Published in final edited form as:
Stat Biopharm Res. 2011 February 1; 3(1): 97–105. doi:10.1198/sbr.2010.09044.

Comparing tumor rates in current and historical control groups in rodent cancer bioassays

Gregg E. Dinse and **Shyamal D. Peddada**[*]
Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

Abstract

When evaluating carcinogenicity, tumor rates from the current study are informally assessed within the context of relevant historical control tumor rates. Current rates outside the range of historical rates raise concerns. We propose a statistical procedure that formally compares tumor rates in current and historical control groups. We use a normal approximation for the null distribution of the proposed test when there are at least 5 historical control groups and the average tumor rate is above 0.5%; otherwise, we apply standard bootstrap techniques. For comparison purposes, we show that formally basing decisions on the range of historical control rates would yield unusually high false positive rates. That is, a range-based decision rule would not maintain the nominal 5% significance level and could produce Type I error rates as high as 67%. In other cases, the power could go to zero. The proposed test, however, controls Type I errors while adjusting for survival and extra variability among the historical studies. We illustrate the methods with data from a study of benzophenone. Compared to a range-based decision rule, the proposed test has several important advantages, including operating at the specified level and being applicable with as few as one historical study.

Keywords

Bootstrap; Carcinogenicity study; Extra variation; Historical range; National Toxicology Program (NTP); Poly-3 survival adjustment; Quantal response

1. INTRODUCTION

To safeguard public health, government and industry researchers routinely conduct rodent cancer bioassays to evaluate the carcinogenicity of chemicals to which humans are exposed. These bioassays involve several treated groups, where animals are exposed to various doses of the test chemical, and a control group, where animals are not exposed but may receive the “vehicle” used to administer the chemical to the treated groups. The information obtained from these control groups provides a rich collection of historical data. For example, since its formation in 1977 as mandated by the U.S. Congress, the National Toxicology Program (NTP) has evaluated over 500 chemicals via 2-year rodent cancer bioassays, thus generating a vast historical control database.

Researchers consider such databases when assessing a chemical’s carcinogenicity in the current experiment. Not only do they evaluate whether tumor incidence increases in treated animals relative to controls in the current study, but they also compare tumor rates in the current dose groups with rates in the historical control database. Although the dose response

[*]corresponding author (peddada@niehs.nih.gov).

in the current study is typically assessed with a formal statistical procedure (see, *e.g.*, Bailer and Portier [1]), comparisons of the current study with the historical database are usually informal. Specifically, if tumor rates in the treated groups fall within the range of rates in a relevant subset of the historical control data, the effect of the chemical may be discounted. In contrast, rates in the treated groups that fall outside the historical range may be considered evidence of a real carcinogenic effect. Moreover, if the current control rate falls outside the historical range, there may be concern about whether the current control group (and study) are consistent with, and comparable to, the previous control groups (and studies).

The proper use of historical control data has been the subject of much discussion among toxicologists and pathologists, including a town hall meeting in June 2008 conducted by the Society of Toxicologic Pathology, an invited “best practices” paper [2], and a paper on developing a “global database” so that control data from various laboratories around the world can be brought together for shared use [3]. These important developments require statisticians to evaluate the common practice of using the historical range to informally assess results from the current study and to derive a formal methodology that is more appropriate and yet simple enough to be adopted for practical use.

For purposes of motivation and illustration, consider the NTP 2-year study of benzophenone (see http://ntp.niehs.nih.gov/go/16328). Among female rats, the rates of mononuclear cell leukemia (MCL) in control, low-, mid-, and high-dose groups were 19/50 (38%), 25/50 (50%), 30/50 (60%), and 29/50 (58%), respectively. Despite the suggestion of a positive trend in MCL rates with dose, the NTP’s trend test was not statistically significant (p = 0.058) at the usual 0.05 level. The NTP noted that MCL rates among controls in 6 recent studies ranged from 12% to 35%, which strengthens the evidence of a trend. As the MCL rate in the current control group (38%) was not within this historical range, however, one might question whether the current study was comparable to these 6 previous studies.

Elmore and Peddada [4] discussed drawbacks of using a historical range of tumor rates to evaluate current experimental data. Their main point was that outliers in the historical data can inflate the range, thus yielding a procedure with little power to detect group differences. Ironically, in the absence of outliers, if one were to use the range of historical control tumor rates to test the null hypothesis of equal tumor rates among one current and k historical control groups, the false positive rate could be as large as $2/(k + 1)$, which varies from 0 (for $k = \infty$) to 0.67 (for $k = 2$). For instance, when $k = 6$, as in the benzophenone example, the Type I error rate could be over 28%. Intrinsically, the range is not designed to control Type I or Type II errors. Thus, although the historical range is a widely used supplemental tool among toxicologists and pathologists, it yields an arbitrary decision rule.

