq 0.7$. The value of ϕ varies such that $\phi = \log(\psi) / \log\{Q_C(t_{\max})\}$, if the survival rate is ψ . Several combinations of the CRSR are considered in order to represent most of the actual animal tumor experiments. The combinations of CRSR considered in this simulation are three common types of mortality patterns: constant CRSR (0.6, 0.6, 0.6, 0.6), decreasing CRSR (0.6, 0.5, 0.4, 0.3), (0.6, 0.6, 0.5, 0.2), (0.5, 0.5, 0.5, 0.2), and umbrella shaped CRSR (0.5, 0.6, 0.5, 0.4), (0.5, 0.7, 0.6, 0.4), (0.5, 0.7, 0.6, 0.5) for the control and three dose groups. These mortality patterns are commonly seen in practice. The rationale is that in NTP feeding studies, the CRSR was about 50% for male F344 rats [24], but it was over 70% for mice and around 60% for female rats. The survival distribution for tumor-induced mortality, T_2 , has the same form as the one for death from competing risks. For each configuration in our simulation study, 5,000 simulated data sets are generated and a nominal significance level of $\alpha = 0.05$ is used. For each data set, 5,000 bootstrap samples are generated. We investigate the improvement of our method over the conventional Poly-3 test in terms of robustness to each tumor onset distribution with various mortality rates. The modified version of the Poly-3 test [17] is used in this comparison.

Table 3 shows the size and the power comparisons of the proposed method to the Poly-3 (P3) test of Bailer and Portier with the variance adjustment suggested by Bieler and Williams with 5000

simulation trials, 5000 bootstrap samples in each trial at the 5% significance level. Our method shows the false positive rate from .033 to .055 across all considered configurations and from .046 to .047 on the average for each value of the Weibull shape parameter (See the last row in Table 3). On the other hand, the P3 test shows the false positive rate from .018 to .072 across all considered configurations and from .039 to .059 on the average. The power varies from .67 to .95 across all considered configurations and from .81 to .88 on the average for our method, while it varies from .64 to .97 across all considered configurations and from .82 to .94 on the average for the P3 test. If CRSRs are the same across the groups, the size of our method ranges from .049 to .055, while the size of the P3 test varies from .050 to .058. The power ranges from .87 to .95 for our method, while it varies from .90 to .96 for the P3 test. If CRSRs are different across the dose groups, the false positive rate ranges from .033 to .054 for the proposed method. However, the false positive rate of the P3 test varies from .018 to .072. The power of the proposed test ranges from .67 to .94, while the power of the P3 test ranges from .64 to .97.

According to our simulation studies, our age-adjusted bootstrap-based Poly- k test appears to have an improvement over the Poly-3 test of Bailer and Portier with the variance adjustment suggested by Bieler and Williams in terms of robustness to various Weibull-family distributions of tumor onset and various competing risks survival rates considered. Use of the proposed approach lessens anticonservatism for the Poly-3 test when the tumor onset distribution is close to the exponential distribution, and improves conservatism of the Poly-3 test when the distribution is far away from the exponential distribution (i.e., Weibull distribution with a shape parameter which is substantially larger than 3).

4 EXAMPLE: THE 2-YEAR GAVAGE STUDY OF FURAN

Furan (C_4H_4O), a clear and colorless liquid serves primarily as an intermediate in the synthesis and preparation of numerous organic compounds [22]. Toxicology and carcinogenesis studies were conducted by administering furan in corn oil by gavage to groups of F344/N rats and B6C3F₁ mice of each sex for two years. Furan was nominated by the National Cancer Institute for evaluation of carcinogenic potential due to its large production volume and use and because of the potential

for widespread human exposure to a variety of furan-containing compounds. Corn oil was chosen as the vehicle for these studies since furan is highly volatile and insoluble in water but soluble in alcohol, ether, and most common organic solvents [25].

In this example, we focus on female F344/N rats for evaluation of carcinogenic potential on incidences of cholangiocarcinoma or hepatocellular neoplasms of the liver and male B6C3F₁ mice on incidences of adenocarcinoma or alveolar/bronchiolar adenoma of the lung. The test results on the incidence of neoplasms from our proposed methods are compared to those from the conventional Poly-3 test [17] which has been adopted by NTP. Groups of 50 rats of each sex were administered 2, 4 or 8 mg furan per kg body weight in corn oil by gavage 5 days per week for 2 years, and groups of 50 mice of each sex received doses of 8 or 15 mg/kg furan 5 days per week for 2 years.

