Published in final edited form as:
J Am Stat Assoc. 2007 January 1; 102(480): 1212–1220.

Incorporating Historical Control Data When Comparing Tumor Incidence Rates *

Shyamal D. Peddada, **Gregg E. Dinse**, and **Grace E. Kissling**

Abstract

Animal carcinogenicity studies, such as those conducted by the U.S. National Toxicology Program (NTP), focus on detecting trends in tumor rates across dose groups. Over time, the NTP has compiled vast amounts of data on tumors in control animals. Currently, this information is used informally, without the benefit of statistical tests for carcinogenicity that directly incorporate historical data on control animals. This article proposes a survival-adjusted test for detecting dose-related trends in tumor incidence rates, which incorporates data on historical control rates and formally accounts for variation in these rates among studies. An extensive simulation, based on a wide range of realistic situations, demonstrates that the proposed test performs well in comparison to the current NTP test, which does not incorporate historical control data. In particular, our test can aid in interpreting the occurrence of a few tumors in treated animals that are rarely seen in controls. One such example, which motivates our work, concerns the analysis of histiocytic sarcoma in the NTP's 2-year cancer bioassay of benzophenone. Whereas the occurrence of three histiocytic sarcomas in female rats was not significant according to the current NTP testing procedure (P = 0.074), it was highly significant (P = 0.004) when control data from six recent historical studies were included and our test was applied to the combined data.

Keywords

Cancer bioassay; Carcinogenicity experiment; National Toxicology Program; Order-restricted inference; Poly-3; Quantal response; Survival adjustment

1. INTRODUCTION

In accord with a mandate by the U.S. Congress, the NTP routinely evaluates the toxicity and carcinogenicity of various chemicals to which humans are exposed. The NTP has performed more than 500 long term rodent studies, involving over 450 chemicals, which has produced a rich collection of information about tumors in control animals. Often this database is used informally when assessing whether differences in tumor rates are dose-related, especially when tumors are rare, but a formal method of incorporating historical control data would be preferable.

As an example, the NTP recently conducted a two-year rodent bioassay to evaluate the toxicity and carcinogenicity of benzophenone, a chemical to which a large number of people are exposed (for details, see http://ntp.niehs.nih.gov). As is typical of NTP bioassays, both sexes of two rodent species were studied, with 50 animals randomly assigned to a control group and 50 to each of three dose groups for all four sex/species combinations. The results

*Shyamal D. Peddada (Corresponding author, peddada@niehs.nih.gov), Gregg E. Dinse (dinse@niehs.nih.gov), and Grace E. Kissling (kissling@niehs.nih.gov) are members of the Biostatistics Branch, National Institute of Environmental Health Sciences, MD A3-03, P.O. Box 12233, Research Triangle Park, NC 27709.

of standard analyses of histiocytic sarcoma in female rats were unsatisfying. The observed numbers of rats with this cancer were 0, 0, 1, and 2 in the control, low-dose, mid-dose, and high-dose groups. The usual trend test reported by the NTP was not statistically significant ($P = 0.074$) at the standard 5% level, but the NTP historical control database shows that spontaneous occurrences of this neoplasm are very rare. None of the 460 female rats among the control groups in 6 recent studies had a histiocytic sarcoma, yet 2 of 50 in the current high-dose group and 1 of 50 in the current mid-dose group developed this cancer. We are interested in determining if the trend in the benzophenone study is significant after formally taking into account that, historically, this tumor is extremely rare.

Over the past few decades, several methods have been proposed for including historical information in the analysis of current data. Many of the procedures focus on lifetime tumor rates and assume a beta-binomial model, which allows for extra-binomial variation among the tumor proportions from the historical control groups (see, for example, Tarone 1982; Yanagawa and Hoel 1985; Hoel and Yanagawa 1986; Tamura and Young 1986; Prentice et al., 1992; Ryan 1993). Similar approaches have been based on generalized binomial (Makuch et al., 1989) and logistic-normal (Dempster et al., 1983; Seewald 1994) models. A major disadvantage of these methods, however, is that they do not adjust for survival. When treatment a ects longevity, the time at risk for developing tumors can differ substantially across dose groups, in which case survival-adjusted procedures are necessary to avoid biases.

Generalizations have been suggested that account for survival, but they have various shortcomings of their own. For example, Ibrahim and Ryan (1996) proposed a Dirichlet-multinomial model that assumes tumors are rapidly lethal, while other methods are oriented toward nonlethal tumors (Ibrahim et al., 1998; Parise et al., 2001). Bayesian procedures have been developed that adjust for survival and do not make extreme lethality assumptions (Dunson and Dinse 2001; French and Ibrahim 2002), but these analyses require investigators to specify prior distributions and hyperparameters.

The need for a formal method of incorporating historical control data in the analysis of the current experiment has long been recognized (Haseman et al., 1984; Haseman 1995), but no procedure has emerged as the clear favorite (Greim et al., 2003). The Technical Reports Review Subcommittee of the NTP Board of Scientific Counselors, which includes two statisticians, has not endorsed any of the current methods and recommended a new procedure be developed for this important problem (http://ntp.niehs.nih.gov/files/TRRSMins0905.pdf). This article presents a new method which is relatively simple, incorporates historical control data, adjusts for survival, and avoids lethality restrictions and distributional assumptions. Section 2 describes the proposed statistical procedure and Section 3 reports the results of an extensive simulation study, which shows that our method has good operating characteristics. The proposed test is illustrated in Section 4 by applying it to two NTP data sets, and several concluding remarks are provided in Section 5.

2. METHODS

Early procedures for analyzing carcinogenicity data, with or without incorporating historical control information, dealt with lifetime tumor rates. In general, though, methods should focus on age-specific tumor rates to avoid biases due to differential survival. Of the various age-specific rates, the most appropriate is tumor incidence (McKnight and Crowley 1984), which corresponds to the hazard function for time to tumor onset. As a function of age, the tumor incidence rate automatically adjusts for differential survival, and as the rate at which

new tumors occur, it is a natural measure of carcinogenesis. Thus, we focus on the detection of a positive trend (across dose groups) in the age-specific tumor incidence rates.

2.1 Notation

Let T be the time to the first of two events, either tumor onset or death, and let $Y(t)$ be an indicator that is 1 if the tumor is present at time t and 0 otherwise. The tumor incidence rate and the tumor-free death rate correspond to the event-specific hazard functions

$$\lambda(t) = \lim_{\epsilon \to 0} Pr(t \leq T < t+\epsilon, Y(T) = 1 | T \geq t) / \epsilon$$

and

$$\beta(t) = \lim_{\epsilon \to 0} Pr(t \leq T < t+\epsilon, Y(T) = 0 | T \geq t) / \epsilon ,$$

respectively. Let π be the expected proportion of animals that develop a tumor during the study (i.e. the lifetime tumor rate), which can be expressed in terms of $\lambda(t)$ and $\beta(t)$:

$$\pi = \int_0^{t_{TS}} \lambda(s) \exp\left(-\int_0^s [\lambda(r) + \beta(r)]\, dr\right) ds , \quad (1)$$

where t_{TS} denotes the terminal sacrifice time (i.e. the length of the study).

Suppose there are k treatment groups in the current experiment, with associated dose levels $d_1 < d_2 < \cdots < d_k$, where animals in the first group are unexposed controls ($d_1 = 0$). More generally, d_i is a dose score assigned to the i^{th} group ($i = 1, 2, \ldots, k$). Let π_i, $\lambda_i(t)$, and $\beta_i(t)$ denote the lifetime tumor rate, the tumor incidence rate, and the tumor-free death rate in the i^{th} group. Tests for a positive trend in tumor incidence rates involve the following null and ordered alternative hypotheses:

$$H_0: \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_k(t)$$

and

$$H_a: \lambda_1(t) \leq \lambda_2(t) \leq \cdots \leq \lambda_k(t) ,$$

respectively, where at least one inequality in H_a is strict for at least one t. Equation (1) shows that tests oriented toward the $\{\pi_i\}$ are not necessarily valid for comparing the $\{\lambda_i(t)\}$, as such tests can be confounded by differences among the $\{\beta_i(t)\}$.