Peddada et al [5] developed a method based on order-restricted inference to evaluate the dose-response in the current study while formally incorporating historical control tumor data. Toxicologists, however, expressed an interest in also comparing the tumor rate in the current control group with that among the historical control groups. We address this concern by developing a simple test for comparing tumor rates in the current and historical control groups. The proposed approach employs the poly-3 survival adjustment [1], a well accepted technique used in long-term carcinogenicity testing to adjust for differences in mortality. Our procedure also accounts for variability within and between studies. Extensive simulations show that our test operates at approximately the nominal level, whereas a pair of decision rules based on the historical range do not. We illustrate these methods with the MCL data from the NTP benzophenone study.

2. STATISTICAL METHODS

2.1. Proposed test

Let n_i be the number of animals in the i^{th} control group and let y_{ij} denote the tumor status of the j^{th} animal in the i^{th} group ($j = 1, \dots, n_i$; $i = 1, \dots, k + 1$), where y_{ij} is 1 if a tumor is present at necropsy and 0 otherwise. Set $y_{i+}=\sum_{j=1}^{n_i} y_{ij}$, which is the number of tumor-bearing animals in the i^{th} group. Let π_i be the tumor rate in the i^{th} of k historical control groups, which are assumed to come from a population of control groups with mean tumor rate π^h. Denote the tumor rate in the current control group by $\pi_{k+1} = \pi^c$. Unless all animals live to the terminal sacrifice time, say t_{TS}, the sample proportion, $\tilde{\pi}_i = y_{i+}/n_i$, tends to underestimate the true tumor rate π_i ($i = 1, \dots, k + 1$). Thus, we use the poly-3 survival adjustment [1] to compensate for the reduced time at risk associated with early deaths. Let t_{ij} denote the death time for the j^{th} animal in the i^{th} group and set $\delta_{ij} = y_{ij} +(1 - y_{ij})(t_{ij}/t_{TS})^3$.

The poly-3 survival-adjusted sample size for the i^{th} group is $n_i^*=\sum_{j=1}^{n_i} \delta_{ij}$. An animal that dies with a tumor receives a full weight of 1, as does an animal that survives to t_{TS}, but otherwise δ_{ij} is a fractional weight that equals the cube of the proportion of the study the animal survived. The corresponding poly-3 survival-adjusted estimator for π_i is $\widehat{\pi}_i=y_{i+}/n_i^*$. Accordingly, we estimate the tumor rates in the current and historical control groups by $\widehat{\pi}^c=\widehat{\pi}_{k+1}=y_{k+1,+}/n_{k+1}^*$ and $\widehat{\pi}^h=\sum_{i=1}^{k} n_i^*\widehat{\pi}_i/\sum_{i=1}^{k} n_i^*=\sum_{i=1}^{k} y_{i+}/\sum_{i=1}^{k} n_i^*$, respectively.

Our goal is to test the null hypothesis H_0: $\pi^c = \pi^h$ against the alternative hypothesis H_a: $\pi^c \neq \pi^h$. Under H_0, the current and historical control groups have the same mean tumor rate, say $\pi^c = \pi^h = \pi$. If no animals die before the end of the study, then $n_i^*=n_i$ and each $\hat{\pi}_i$ reduces to $\tilde{\pi}_i$ and has expected value π ($i = 1, \dots, k + 1$). Otherwise, we substitute n_i^* for n_i and, unlike the standard binomial proportions problem, we must account for the fact that the sample sizes are random variables when calculating the variances of $\hat{\pi}^c$ and $\hat{\pi}^h$. This issue was addressed by Bieler and Williams [6]; their variance estimator is widely used in this context and we apply it here. Following the approach of Peddada et al [5], we allow the variance of $\hat{\pi}^h$ to have two components; namely, the variability within each historical control group and also the variability between historical control groups.