Table 4 shows test results on the carcinogenic activity of furan in female F344/N rats based on increased incidences of cholangiocarcinoma and hepatocellular neoplasms of the liver. Data are summarized in the first half of Table 1. We test the carcinogenic activity of furan with groups 2, 3 and 4 via our test and the P3 test. Overall test results, results with 3 combinations of 3 groups, and results with 2 combinations (0, 4 and 0, 8 mg/kg) of two groups from our test and the P3 test agree and indicate that there is a significant difference on incidences of neoplasms of the liver due to carcinogenic activity of furan. On the other hand, the two-group comparison of the vehicle control versus the 2 mg/kg dose group shows different conclusions between our test and the P3 test. In contrast with the result of the P3 test ($p = .072$), our test indicates that there is marginal carcinogenic activity of furan ($p = .040$). NTP concluded that under the conditions of these 2-year gavage studies, there was clear evidence of carcinogenic activity of furan in female F344/N rats based on increased incidences of cholangiocarcinoma and hepatocellular neoplasms of the liver. Results from our test applied to any combination of experimental groups agree with the conclusions of NTP.

Table 5 shows the results from our test and the P3 test for the carcinogenic potential of furan on incidences of adenocarcinoma and alveolar/bronchiolar adenoma of the lung in male B6C3F₁ mice. Data are summarized in the second half of Table 1. The test results from our test and the P3 test show different conclusions except one from a two-group test between the vehicle control and

the 8 mg/kg dose group. All the results from our test clearly do not show significant carcinogenic activity of furan on incidences of neoplasms in the lung ($p > .05$), and the test results agree with the conclusions from NTP. On the other hand, the P3 test shows marginally significant carcinogenic activity of furan among three dose groups ($p = .045$) and between the vehicle control and the 15 mg/kg dose group ($p = .046$).

5 DISCUSSION

The importance of the Poly- k test has been highlighted recently. In May 2001, FDA's Center for Drug Evaluation and Research (CDER) published in the Federal Register a guidance document [8]. In that document, CDER endorsed the use of the Peto procedure for analyzing tumorigenicity data, although the Poly-3 test was mentioned as a possible alternative when cause-of-death (COD) data (equivalently, context-of-observation data) are not available. The Society of Toxicologic Pathologists (STP) published in its journal a commentary on CDER's guidance that was critical of Peto's procedure and that largely endorsed the Poly-3 procedure [6]. The Poly-3 procedure does not require COD data but does require a critical assumption about the shape of the tumor onset distribution. Later, the Society withdrew its criticisms of the CDER guidance document on the Peto approach, while still recognizing the appropriateness of the Poly-3 test in certain situations [7].

It is clear that the statistical analysis of tumorigenicity data from animal bioassays remains an important regulatory issue to FDA and the pharmaceutical industry. The present research builds on previous research by the authors to further refine the Poly-3 test in order to make it more robust to a variety of Weibull-family tumor onset distributions and various intercurrent mortality rates in rodent bioassays with a single terminal sacrifice via the proposed age-adjusted bootstrap method although the shape of Weibull-family tumor onset distribution is different from 3. Our simulation shows that our bootstrap-based age-adjusted Poly-3 test for dose-related trend is robust to a variety of Weibull-family tumor onset distributions and various competing risks survival rates. It appears to control the false positive rate better than the Poly-3 test, thus having enhanced performance in identifying dose-related trends. Because the Poly-3 test does not require COD data or any information on tumor lethality, the improved version can be used confidently when Peto's test can

not be implemented due to lack of cause-of-death information.