Suppose n_{ci} animals in the current study are randomized to the i^{th} group and receive dose d_i of the treatment ($i = 1, 2, \ldots, k$). We use c to index variables and parameters related to the current study, and later we use h to index similar quantities related to the historical data. Let y_{cij} and t_{cij} denote the tumor status and death time, respectively, for the j^{th} animal in the i^{th} group of the current study ($j = 1, 2, \ldots, n_{ci}$; $i = 1, 2, \ldots, k$). Thus, y_{cij} is 1 if a tumor is present at necropsy and 0 otherwise. Furthermore, $n_{c+} = \sum_{i=1}^{k} n_{ci}$ is the total number of animals in the current study, and $y_{ci+} = \sum_{j=1}^{n_{ci}} y_{cij}$ and $y_{c++} = \sum_{i=1}^{k} y_{ci+}$ are the numbers of tumor-bearing animals in the i^{th} group and in the entire study, respectively.

2.2 Background

Many early analyses used the linear trend test of Cochran (1954) and Armitage (1955) to compare lifetime tumor rates. Inferences based on lifetime rates can be misleading, though, if survival differs across treatment groups. For example, unless all animals live to the end of the study, the sample proportion y_{ci+}/n_{ci} tends to underestimate π_{ci} because it does not account for early deaths reducing the time at risk for developing tumors. Bias arises when the amount of reduction in the time at risk varies with dose (i.e. differential mortality).

Bailer and Portier (1988) addressed the survival issue by defining a modified sample size $n^*_{ci}=\sum_{j=1}^{n_{ci}}\delta_{cij}$, where δ_{cij} is a weight for animal j in group i. Clearly an animal that developed a tumor was at risk of tumor onset and thus receives a weight of 1. An animal that dies without a tumor, however, is assigned a weight linked to the fraction of the study it survived. Bailer and Portier (1988) suggested using $(t_{cij}/t_{TS})^3$ for this fractional weight, which is known as the Poly-3 weight; then they applied the Cochran-Armitage test after substituting n^*_{ci} for n_{ci}. The resulting survival-adjusted estimate of π_{ci} is $\tilde{\pi}_{ci}=y_{ci+}/n^*_{ci}$. Bieler and Williams (1993) noted that n^*_{ci} is a random variable and proposed a corrected variance estimator, which under the assumption that $\pi_{c1} = \pi_{c2} = \cdots = \pi_{ck}$ is given by

$$var\left(\tilde{\pi}_{ci}\right)=S_1^2/w_{ci}\,,$$

where

$$S_1^2=\sum_{ij}\left(r_{cij}-\bar{r}_{ci}\right)^2/\left(n_{c+}-k\right),\quad r_{cij}=y_{cij}-\tilde{\pi}_c\delta_{cij},\quad \bar{r}_{ci}=r_{ci+}/n_{ci},\quad \tilde{\pi}_c=y_{c++}/n^*_{c+},\quad w_{ci}=n^{*2}_{ci}/n_{ci}$$

, and the plus sign indicates summation over the missing index.

The usual Poly-3 trend test, with the Bieler-Williams variance adjustment, is

$$W_{BW}=\frac{\sum_{i=1}^{k}w_{ci}d_i\tilde{\pi}_{ci}-\left(\sum_{i=1}^{k}w_{ci}d_i\right)\left(\sum_{i=1}^{k}w_{ci}\tilde{\pi}_{ci}\right)/\sum_{i=1}^{k}w_{ci}}{S_1\sqrt{\sum_{i=1}^{k}w_{ci}d_i^2-\left(\sum_{i=1}^{k}w_{ci}d_i\right)^2/\sum_{i=1}^{k}w_{ci}}}=\frac{A}{B}\,.$$

The NTP currently uses the following (conservative) variation of W_{BW}:

$$W_{NTP}=sign\,(A)\left\{\frac{|A|-CF}{B}\right\}\,,$$

which shrinks W_{BW} toward 0. The continuity correction factor in the above formula is

$$CF=\frac{1}{2}max_{i=2,\ldots,k}\left\{\frac{n^*_{ci}\left(d_i-\bar{d}\right)}{n_{ci}}-\frac{n^*_{c,i-1}\left(d_{i-1}-\bar{d}\right)}{n_{c,i-1}}\right\}\,,$$

where $\bar{d}=\sum_{i=1}^{k}w_{ci}d_i/\sum_{i=1}^{k}w_{ci}$ is a weighted average of the dose scores. In the absence of survival differences, CF reduces to half the largest difference between successive dose

scores. Both W_{NTP} and W_{BW} are approximately normal when $\pi_{c1} = \pi_{c2} = \cdots = \pi_{ck}$. Bailer and Portier showed that while the Poly-3 test is oriented toward survival-adjusted comparisons of the $\{\pi_{ci}\}$, it also provides an approximately valid comparison of the age-specific tumor incidence rates when each $\lambda_{ci}(t)$ follows a Weibull model with a shape parameter of 3. However, simulation studies demonstrate that the Poly-3 test is sensitive to certain departures from the underlying Weibull model (Mancuso et al., 2002; Peddada et al., 2005).

Peddada et al. (2005) suggested an alternative trend test which uses the Poly-3 weights, but does not assume the tumor rates are linear in dose, does not depend on the numerical scores assigned to the dose groups, and does not incorporate a continuity correction. Let $\widehat{\pi}_{ci}$ denote the survival-adjusted isotonic regression estimate of π_{ci} ($i = 1, 2, \ldots, k$) under the constraint that $\pi_{c1} \leq \pi_{c2} \leq \cdots \leq \pi_{ck}$. These estimates can be obtained, for example, by applying the pool-adjacent-violators algorithm (PAVA) (Barlow et al., 1972) to the $\{\widehat{\pi}_{ci}\}$ with weights $\{n^*_{ci}\}$. As a simple illustration of PAVA, suppose there are $k = 4$ groups, all weights are equal, the true (unknown) proportions are constrained by $\pi_1 \leq \pi_2 \leq \pi_3 \leq \pi_4$, and the sample proportions are $\tilde{\pi}_1 = .1, \tilde{\pi}_2 = .3, \tilde{\pi}_3 = .2$ and $\tilde{\pi}_4 = .4$, respectively. Clearly, $\tilde{\pi}_2$ and $\tilde{\pi}_3$ violate the constraint. The isotonic solution can be obtained by retaining those estimates which satisfy the inequality constraints while averaging the estimates which are in violation. Thus, in this illustration, the "isotonized" estimates are $\widehat{\pi}_1 = .1, \widehat{\pi}_2 = \widehat{\pi}_3 = (.3 + .2)/2 = .25$, and $\widehat{\pi}_4 = .4$. Mathematically, in the general case, the "isotonized" values $\widehat{\pi}_{ci}$ ($i = 1, 2, \ldots, k$) under the above inequality constraint can be computed using the following min-max formula:

$$\widehat{\pi}_{ci} = \frac{\min_{s \geq i} \max_{t \leq i} \sum_{j=t}^{s} n^*_{cj} \tilde{\pi}_{cj}}{\min_{x \geq i} \max_{t \leq i} \sum_{j=t}^{s} n^*_{cj}}.$$

The isotonic test statistic proposed by Peddada et al. (2005) can be written

$$W_{ISO} = \frac{\widehat{\pi}_{ck} - \widehat{\pi}_{c1}}{S_1 \sqrt{1/w_{c1} + 1/w_{ck}}},$$

which was motivated by a procedure that Williams (1977) derived to test the equality of normal means. Define k standard normal random variables: $Z_i \sim N(0, 1)$ for $i = 1, \ldots, k$. Given equal $\{\pi_{ci}\}$, the distribution of W_{ISO} can be approximated by the distribution of

$$V_{ISO} = \frac{\widehat{Z}_k - \widehat{Z}_1}{\sqrt{2}},$$

where $\widehat{Z}_1 \leq \widehat{Z}_2 \leq \cdots \leq \widehat{Z}_k$ are the isotonized values of the $\{Z_i\}$ using unit weights. Peddada et al. (2005) conducted extensive simulations which showed that their test generally was more powerful than the Bailer-Portier/Bieler-Williams test when comparing age-specific incidence rates, even though both tests are more naturally oriented toward the $\{\pi_{ci}\}$ than the $\{\lambda_{ci}(t)\}$.