We propose testing H_0 against H_a with the following Wald-type statistic:

$$Q=\frac{\widehat{\pi}^h - \widehat{\pi}^c}{\sqrt{\widehat{\pi(1-\pi)}(\widehat{\sigma}^2/w^h+1/w^c)}}.$$

Under the null hypothesis, the Bieler-Williams estimator for within-studies variation is

$$\widehat{\pi(1-\pi)}=\sum_{i=1}^{k+1}\sum_{j=1}^{n_i}(r_{ij} - \overline{r}_i)^2/\sum_{i=1}^{k+1}(n_i - 1),$$

where $r_{ij} = y_{ij} - \delta_{ij}\hat{\pi}$, $\overline{r}_i=\sum_{j=1}^{n_i} r_{ij}/n_i$, $\widehat{\pi}=\sum_{i=1}^{k+1} y_{i+}/\sum_{i=1}^{k+1} n_i^*$, $w^h=\sum_{i=1}^{k}(n_i^*)^2/n_i$, and $w^c=(n_{k+1}^*)^2/n_{k+1}$. For $k = 1$, we set $\hat{\sigma}^2 = 1$; otherwise, we estimate the between-studies variation by

$$\widehat{\sigma}^2=\frac{\frac{1}{k-1}\sum_{i=1}^{k}(y_{i+}-n_i^*\widehat{\pi}^h)^2/n_i^*}{\widehat{\pi}^h(1-\widehat{\pi}^h)}.$$

We approximate the null distribution of Q by the standard normal distribution and often we can test H_0 by comparing the observed value of Q to the percentage points of the standard normal distribution. However, the approximation does not work well in certain extreme situations. Thus, for $k < 5$ or $\hat{\pi}^h \leq 0.005$, we derive the null distribution of Q using a standard nonparametric bootstrap methodology along the lines of Peddada et al [7]. In this case, we assume that all $k + 1$ control groups are homogeneous under H_0 and that any differences are strictly due to random variation in the data. From the pooled collection of all observed pairs $\{(y_{ij}, t_{ij}): j = 1, \ldots, n_i; i = 1, \ldots, k + 1\}$, we select a simple random sample (with replacement) of n_i pairs and assign them to the i^{th} group for each $i = 1, \ldots, k + 1$. We repeat this procedure B times and for the b^{th} of these bootstrap samples, we calculate the statistic Q and denote it $Q_b^*\ (b=1,\ldots,B)$. The bootstrap estimate of the p-value associated with the observed data is $\widehat{p}_B=\#(Q_b^*\geq Q_{obs})/B$, where Q_{obs} is the test statistic based on the observed data. Since $\hat{p}_B$ depends on the number of bootstrap samples, we recommend using a large value for B to reduce the variability in the estimated p-values. Our simulation studies indicate that $B = 10,000$ is sufficiently large.

Extensive simulations, reported in the next section, demonstrate that the proposed test operates at approximately the nominal level across a wide range of realistic situations.

2.2. Range-based decision rule

Tumor rates from the current study are often evaluated, at least informally, in the context of the range of historical rates; for a discussion, see Keenan et al [2]. To assess this approach, we define a decision rule R, based on the range of unadjusted rates among the historical control groups, which rejects H_0 if $\tilde{\pi}_{k+1} < min(\tilde{\pi}_1, \ldots, \tilde{\pi}_k)$ or $\tilde{\pi}_{k+1} > max(\tilde{\pi}_1, \ldots, \tilde{\pi}_k)$. In addition, define a similar decision rule R^*, based on the poly-3 survival-adjusted rates, which rejects H_0 if $\hat{\pi}_{k+1} < min(\hat{\pi}_1, \ldots, \hat{\pi}_k)$ or $\hat{\pi}_{k+1} > max(\hat{\pi}_1, \ldots, \hat{\pi}_k)$. In the absence of ties among the tumor rates, each of the k+1 control groups is equally likely to have the smallest (or largest) tumor rate under H_0, and thus the probability of a Type I error for a 2-sided range-based decision rule is $2/(k + 1)$. As explained later, ties are more likely for R than R^*. Also, the number of ties increases as the true tumor rate approaches 0 or 1.

3. SIMULATION RESULTS

3.1. Simulation study design

Data were simulated from a variety of situations typically encountered in the 2-year NTP bioassays. We generated two latent variables for each animal: T_1, the time to tumor onset, and T_2, the time to natural death. A simulated animal developed a tumor before death if $T_1 < \min(T_2, t_{TS})$, where $\min(T_2, t_{TS})$ is the observed death time. Time was measured in months, and thus $t_{TS} = 24$. Data were simulated for 50 animals per group, which is the standard sample size in most NTP long-term bioassays.