For illustration, our test was applied to NTP data sets of the 2-year gavage study of furan. Comparison of the proposed test to the Poly-3 test of Bieler and Williams was made in the application to a data set of female F344/N rats on incidences of cholangiocarcinoma and hepatocellular neoplasms of the liver and a data set of male B6C3F₁ mice on incidences of adenocarcinoma and alveolar/bronchiolar adenoma of the lung. Contrary to the conclusions obtained from the Poly-3 test of Bailer and Portier with variance adjustment suggested by Bieler and Williams, the proposed test detected significant carcinogenic activity in the liver of rats, but did not detect carcinogenic activity in the lung of mice, in agreement with the conclusion of the National Toxicology Program.

ACKNOWLEDGEMENTS

Hongshik Ahn's work was supported by NIH grant 1 R29 CA77289-06 and was partially supported by the Faculty Research Participation Program at the National Center for Toxicological Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between USDOE and USFDA.

REFERENCES

1. OFR (Office of the Federal Register). Chemical Carcinogens; A Review of the Science and Its Associated Principles. Office of Science and Technology Policy. *Federal Register* 1985; **II**:47-58.
2. Kodell RL, Lin KK, Thorn BT, Chen, JJ. Bioassays of Shortened Duration for Drugs: Statistical Implications. *Toxicological Sciences* 2000; **55**:415-432.
3. Ahn H, Kodell RL. Analysis of Long-Term Carcinogenicity Studies. In: *Design and Analysis of Animal Studies in Pharmaceutical Development*, Chow, S. C. and Liu, JP. (Eds.). Marcel Dekker, Inc.: New York, 1998; pp259-289.
4. Peto R. Guidelines on the analysis of tumor rates and death rates in experimental animals. *British Journal of Cancer* 1974; **29**:101-105.
5. Lee PN, Fry JS, Fairweather WR, Haseman JK, Kodell RL, Chen JJ, Roth AJ, Soper K, Morton, D. Current Issues: Statistical Methods for Carcinogenicity Studies. *Toxicological Pathology* 2002; **30**:403-414.

6. STP Peto Analysis Working Group. The Society of Toxicologic Pathology's Position on Statistical Methods for Rodent Carcinogenicity Studies. *Toxicologic Pathology* 2001; **29**(6):670-672.
7. STP Peto Analysis Working Group. The Society of Toxicologic Pathology's Recommendations on Rodent Carcinogenicity Studies. *Toxicologic Pathology* 2002; **30**:415-418.
8. U.S. FDA. Guidance for Industry: Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals. *Federal Register* 2001; **66**(89):23266-23267.
9. Kodell RL, Balckwell BN, Bucci TJ, Greenman DL. Cause-of-Death Assignment at the National Center for Toxicological Research. *Toxicologic Pathology* 1995; **23**:241-247.
10. Bailer AJ, Portier CJ. Effects of Treatment-Induced Mortality and Tumor-Induced Mortality on Tests for Carcinogenicity in Small Samples. *Biometrics* 1988; **44**:417-431.
11. Piegorsch WW, Bailer AJ. *Statistics for Environmental Biology and Toxicology*, Chapman and Hall; London, 1997.
12. Cochran WG. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics* 1954; **10**:417-451.
13. Armitage P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* 1955; **11**:375-386.
14. Armitage P. *Statistical Methods in Medical Research*. John Wiley: New York, 1971.
15. Moon H, Ahn H, Kodell RL, Lee JJ. Estimation of k for the Poly- k Test. *Statistics in Medicine* 2003; **22**:2619-2636.
16. Portier C, Hedges J, Hoel DG. Age-specific Models of Mortality and Tumor Onset for Historical Control Animals in the National Toxicology Program's Carcinogenicity Experiments. *Cancer Research* 1986; **46**:4372-4378.
17. Bieler GS, Williams RL. Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity. *Biometrics* 1993; **49**:793-801.
18. Woodruff RS. A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association* 1971; **66**:411-414.
19. Chen JJ, Lin KK, Huque MF, Arani RB. Weighted p -value for Animals Carcinogenicity Trend Test. *Biometrics* 2000; **56**:586-592.
20. Farrar DB, Crump KS. Exact Statistical Tests for Any Carcinogenic Effect in animal Bioassays, II. Age-Adjusted Tests. *Fundamental and Applied Toxicology* 1990; **15**:710-721.
21. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*, Chapman and Hall; New York, 1993.
22. National Toxicology Program. Toxicology and Carcinogenesis Studies of Furan in F344/N Rats and B6C3F₁ Mice (Gavage Studies). *NTP Technical Report* 1993; **402**, Research Triangle Park, NC.