2.3 Proposed Test

Suppose there are p relevant studies in the historical control database, where this number depends on several factors. Generally, we would prefer to include only those historical

studies for which the route of exposure (e.g. feed, gavage, inhalation, skin painting) is the same as in the current study. Also, we typically restrict our comparison to studies which were performed relatively recently, since background tumor rates are known to "drift" over time (Haseman, 1992). Thus, in practice, we anticipate incorporating control data from approximately 5 to 10 historical studies.

Let n_{hm} be the number of animals in the m^{th} historical control group, and let y_{hmj} and t_{hmj} denote the tumor status and death time, respectively, for animal j in group m ($j = 1, 2, \ldots, n_{hm}$; $m = 1, 2, \ldots, p$). In addition, set $y_{hm+} = \sum_{j=1}^{n_{hm}} y_{hmj}$ and $y_{h++} = \sum_{m=1}^{p} y_{hm+}$. Finally, let π_{hm} denote the proportion of animals [summationtext] in the m^{th} historical control group that are expected to develop a tumor if surviving the entire study (i.e. the lifetime tumor rate).

The problem of detecting trends in tumor rates can be viewed as a special case of order restricted inference and can be solved by maximizing certain distances in a directed graph. Figure 1 shows a directed graph appropriate for the usual NTP situation, where there is a high-dose (HD) group, a mid-dose (MD) group, a lowdose (LD) group, a current control (CC) group, and a collection of historical control (HC) groups. The arrows indicate the restrictions that the tumor rates in the control groups do not exceed the rate in the low-dose group, which does not exceed the rate in the mid-dose group, which does not exceed the rate in the high-dose group. The control animals, both current and historical, are assumed to come from a common population, and thus there are no arrows connecting CC and HC.

In the spirit of Kolmogorov distance measures, we follow Peddada et al. (2001) and define a test statistic for the graph in Figure 1 as the maximum distance between connected points. Specifically, our test is $max(D_1, D_2)$, where D_1 is the distance between the current control group and current dose groups, and D_2 is the distance between the historical control groups and current dose groups. Each of these two distances, D_1 and D_2, is itself the maximum of two quantities. Distance D_1 is the maximum of the Bieler-Williams version of the Poly-3 statistic and the order restricted inference statistic of Peddada et al. (2005) based on the current control group. Our choice for D_1 was motivated by a test proposed by Peddada and Kissling (2006), who selected this pair of statistics because the former performs well against linear trends, while the latter performs well when trends are not linear. By taking the maximum of the two, we obtain a test with reasonably good power for detecting linear or nonlinear (but still non-decreasing) trends. Distance D_2 has exactly the same form as D_1, except that it is based on the historical control groups rather than the current control group. Thus, the proposed test statistic is the maximum of four components. Two of the components are W_{BW} and W_{ISO}, as described in the previous subsection, and now we derive the other two analogous components, say W^*_{BW} and W^*_{ISO}.

Our approach views the lifetime tumor rates in the current and historical control groups (π_{c1}, $\pi_{h1}, \ldots, \pi_{hp}$) as random realizations from a common control population with mean π. If no animals die before the end of the study, the sample proportion y_{hm+}/n_{hm} has expected value π ($m = 1, 2, \ldots, p$) and we assume the variance can be expressed as

$$V(y_{hm+}/n_{hm}) = \sigma^2 \pi(1-\pi)/n_{hm}, \tag{2}$$

where σ^2 is an unknown parameter that allows for extra-binomial variation.

Now suppose that some animals die before t_{TS} and we want to account for this by using Poly-3 adjusted sample sizes. Allowing for the random nature of the sample size in variance formula (2), we propose separate estimators for σ^2 and $\pi(1 — \pi)$, where a Bieler-Williams

type adjustment is made to the latter. Under the assumption that the lifetime tumor rates are equal in all control and treated groups, we estimate σ^2 by $\widehat{\sigma}^2 = S_2^2 / \left| \tilde{\pi}\left(1-\tilde{\pi}\right) \right|$, where

$$S_2^2 = \frac{1}{p+k-1}\left[\sum_{m=1}^{p}\frac{\left(y_{hm+} - n_{hm}^{*}\,\tilde{\pi}\right)^2}{n_{hm}^{*}} + \sum_{i=1}^{k}\frac{\left(y_{ci+} - n_{ci}^{*}\,\tilde{\pi}\right)^2}{n_{ci}^{*}}\right]$$

and

$$\tilde{\pi} = \frac{y_{h++}+y_{c++}}{n_{h+}^{*}+n_{c+}^{*}}.$$

The Bieler-Williams estimate of $\pi(1 - \pi)$, say $\pi(\widehat{1-\pi})$, is obtained by applying the formula for S_1^2 to the combined set of p historical control groups and k current groups. We estimate the variance of the Poly-3 adjusted tumor proportions in the current high-dose group, $\tilde{\pi}_{ck}$, and in the pooled historical control groups, $\tilde{\pi}_h = y_{h++}/n_{h+}^{*}$, by

$$var\left(\tilde{\pi}_{ck}\right) = \frac{\widehat{\sigma}^2\pi(\widehat{1-\pi})}{w_{ck}}$$

and

$$var\left(\tilde{\pi}_{h}\right) = \frac{\widehat{\sigma}^2\pi(\widehat{1-\pi})}{w_{h+}},$$

respectively, where $w_{h+} = n_{h+}^{*2}/n_{h+}$. Note that although the extra binomial variation among the historical control groups should not extend to the current dose groups, we purposely inflate the estimated variance of $\tilde{\pi}_{ck}$ by $\widehat{\sigma}$ in the above formula to avoid the Behrens-Fisher problem. This modification leads to a slightly conservative test (described below), but we view it as a small price to pay for the resulting simplification.

Thus, we define a new pair of test statistics, say W_{BW}^{*} and W_{ISO}^{*}, which are analogous to W_{BW} and W_{ISO}, except that they are oriented toward detecting a trend in the current dose groups relative to the historical control groups rather than relative to the current control group. Specifically, define W_{BW}^{*} exactly the same as W_{BW}, except for replacing the current control group quantities w_{c1} and $\tilde{\pi}_{c1}$ with the historical control group summaries w_{h+} and $\tilde{\pi}_h$, as well as replacing S_1 with $S_1\widehat{\sigma}$. Similarly, let W_{ISO}^{*} be the following analog of W_{ISO}:

$$W_{ISO}^{*} = \frac{\check{\pi}_{ck} - \check{\pi}_{h}}{\widehat{\sigma}\sqrt{\pi(\widehat{1-\pi})\left(1/w_{h+}+1/w_{ck}\right)}},$$

where $\check{\pi}_h \le \check{\pi}_{c2} \le \check{\pi}_{c3} \le \cdots \le \check{\pi}_{ck}$ are the isotonized values of $\{\tilde{\pi}_h, \tilde{\pi}_{c2}, \tilde{\pi}_{c3}, \ldots, \tilde{\pi}_{ck}\}$ under the constraint that $\pi \le \pi_{c2} \le \pi_{c3} \cdots \pi_{ck}$, with weights $n_{h+}^{*}, n_{c2}^{*}, n_{c3}^{*}, \ldots, n_{ck}^{*}$.

Finally, the proposed survival-adjusted test statistic for detecting a dose-related trend in the tumor rates, using both current and historical controls, is given by

$$W=\max\left(W_{BW}, W_{ISO}, W^{*}_{BW}, W^{*}_{ISO}\right) .$$

Deriving the exact null distribution of W is not trivial. Hence we use critical values based on the approximations described below. In the next section, simulation studies reveal that our approximations control the Type I error rate very well.

Initially, assume that all animals survive to the end of the study and the lifetime tumor rates in all control and treated groups are equal to π. Also, let v_g denote the estimated variance of $\widehat{\pi}_g$ for $g = h, c1, c2, \ldots, ck$. These assumptions, together with the consistency of the $\{v_g\}$, suggest the following approximations:

$$\frac{\widehat{\pi}_g - \pi}{\sqrt{v_g}} \xrightarrow{dist} N(0,1) .$$

Denote the above limiting standard normal random variables by $\{Z_0, Z_1, \ldots, Z_k\}$, where the first term ($Z_0$) corresponds to the pooled set of p historical control groups and the remaining terms correspond to the k groups in the current study. Next, define $U_0=Z_0/\sqrt{p}$ and $U_i = Z_i$ for $i = 1, \ldots, k$. In our application, the observations in all control and treated groups are independent; thus, the $\{Z_i\}$ are independently distributed, as are the $\{U_i\}$. Note that all 4 components of W involve $\widehat{\pi}_{c2}, \ldots, \widehat{\pi}_{ck}$, but in addition to these terms W_{BW} and W_{ISO} are also functions of $\widehat{\pi}_{c1}$, whereas W^{*}_{BW} and W^{*}_{ISO} are also functions of $\widehat{\pi}_h$. Thus, we approximate the null distribution of W_{BW} and W_{ISO} by using the random variables $Z_1, Z_2, \ldots, Z_k$, and we approximate the null distribution of W^{*}_{BW} and W^{*}_{ISO} by using the random variables, $U_0, U_2, U_3, \ldots, U_k$. As W_{BW} can be obtained by regressing the lifetime tumor rates on dose, it should be distributed approximately the same as the regression of $Z_1, Z_2, \ldots, Z_k$ on dose:

$$W_{BW} \overset{dist}{\approx} \frac{\sum_{i=1}^{k}\left(d_i-\bar{d}\right)Z_i}{\sqrt{\sum_{i=1}^{k}\left(d_i-\bar{d}\right)^2}} .$$

We approximate the distribution of W^{*}_{BW} using the analogous function of $U_0, U_2, \ldots, U_k$:

$$W^{*}_{BW} \overset{dist}{\approx} \frac{\sum_{i=0,i\neq 1}^{k}\left(d_i-\bar{d}\right)U_i}{\sqrt{\sum_{i=0,i\neq 1}^{k}\left(d_i-\bar{d}\right)^2}} .$$

Both W_{ISO} and W^{*}_{ISO} resemble the statistic proposed by Williams (1977) to test for a monotone trend in dose. Hence, using the asymptotic approximations given above, the null distributions for these statistics can be approximated by

$$\frac{\widehat{Z}_k - \widehat{Z}_1}{\sqrt{2}}$$

and

$$\frac{\widehat{U}_k - \widehat{U}_0}{\sqrt{2}}.$$

respectively, where $\widehat{Z}_1 \le \widehat{Z}_2 \le \cdots \le \widehat{Z}_k$ are the isotonized values of $\{Z_1, Z_2, \ldots, Z_k\}$ with unit weights and $\widehat{U}_0 \le \widehat{U}_2 \le \cdots \le \widehat{U}_k$ are the isotonized values of $\{U_0, U_2, \ldots, U_k\}$ with a weight of p for U_0 and unit weights for the others.

Therefore, in the absence of deaths prior to the end of the experiment, the null distribution of our test statistic W can be approximated by the distribution of

$$V = \max\left(\frac{\sum_{i=1}^{k}\left(d_i - \bar{d}\right)Z_i}{\sqrt{\sum_{i=1}^{k}\left(d_i - \bar{d}\right)^2}}, \frac{\widehat{Z}_k - \widehat{Z}_1}{\sqrt{2}}, \frac{\sum_{i=0,i\neq 1}^{k}\left(d_i - \bar{d}\right)U_i}{\sqrt{\sum_{i=0,i\neq 1}^{k}\left(d_i - \bar{d}\right)^2}}, \frac{\widehat{U}_k - \widehat{U}_0}{\sqrt{2}} \right).$$

As V is a function of independent N(0, 1) random variables, its distribution can be simulated by repeatedly generating $k + 1$ independent standard normals. The level α critical value for W is the $100(1 - \alpha)^{th}$ percentile of the simulated distribution of V. In practice, we take a million random samples to approximate the distribution of W.

More generally, when some animals die early, the components of W are functions of the Poly-3 weights, which provide a simple survival adjustment. However, despite possible survival effects, we approximate the distribution of W by the same simulated distribution of V described above. We demonstrate that the proposed test performs well in the more general case by simulating a variety of realistic data sets with many deaths before t_{TS}. The simulation results show that our test operates at (or below) the nominal level in a wide range of situations typically observed in practice.

3. SIMULATION STUDY

3.1 Study Design

Following Peddada et al. (2005), data were simulated from a variety of realistic situations commonly encountered in NTP studies, which typically run for two years and end with a terminal sacrifice. We generated two latent random variables for each animal: T_1, the time to tumor onset, and T_2, the time to natural death. A simulated animal develops a tumor before death if and only if $T_1 < \min(T_2, t_{TS})$, where $\min(T_2, t_{TS})$ is the observed death time. Our simulation measured time in months, and thus $t_{TS} = 24$. Data were simulated for 50 animals per group, which is the standard sample size in most NTP 2-year bioassays. Poly-3 based tests assign unit weights to all animals that die with a tumor, regardless of the tumor's lethality. Thus, we did not simulate different levels of tumor lethality because such information is not used by the quantal response tests considered here.

We generated data for p historical control groups, a current control group, and three current dose groups. The latent times to tumor onset and death, T_1 and T_2, for animals in all groups were generated from independent Weibull distributions with survival functions of the form: $P(T > t) = exp(-\psi t^{\gamma})$. We assumed that all dose groups had the same shape parameters for T_1 and T_2, say γ_1 and γ_2, which govern the steepness of the tumor incidence and mortality curves, respectively. Any differences among groups were assumed to arise only through the scale parameters. Let the scale parameters for T_1 and T_2 be denoted by ψ_{1ci} and ψ_{2ci} in the i-th group in the current study ($i = 1, 2, 3, 4$) and by ψ_{1hm} and ψ_{2hm} in the m-th historical control group ($m = 1, \ldots, p$). We generated T_1 and T_2 values for the various groups according to the following models for the scale parameters:

$$\psi_{1ci}=\psi_1\phi_{1i}^{d_i}$$

and

$$\psi_{2ci}=\psi_2\phi_2^{d_i},$$

$$\psi_{1hm}=\psi_1|1+\tau X_m|$$

and

$$\psi_{2hm}=\psi_2,$$

where ψ_1 and ψ_2 are baseline values, ϕ_{1i} is specific to dose group i, ϕ_2 is common to all dose groups, and $X_m \overset{iid}{\sim} N(0, 1)$ for $m = 1, \ldots, p$. Thus, all control groups (current and historical) shared a common mortality scale parameter, ψ_2, but each had a distinct incidence scale parameter. The incidence scale parameters in the historical groups ($\psi_{1h1}, \psi_{1h2}, \ldots, \psi_{1hp}$) were perturbed from the current control value ($\psi_{1c1} = \psi_1$) in proportion to τ, to induce extra variation. Our primary focus is on the ratio of the tumor incidence rate in dose group i relative to the tumor incidence rate in the current control group, which is represented by $\theta_i=\phi_{1i}^{d_i}$.