23. Mancuso JY, Ahn H, Chen JJ, Mancuso JP. Age-Adjusted Exact Trend Tests in the Event of Rare Occurrences. *Biometrics* 2002; **58**:403-412.
24. Haseman JK, Hailey JR, Morris RW. Spontaneous Neoplasm Incidences in Fischer 344 Rats and B6C3F₁ Mice in Two-Year Carcinogenicity Studies: A National Toxicology Program Update. *Toxicologic Pathology* 1998; **26**:428-441.
25. The Merck Index. 10th ed. (M. Windholz, Ed.), Merck & Company; Rahway, NJ; 1983.

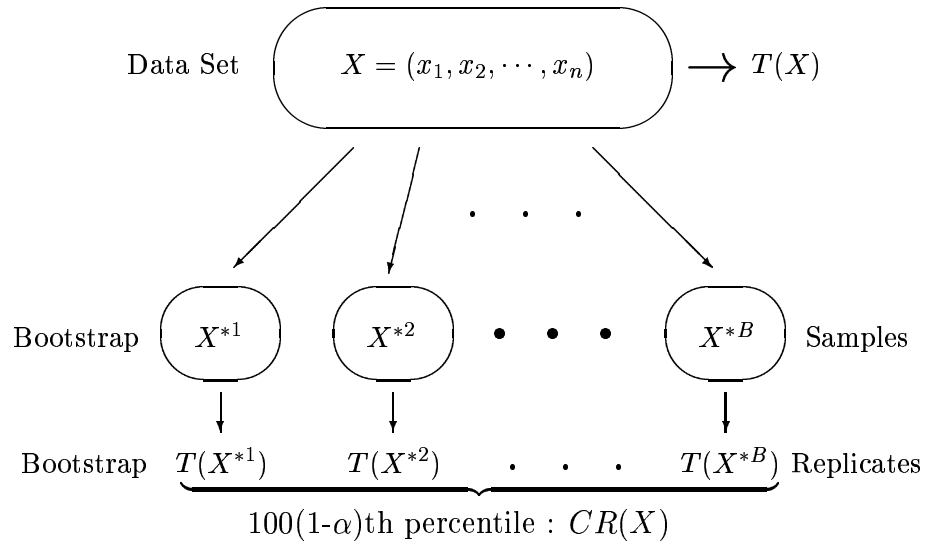


Figure 1: Bootstrap Method

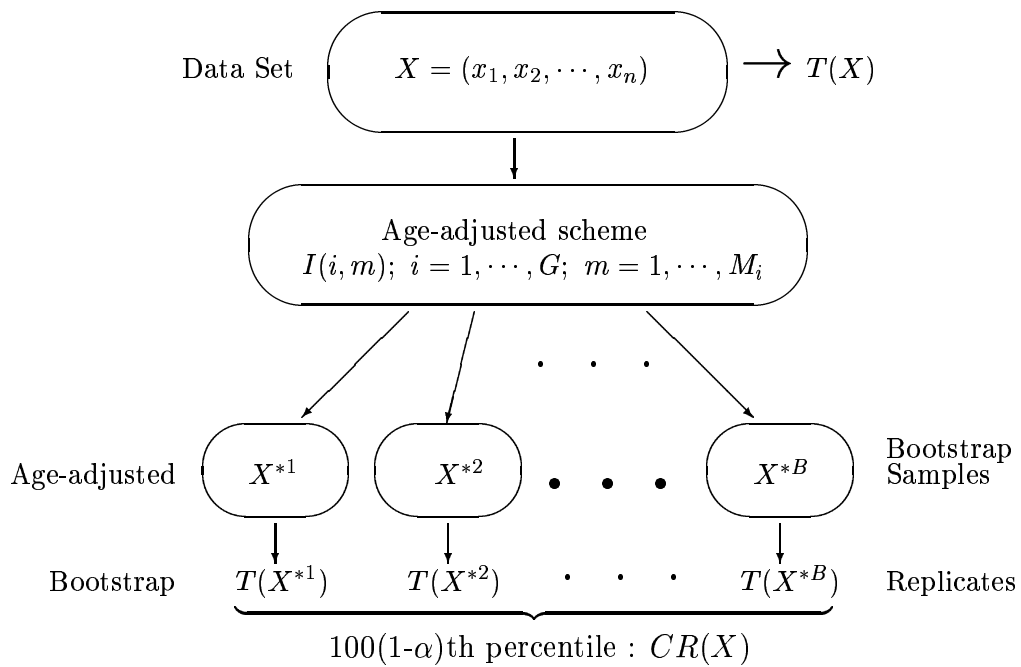


Figure 2: Age-Adjusted Bootstrap Scheme

Table 1: Animal tumor pathology of female F344/N rats (liver) and male B6C3F₁ mice (lung) in the 2-year gavage study of furan

	Group	Animal Tumor Pathology ^a
Liver ^b	Vehicle Control	1(0), 2(16), 3(0), 4(34)
	2 mg/kg	1(1), 2(17), 3(1), 4(31)
	4 mg/kg	1(3), 2(19), 3(3), 4(25)
	8 mg/kg	1(4), 2(27), 3(6), 4(13)
Lung ^c	Vehicle Control	1(3), 2(14), 3(4), 4(29)
	8 mg/kg	1(4), 2(24), 3(3), 4(19)
	15 mg/kg	1(7), 2(23), 3(6), 4(14)

^a $x(y)$: x 's are 1 for death with tumor of interest, 2 for death without tumor, 3 for sacrifice with tumor and 4 for sacrifice without tumor, and y represents the number of animals in the tumor pathology.

^bCholangiocarcinoma or hepatocellular neoplasms of the liver in female F344/N rats

^cAdenocarcinoma or alveolar/bronchiolar adenoma of the lung in male B6C3F₁ mice

Table 2: Death times (in days) with tumor status in parenthesis in a hypothetical animal carcinogenicity data set with four groups. Each number in parenthesis represents 1 for death with tumor, 2 for death without tumor.

ID	Group 1	Group 2	Group 3	Group 4
A				74 (2)
B		150 (2)		
C		176 (2)		
D	185 (2)			
E				243 (1)
F			300 (2)	
G			316 (2)	
H				324 (2)
I				340 (1)
J				341 (2)
K		343 (2)		
L	345 (2)			