The simulation specified multiple values for some parameters, which led to 540 configurations. We generated data for p = 5 historical control groups. Results are reported for a typical pattern of dose scores: $(d_1, d_2, d_3, d_4) = (0, 0.5, 1.0, 2.0)$, which we refer to as 2-fold spacings. We also investigated 5-fold dose spacings, but the results are very similar to those obtained for 2-fold spacings (if not even more favorable for our test), and thus they are not shown. As control death rates in NTP studies are well estimated, and our focus is on tumor incidence, very few mortality configurations were examined. The mortality shape parameter was fixed at $\gamma_2 = 5$ and we derived the baseline scale parameter value that gave a typical control survival rate of 70% at 2 years: $\psi_2 = 4.48 \times 10^{-8}$. Two mortality scale multipliers were used, $\phi_2 = 1$ and 1.5, which correspond to dose effects on mortality that can be labeled as “none” and “moderate” (i.e. 70% and 45% survival at 2 years in the high-dose group).

In contrast, we varied many factors influencing tumor incidence rates. We investigated three values for the shape of the tumor onset curve, $\gamma_1 = 1.5$, 3, and 6, and five values for the lifetime tumor rate in the current control group, $\pi_1 = 0.001$, 0.01, 0.05, 0.15, and 0.30. For

fixed values of π_1, γ_1, γ_2, and ψ_2, we used equation (1) to solve for ψ_1, the baseline incidence scale parameter for an "average" control group (i.e. $X_m = 0$). We examined three levels of heterogeneity among the historical control groups, where the heterogeneity parameter τ varied with the background tumor rate (π_1). The τ values were selected so that the resulting standard deviation of π_1 among the simulated historical control groups reasonably approximated low, medium, and high values within the range of sample standard deviations observed in the NTP historical control database. Finally, we chose ϕ_{i1} values that gave certain patterns for the incidence ratio, $\boldsymbol{\theta}' = (\theta_1, \theta_2, \theta_3, \theta_4)$. The null case for comparing incidence rates is $\boldsymbol{\theta}' = (1, 1, 1, 1)$, and we considered five non-null patterns, which varied with the background tumor rate. The specific values for ψ_1, τ, and $\boldsymbol{\theta}$ are given in Table 1.

Overall, we investigated 540 configurations by taking all combinations of several factors: mortality scale multiplier (2 levels), shape of incidence curve (3 levels), background tumor rate (5 levels), heterogeneity of control groups (3 levels), and incidence ratio pattern (6 levels). Type I error rates were evaluated for the 90 configurations associated with the null hypothesis of equal incidence rates, and powers were estimated for the 450 configurations associated with various alternative hypotheses. The operating characteristics of W and W_{NTP} were compared in each situation. Though initially it may seem odd to compare a test which uses historical control data with another test which does not, we view this as a comparison of a new method with the current standard method.

For each configuration, we generated 10,000 sets of current study data (i.e. a control group and three dose groups), and for each set of current study data, we generated a set of data from p historical control groups. We estimated the Type I error rates for W and W_{NTP} by the empirical proportions of the 10,000 trials for which H_0 was rejected. The nominal level for both trend tests was 0.05.

3.2 Results

This section summarizes the operating characteristics of the proposed test (W) and the NTP test (W_{NTP}) as estimated in our simulation study. Both tests maintained the nominal level (see Figure 2). The Type I error rate did not exceed 0.05 in any of the 90 null cases for W_{NTP} and only slightly exceeded the nominal level in one null case for W, and that single excess was not statistically significant. The Type I error rate generally increased with the background tumor rate (π_1), especially for W_{NTP}, as illustrated by the different symbols in Figure 2. The Type I error rates also varied with the shape of the incidence curve (γ_1) and the dose effect on mortality (ϕ_2), but these effects were small. Control heterogeneity (τ) had no impact on W_{NTP}, as the NTP test does not incorporate historical data. In contrast, the Type I error rates for W decreased as variability among historical control groups increased, though this effect was not strong for typical levels of heterogeneity, such as those used here.

The proposed test was more powerful than the standard test in almost all situations, with W having a fairly large advantage relative to W_{NTP} in many cases (see Figure 3). For some configurations, the two tests had approximately the same power, as shown by the symbols along the diagonal line. When the rejection rates differed, however, W tended to be more powerful than W_{NTP} , as indicated by the large proportion of symbols above the diagonal reference line. In fact, the proposed test was 4.5 times as powerful as the standard test in some cases and was never less than 94% as powerful. Analogous to the results for Type I error rates, the observed power tended to increase with background tumor rate, as illustrated by the different symbols in Figure 3. Similarly, power varied with the shape of the incidence curve and the dose effect on mortality, as well as the control group heterogeneity for W, but these effects usually were relatively small.

The proposed test performed well compared to the standard NTP test in general, but the most obvious improvement of W over W_{NTP} was when tumors were rare. For example, when the background tumor rate was 0.001 (i.e. a very rare tumor), W_{NTP} was extremely conservative, with Type I error rates close to 0 and powers between 3% and 9%. In contrast, W was less conservative than W_{NTP}, with Type I error rates as high as 2% and powers ranging from 5% to 26%. As for the other factors, the power of both tests decreased as differential mortality increased and as tumor development shifted to later in life (i.e. as γ_1 increased). In addition, the power of W decreased as the historical control groups became more heterogeneous. The effects of these other factors were relatively small, however, compared to the effect of the background tumor rate, at least at the factor levels reported here, which were chosen to simulate data representative of what the NTP observes in practice.

4. EXAMPLES

This section illustrates the proposed methodology with real data from two NTP studies. The first analysis focuses on a rare tumor, which is the situation that motivated our research, and the second analysis presents a contrasting example involving a common tumor. In one case, our results strengthen an NTP conclusion that was reached despite rather weak statistical evidence. The other case demonstrates how the increased power of our formal procedure for incorporating data from past studies provides an appealing alternative to the NTP's informal practice of comparing tumor rates in current dose groups with the range of rates from historical control groups.

4.1 Analysis of Benzophenone Data

Benzophenone is an aryl ketone which is produced in large quantities in the U.S., with the potential for widespread occupational and consumer exposures through its use as a fragrance enhancer, flavor additive, photoinitiator, and ultraviolet curing agent (NTP, 2006). It is also used in the manufacture of pharmaceuticals, insecticides, and agricultural chemicals, as well as being an additive in cosmetics, plastics, and adhesives. Benzophenone does not appear to be mutagenic (NTP, 2006, p. 19), although it could potentially be carcinogenic. Short-term animal studies suggested that the liver and the kidneys were the target organs, though toxicity also was observed in the hematopoietic system.

As a consequence of its widespread use, the NTP conducted a two-year study of male and female mice and rats exposed to benzophenone. This example focuses on histiocytic sarcomas in female rats. Groups of size 50 received doses of 0, 312, 625, or 1250 ppm of benzophenone in their food throughout the study. One rat in the 625 ppm group and two rats in the 1250 ppm group developed histiocytic sarcomas. Based only on the current study data, the NTP's Poly-3 trend test led to a significance value of $P = 0.074$. Though not statistically significant at the usual 0.05 level, the NTP concluded that there was equivocal evidence of carcinogenic activity of benzophenone in female rats for several reasons (NTP, 2006). First, histiocytic sarcoma is a rare neoplasm in female rats and had not been observed in any controls (0/460) in six recent NTP feed studies. In addition, the same neoplasm was positively correlated with the dose of benzophenone in the current study of female mice ($P = 0.032$), and there was a marginally significant trend ($P = 0.058$) observed for another hematopoietic system lesion (mononuclear leukemia) in female rats.

Using the proposed methodology, which formally accounts for the historical control data from the six recent feed studies, we find a significant dose-related trend ($P = 0.004$) in histiocytic sarcoma among female rats. Thus, our method uses historical information to bolster the weak statistical evidence which would otherwise be obtained if considering only

the concurrent data. Our results provide formal support for the biologic evidence that fueled the NTP's suspicion that benzophenone is carcinogenic in female rats.

4.2 Analysis of Gallium Arsenide Data

Gallium arsenide is a by-product of aluminum and zinc extraction (NTP, 2000). Due to its photovoltaic properties, this chemical is used extensively when producing lasers, photodetectors, light-emitting diodes, microwave devices, solar cells, and semicon-ductors. The most common exposure to gallium arsenide occurs when workers inhale small particles during the manufacturing of these devices. As in the case of benzophenone, gallium arsenide does not appear to be mutagenic (NTP, 2000, p. 27), even though it could potentially be carcinogenic. While little is known about the effects of gallium arsenide on humans, short-term rodent studies suggest that the lung is the primary target organ, with toxic effects also seen in the kidney, liver, immune system, and male reproductive system, as well as in fetal development and heme biosynthesis.

As a result of its extensive use in the microelectronics industry, the NTP conducted a two-year inhalation study of male and female mice and rats exposed to gallium arsenide (NTP, 2000). In this example, we focus on malignant lymphomas in female mice. Groups of 50 female mice were exposed in inhalation chambers to doses of 0, 0.1, 0.5, or 1.0 mg/m^3 of gallium arsenide. The corresponding numbers of mice with a malignant lymphoma were 3, 8, 11, and 10, respectively, and the NTP trend test produced a significance value of $P = 0.035$. Even if a statistically significant result is observed in the current study, the NTP often re-evaluates the significance of dose effects on tumor rates in the current study by comparing the current rates with the range of rates observed in control animals from past studies. For instance, in our gallium arsenide example, the survival-adjusted rates of malignant lymphoma in the control and three dose groups are 0.067, 0.175, 0.269, and 0.234, respectively. The NTP compared these rates with the background malignant lymphoma rates from 20 inhalation studies in their historical control database, which ranged from 0.06 to 0.32. As all of the lifetime tumor rates in the current study were within the range of historical rates, the NTP concluded that "the incidences of malignant lymphoma in exposed females were not considered to be exposure related" (NTP, 2000, p. 86), despite a statistically significant result for the current data at the usual 0.05 level of significance. It is easy to verify, however, that methods based on the range of historical tumor rates become less powerful as the number of historical studies increases.

As an alternative to the NTP strategy, we applied our test using 19 of the same 20 historical control groups (data from one study were unavailable). The significance value for our test was $P = 0.021$, which shows that even after incorporating historical control information, the observed dose effects remained statistically significant. Thus, this example suggests that the historical range may not be a good criterion for evaluating the statistical significance of dose effects in the current study.

We also assessed the robustness of our method to the choice of historical data by considering a variety of subsets of 5 historical studies. In each case, the resulting significance value was $P = 0.02$, with differences only in the third decimal place. Furthermore, the inclusion of historical control data from as few as 5 recent studies when analyzing the current data may be su cient, as the large sample approximation seems to work reasonably well even in that situation.

5. DISCUSSION

The present NTP approach to evaluating animal carcinogenicity experiments applies formal statistical tests only to concurrent data. Historical control data enter the evaluation only

informally – mainly in a comparison of the tumor rates in the current study to the range of background tumor rates among the historical controls. Incorporation of historical data into a formal statistical procedure would represent a much needed tool in evaluating results from cancer bioassays. In fact, the Technical Reports Review Subcommittee of the NTP Board of Scientific Counselors recently recommended that a new method be developed for this important problem (see http://ntp.niehs.nih.gov/files/TRRSMins0905.pdf). In this paper we propose such a procedure, which formally incorporates historical control data, adjusts for survival, makes no lethality assumptions, and avoids parametric models.

We want to re-emphasize that our focus is on age-specific tumor incidence, even though this endpoint generally can not be estimated directly. The proposed test and the standard NTP test are most naturally oriented toward comparisons of lifetime rates, but we evaluate their performance with respect to hypotheses about age-specific incidence rates. Both tests use the Poly-3 adjustment to account for dose-related survival diff erences, and structurally they are expressed in terms of survival-adjusted versions of lifetime tumor rates $\{\pi_i\}$ rather than age-specific tumor incidence rates $\{\lambda_i(t)\}$. Nevertheless, we conducted extensive simulations to characterize the rejection rates of these two tests with respect to hypotheses about the $\{\lambda_i(t)\}$, where we considered the null configurations to be those having equal age-specific incidence rates rather than those having equal lifetime tumor rates.

As noted in Haseman (1992), extra-binomial variation among historical controls can arise from calendar time drift, inter-laboratory variability, and different types of controls. For example, the variability of tumor rates for vehicle controls in inhalation studies often differs from that of control tumor rates in feed studies. In order to minimize extra variability, the current NTP practice is to use the most recent historical control data and also, as far as possible, to limit the data to studies involving the same route of exposure. We believe that the NTP should continue to base their decisions about carcinogenicity on such a strategy of selecting relevant historical control data, as it is not only scientifically meaningful but will also increase the power of our test considerably. Too much variability in the historical control data can result in W having less power than W_{NTP} in certain extreme situations.

We believe that the proposed test should be used whenever historical control data are incorporated in the analysis. Specifically, the informal range-based method presently used by the NTP should be replaced by a formal statistical test such as W. In most practical situations, W is at least as powerful as W_{NTP} and often is much more powerful. Even when W has no power advantage over W_{NTP}, the former still represents a more realistic approach to understanding the data than the latter. We note, however, that although on a relative scale our test enjoys its greatest power advantage over the NTP test when the tumor of interest is rare, there is still room to improve the absolute power of our test in this situation.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The authors thank Dr. Kenny Crump, Dr. Beth Gladen, Dr. Joseph Haseman, the referees, and the editors for their helpful comments and suggestions. The authors also thank Mr. Kevin McGowan for extracting the NTP data from the database which were analyzed in this paper.

REFERENCES

Armitage P. Tests for Linear Trends in Proportions and Frequencies. Biometrics 1955;11:375–386.

Bailer A, Portier C. Effects of Treatment-induced Mortality and Tumor-induced Mortality on Tests for Carcinogenicity in Small Samples. Biometrics 1988;44:417–431. [PubMed: 3390507]

Barlow, R.; Bartholomew, D.; Bremner, J.; Brunk, H. Statistical Inference Under Order Restrictions. John Wiley & Sons; London: 1972.

Bieler G, Williams R. Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity. Biometrics 1993;49:793–801. [PubMed: 8241374]

Cochran W. Some Methods for Strengthening the Common χ^2 Tests. Biometrics 1954;10:417–451.

Dempster AP, Selwyn MR, Weeks BJ. Combining Historical and Randomized Controls for Assessing Trends in Proportions. Journal of the American Statistical Association 1983;87:221–227.

Dunson D, Dinse G. Bayesian Incidence Analysis of Animal Tumorigenicity Data. Applied Statistics 2001;50:125–141.

French JL, Ibrahim JG. Bayesian Methods for a Three-state Model for Rodent Carcinogenicity Studies. Biometrics 2002;58:906–916. [PubMed: 12495145]

Greim H, Gelbke H-P, Reuter U, Thielmann HW, Edler L. Evaluation of Historical Control Data in Carcinogenicity Studies. Human and Experimental Toxicology 2003;22:541–549. [PubMed: 14655720]

Haseman JK. Value of Historical Controls in the Interpretation of Rodent Tumor Data. Drug Information Journal 1992;26:191–200.