M				351 (2)
N			385 (2)	
			⋮	

Table 3: Size and power evaluation with 5000 simulation trials and 5000 bootstrap samples in each trial (5% significance level)

TR ^a	CRSR ^b	Size						Power					
		Weib1.5		Weib3.0		Weib6.0		Weib1.5		Weib3.0		Weib6.0	
		B ^c	N ^d	B	N	B	N	B	N	B	N	B	N
.05	.6, .6, .6, .6	.052	.058	.050	.056	.054	.055	.90	.92	.89	.91	.87	.90
	.5, .5, .5, .2	.046	.059	.047	.051	.046	.037	.82	.91	.77	.85	.71	.72
	.6, .6, .5, .2	.042	.064	.043	.047	.040	.035	.79	.92	.74	.85	.68	.72
	.6, .5, .4, .3	.048	.061	.049	.055	.048	.047	.86	.92	.82	.88	.77	.79
	.5, .6, .5, .4	.049	.057	.050	.055	.047	.050	.88	.92	.85	.89	.81	.83
	.5, .7, .6, .4	.047	.056	.049	.051	.046	.047	.87	.92	.84	.89	.80	.83
	.5, .7, .6, .5	.050	.058	.050	.052	.050	.052	.89	.92	.87	.90	.84	.86
.15	.6, .6, .6, .6	.049	.053	.052	.051	.051	.054	.95	.96	.95	.96	.94	.94
	.5, .5, .5, .2	.045	.062	.042	.045	.041	.030	.89	.96	.85	.91	.80	.78
	.6, .6, .5, .2	.036	.066	.038	.044	.037	.027	.86	.96	.82	.90	.76	.77
	.6, .5, .4, .3	.047	.062	.045	.048	.043	.035	.91	.97	.89	.93	.85	.85
	.5, .6, .5, .4	.047	.056	.049	.047	.047	.042	.93	.96	.91	.94	.89	.89
	.5, .7, .6, .4	.045	.056	.044	.044	.044	.039	.93	.96	.91	.94	.88	.89
	.5, .7, .6, .5	.049	.051	.049	.047	.050	.046	.94	.96	.93	.95	.91	.92
.30	.6, .6, .6, .6	.053	.050	.054	.050	.055	.052	.92	.93	.91	.92	.89	.90
	.5, .5, .5, .2	.044	.066	.044	.041	.040	.021	.84	.93	.78	.85	.73	.67
	.6, .6, .5, .2	.036	.072	.033	.037	.033	.018	.79	.94	.73	.85	.67	.64
	.6, .5, .4, .3	.047	.069	.043	.045	.040	.024	.86	.94	.83	.88	.77	.75
	.5, .6, .5, .4	.049	.055	.050	.048	.048	.037	.89	.93	.87	.90	.83	.82
	.5, .7, .6, .4	.046	.053	.048	.046	.045	.036	.88	.93	.86	.89	.82	.81
	.5, .7, .6, .5	.054	.050	.051	.047	.054	.044	.90	.93	.88	.91	.86	.87
Average		.047	.059	.047	.048	.046	.039	.88	.94	.85	.90	.81	.82