Haseman JK. Data Analysis: Statistical Analysis and Use of Historical Control Data. Regulatory Toxicology and Pharmacology 1995;21:52–59. [PubMed: 7784636]

Haseman JK, Hu J, Boorman GA. Use of Historical Control Data in Carcinogenicity Studies in Rodents. Toxicologic Pathology 1984;12:126–135. [PubMed: 11478313]

Hoel DG, Yanagawa T. Incorporating Historical Controls in Testing for a Trend in Proportions. Journal of the American Statistical Association 1986;81:1095–1099.

Ibrahim JG, Ryan LM. Use of Historical Controls in Time-adjusted Trend Tests for Carcinogenicity. Biometrics 1996;52:1478–1485. [PubMed: 8962464]

Ibrahim JG, Ryan LM, Chen M-H. Using Historical Controls to Adjust for Covariates in Trend Tests for Binary Data. Journal of the American Statistical Association 1998;93:1282–1293.

Makuch RW, Stephens MA, Escobar M. Generalised Binomial Models to Examine the Historical Control Assumption in Active Control Equivalence Studies. The Statistician 1989;38:61–70.

Mancuso J, Ahn H, Chen J, Mancuso J. Age-adjusted Exact Trend Tests in the Event of Rare Occurrences. Biometrics 2002;58:403–412. [PubMed: 12071414]

McKnight B, Crowley J. Tests for Differences in Tumor Incidence Based on Animal Carcinogenesis Experiments. Journal of the American Statistical Association 1984;79:639–648.

National Toxicology Program. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health; RTP, NC: 2000. NTP Technical Report on the Toxicology and Carcinogenesis Studies of Gallium Arsenide (CAS No. 1303-00-0) in *F*344/*N* Rats and *B*6*C*3F_1 Mice (Inhalation Studies). Technical Report Series No. 492NIH Publication No. 00-3951

National Toxicology Program. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health; RTP, NC: 2006. NTP Technical Report on the Toxicology and Carcinogenesis Studies of Benzophenone (CAS No. 119-61-9) in *F*344/*N* Rats and *B*6*C*3F_1 Mice (Feed Studies). Technical Report Series No. 533NIH Publication No. 05-4469

Parise H, Wand MP, Ruppert D, Ryan L. Incorporation of Historical Controls Using Semiparametric Mixed Models. Applied Statistics 2001;50:31–42.

Peddada S, Dinse G, Haseman J. A Survival-adjusted Quantal Response Test for Comparing Tumor Incidence Rates. Applied Statistics 2005;54:51–61.

Peddada S, Kissling G. A Survival-adjusted Quantal-Response Test for Analysis of Tumor Incidence Rates in Animal Carcinogenicity Studies. Environmental Health Perspectives 2006;114:537–541. [PubMed: 16581542]

Peddada S, Prescott K, Conaway M. Tests for Order Restrictions in Binary Data. Biometrics 2001;57:1219–1227. [PubMed: 11764263]

Prentice RL, Smythe RT, Krewski D, Mason M. On the Use of Historical Control Data to Estimate Dose Response Trends in Quantal Bioassay. Biometrics 1992;48:459–478. [PubMed: 1637972]

Ryan L. Using Historical Controls in the Analysis of Developmental Toxicity Data. Biometrics 1993;49:1126–1135. [PubMed: 8117906]

Seewald W. Time Trend in Historical Controls for Tumour Incidences in Long-term Animal Studies. Applied Statistics 1994;43:127–137.

Tamura RN, Young SS. The Incorporation of Historical Control Information in Tests of Proportions: Simulation Study of Tarone's Procedure. Biometrics 1986;42:343–349. [PubMed: 3755626]

Tarone RE. The Use of Historical Control Information in Testing for a Trend in Proportions. Biometrics 1982;38:214–220.

Williams D. Some Inference Procedures for Monotonically Ordered Normal Means. Biometrika 1977;64:9–14.

Yanagawa T, Hoel DG. Use of Historical Controls for Animal Experiments. Environmental Health Perspectives 1985;63:217–224. [PubMed: 4076086]

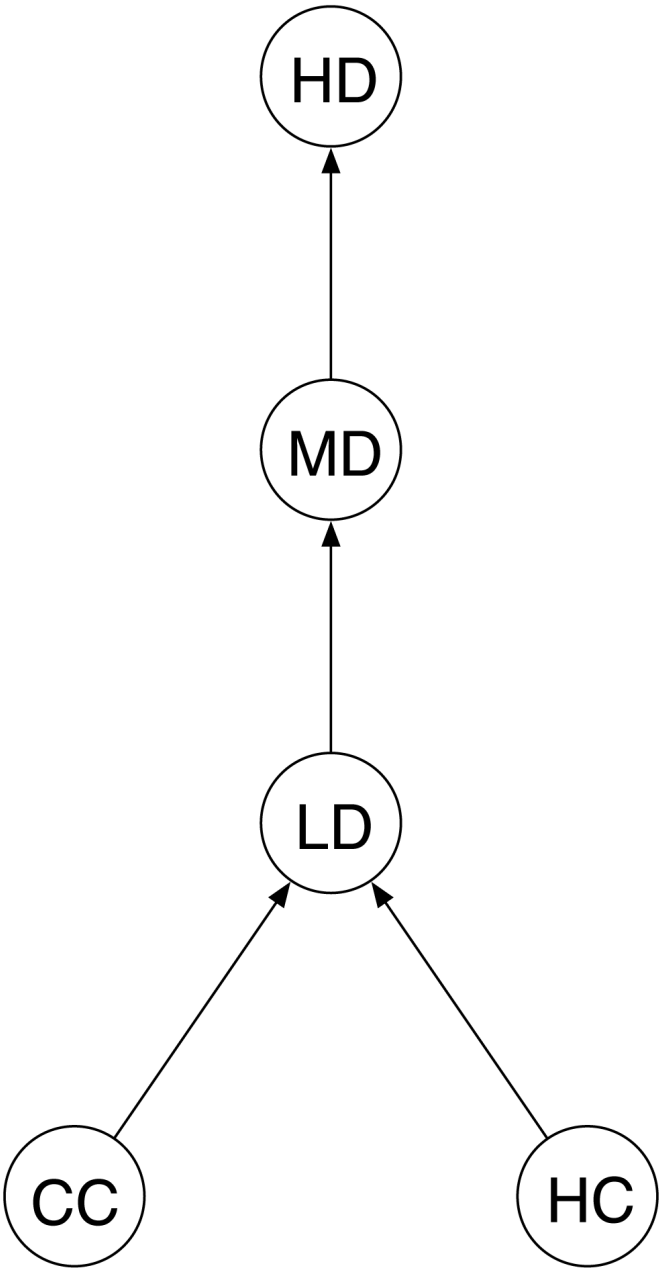

Figure 1.
Directed Graph Representing the Trends of Interest. The circles denote parameters in the historical control (HC) group, the current control (CC) group, the current low-dose (LD) group, the current mid-dose (MD) group, and the current high-dose (HD) group. The arrow between a pair of circles points toward the larger parameter.

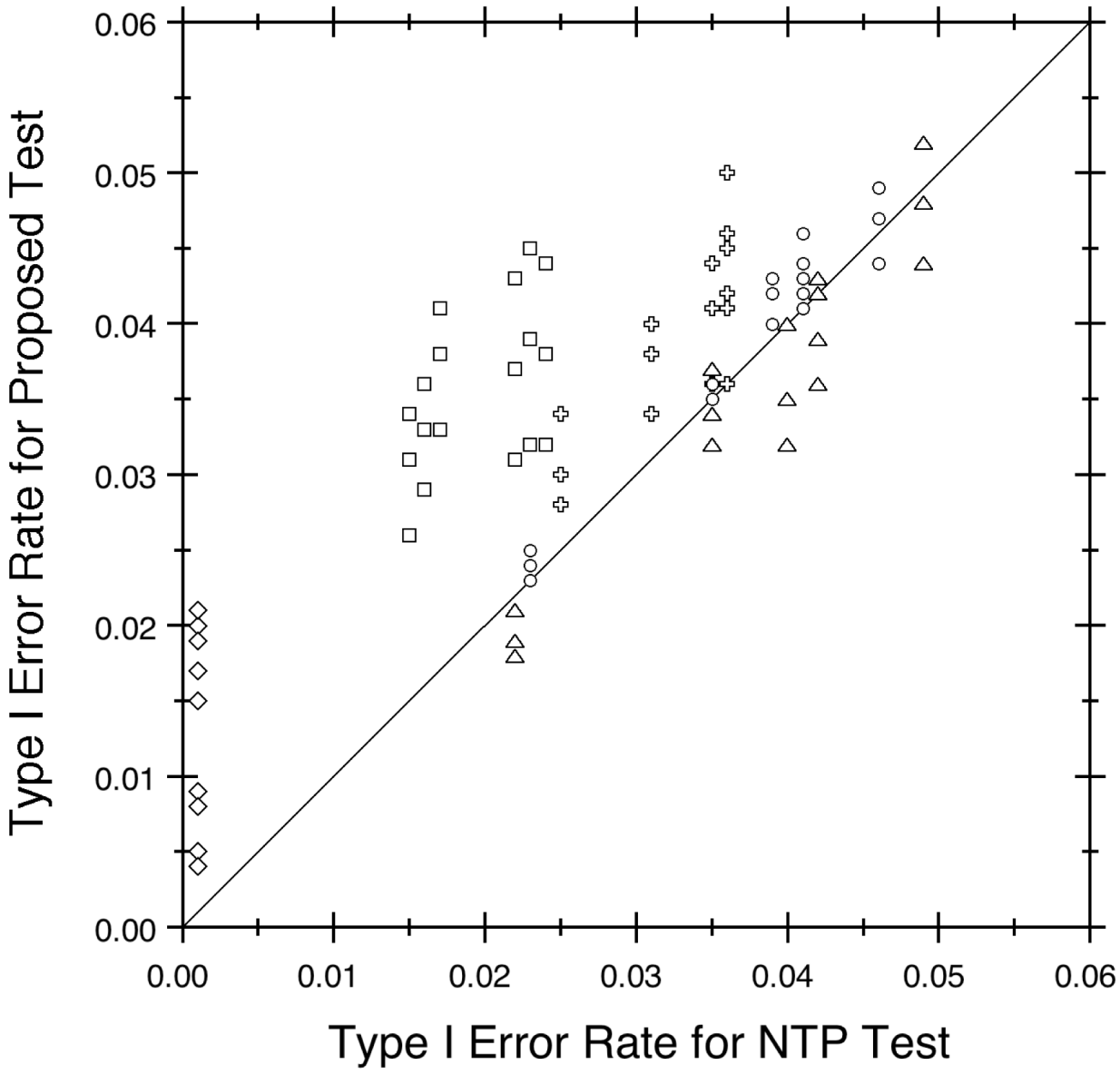

Figure 2.
Type I Error Rate Comparison for Proposed Test (W) Versus NTP Test (W_{NTP}). Different symbols identify the various background tumor rates: 0.001 (diamonds), 0.01 (squares), 0.05 (pluses), 0.15 (circles), and 0.30 (triangles). The diagonal reference line shows where the two trend tests have equal Type I error rates. The nominal level is 0.05.

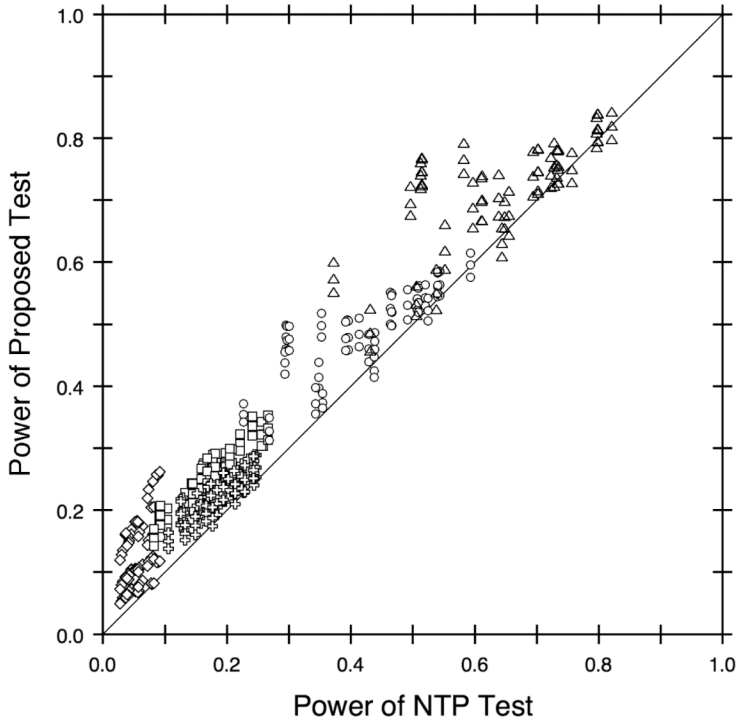

Figure 3.
Power Comparison for Proposed Test (W) Versus NTP Test (W_{NTP}). Different symbols identify the various background tumor rates: 0.001 (diamonds), 0.01 (squares), 0.05 (pluses), 0.15 (circles), and 0.30 (triangles). The diagonal reference line shows where the two trend tests have equal power. The nominal level is 0.05.

Table 1

Parameter Values Used in the Simulations. The entries in the last 5 columns correspond to the 5 values of the background tumor rate (π_1). The entries in the first 3 rows are the values of the baseline incidence scale parameter (ψ_1) for each of the 3 values of the incidence shape parameter (γ_1). The entries in the middle 3 rows are the values of the historical control heterogeneity parameter (τ) for each of the 3 levels of heterogeneity. The entries in the last 5 rows are the values in the vector of incidence ratios ($\boldsymbol{\theta}$) for each of the 5 patterns of incidence ratios across dose.

		Background Tumor Rate (π_1)				
		0.001	**0.01**	**0.05**	**0.15**	**0.30**
		Value of Baseline Incidence Scale Parameter (ψ1)				
Incidence Shape (γ_1)	1.5	9.00×10^{-6}	9.00×10^{-5}	4.70×10^{-4}	1.50×10^{-3}	3.30×10^{-3}
	3.0	8.00×10^{-8}	8.00×10^{-7}	4.20×10^{-6}	1.34×10^{-5}	2.97×10^{-5}
	6.0	6.50×10^{-12}	6.50×10^{-11}	3.25×10^{-10}	1.04×10^{-9}	2.32×10^{-9}
		Value of Historical Control Heterogeneity Parameter (τ)				
Level of Heterogeneity	low	5.0	1.50	0.75	0.3	0.2
	med	10.0	1.75	1.00	0.4	0.3
	high	15.0	2.00	1.25	0.5	0.4
		Vector of Dose-specific Incidence Ratios ($\boldsymbol{\theta}$)				
Pattern 1		(1,1,1,10)	(1,1,1,4)	(1,1,1,2)	(1,1,1,2)	(1,1,1,2)
Pattern 2		(1,1,10,10)	(1,1,4,4)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)
Pattern 3		(1,10,10,10)	(1,4,4,4)	(1,2,2,2)	(1,2,2,2)	(1,2,2,2)
Pattern 4		(1,5,5,10)	(1,2,2,4)	(1,1.5,1.5,2)	(1,1.5,1.5,2)	(1,1.5,1.5,2)
Pattern 5		(1,5,10,15)	(1,2,3,4)	(1,1.25,1.75,2)	(1,1.25,1.75,2)	(1,1.25,1.75,2)