^aTR: Tumor rate^bCRSR: Competing risks survival rate for the control and three dose groups^cB: The proposed method^dN: The Poly-3 test of Bailer and Portier with variance adjustment suggested by Bieler and Williams

Table 4: Animal tumor pathology of female F344/N rats in the 2-year gavage study of furan based on increased incidences of cholangiocarcinoma or hepatocellular neoplasms of the liver (Reject when $T(X) \geq CR(X)$; Boldface represents the places where our test results are different from the P3 test results.)

mg/kg	$T(X)_{\text{BW}}^a$	$CR(X)_{\text{Normal}}^b$	$p\text{-value}_{\text{Normal}}$	$CR(X)_{\text{Bootstrap}}^c$	$p\text{-value}_{\text{Bootstrap}}$
Overall	4.1617	1.6449	< .001	2.0141	< .001
0, 2, 4	2.7705	1.6449	.003	1.9584	.004
0, 2, 8	4.3559	1.6449	< .001	2.0603	< .001
0, 4, 8	3.6632	1.6449	< .001	1.8214	< .001
^d 0, 2	1.4641	1.6449	.072	1.4625	.040
0, 4	2.6542	1.6449	.004	1.5905	.001
0, 8	3.8420	1.6449	< .001	1.7423	< .001

^aThe P3 test statistic obtained from the data

^bStandard normal critical value at the significance level 0.05

^cCritical value estimated by the 95th percentile of $T(X)$'s from our method

^dThe proposed method shows significant effect ($p\text{-value} = .040$) at the level of .05, while the P3 test shows non-significant effect ($p\text{-value} = .072$) at the level of .05.

Table 5: Animal tumor pathology of male B6C3F₁ mice in the 2-year gavage study of furan based on incidences of adenocarcinoma or alveolar/bronchiola adenoma of the lung in the respiratory system (Reject when $T(X) \geq CR(X)$; Boldface represents the places where our test results are different from the P3 test results.)

mg/kg	$T(X)_{\text{BW}}^a$	$CR(X)_{\text{Normal}}^b$	$p\text{-value}_{\text{Normal}}$	$CR(X)_{\text{Bootstrap}}^c$	$p\text{-value}_{\text{Bootstrap}}$
^d Overall	1.6995	1.6449	.045	1.7774	.058
^e 0, 15	1.6805	1.6449	.046	1.6938	.052
0, 8	0.2229	1.6449	.41	1.9248	.53

^aThe P3 test statistic obtained from the data

^bStandard normal critical value at the significance level 0.05

^cCritical value estimated by the 95th percentile of $T(X)$'s from our method

^dThe proposed method shows non-significant effect ($p\text{-value} = .058$) at the level of .05, while the P3 test shows significant effect ($p\text{-value} = .045$) at the level of .05.

^eThe proposed method shows non-significant effect ($p\text{-value} = .052$) at the level of .05, while the P3 test shows significant effect ($p\text{-value} = .046$) at the level of .05.