We generated data for k historical control groups and one current control group. For each group, latent times to tumor onset and natural death, T_1 and T_2, were generated from a pair of independent Weibull distributions with survival functions of the form: $P(T > t) = exp(-\psi t^{\gamma})$. The tests are not affected by tumor lethality, so there was no need to consider dependent T_1 and T_2. We assumed the shape parameters for T_1 and T_2, say $\gamma_1 > 0$ and $\gamma_2 > 0$, which determine the steepness of the tumor incidence and mortality curves, did not vary across the $k + 1$ groups. Any differences among groups were assumed to arise only through the scale parameters for T_1 and T_2, say $\psi_{1i} > 0$ and $\psi_{2i} > 0$, which are inversely proportional to the mean times to tumor onset and natural death, respectively. We introduced extra variability by modeling ψ_{hi} as $\psi_h e^{\tau Z}$, where $\psi_h > 0$ is a baseline scale parameter, $\tau \geq 0$ is a heterogeneity parameter, and Z is a N(0, 1) random variable truncated on the (−2,2) interval ($h = 1, 2$; $i = 1, \ldots, k + 1$). Therefore, each control group had distinct incidence and mortality scale parameters that varied about baseline values ψ_1 and ψ_2, respectively, as functions of τ. We refer to the control groups as being homogeneous for $\tau = 0$ (i.e., $\psi_{1i} \equiv \psi_1$ and $\psi_{2i} \equiv \psi_2$) and as having extra variation (or being heterogeneous) for $\tau > 0$.

Our simulation study investigated 288 configurations by taking all combinations of five factors: number of historical control groups (4 levels), shape of the incidence curve (3 levels), mean historical control tumor rate (4 levels), heterogeneity of the control groups (2 levels), and difference between the current and historical rates (3 levels). As control death rates in NTP studies are well estimated and our focus is on tumor rates, we used a single baseline mortality distribution in all simulations. The mortality shape parameter and baseline scale parameter were fixed at $\gamma_2 = 5$ and $\psi_2 = 4.48 \times 10^{-8}$, which produce an average survival rate of 70% at 2 years, a common value in NTP long-term studies.

In contrast, we varied several factors influencing tumor rates. We considered three values for the shape of the incidence curve ($\gamma_1 = 1.5$, 3, and 6), ranging from early-onset to late-onset tumors, and four values for the mean tumor rate among the historical control groups ($\pi^h = 0.01$, 0.05, 0.15, and 0.30), ranging from rare to common tumors. For given values of γ_1, γ_2, and ψ_2, we used equation (1) in Peddada et al [5] to determine what value of ψ_1, the baseline incidence scale parameter for an “average” control group ($Z = 0$), yields the desired value of π^h. We examined the homogeneous case ($\tau = 0$) and a heterogenous case based on a value of τ that made the variance of the tumor rates 20% larger than in the homogeneous case. We investigated three differences between π^c and π^h: none, small, and large. Small and large differences corresponded to values of π^c that were 25% and 50% larger than π^h, respectively, on the log scale. Finally, we generated data for $k = 1$, 2, 5, and 10 historical control groups. Table 1 gives the values of ψ_1 and τ used in the simulation.

For each of the 96 null configurations, where $\pi^c = \pi^h = \pi$, we generated 10,000 sets of data for one current and k historical control groups. We estimated the Type I error rates for Q, R and R^* by the empirical proportions of the 10,000 trials for which H_0 was rejected at the nominal 0.05 level. With respect to our test, Q, if $k < 5$ or $\hat{\pi}^h \leq 0.005$, we generated $B = 10{,}000$ bootstrap samples for each set of observed data and rejected H_0 if $\hat{p}_B < 0.05$. Otherwise, we used a normal approximation and rejected H_0 if $|Q| > 1.96$.

3.2. Type I error rates for range-based decision rules

The decisions based on the historical range of tumor rates performed poorly. This was true whether using unadjusted rates or poly-3 survival-adjusted rates. For $k = 2$, the simulated Type I error rates varied from 27.9% to 58.7% for R and 36.9% to 67.0% for R^* (Table 2). Even for $k = 10$, the Type I error rates were as high as 14.5% for R and 18.5% for R^*. These error rates are unacceptably high. In contrast, the Type I error rates can become vanishingly small for very large k, yielding extremely conservative procedures (results not shown).

If H_0 is true but the observed tumor rates are distinct (i.e., no ties), the Type I error rate for a decision rule that rejects H_0 if the current control tumor rate falls outside the historical range is $2/(k + 1)$, which can differ greatly from the usual 5% significance level, depending on the value of k. This formula is derived from the fact that under the null hypothesis of no differences among the $k + 1$ control groups, any group is equally likely to have the smallest (or largest) tumor rate. If multiple tumor rates coincide with the minimum or maximum, the use of strict inequalities in the definitions of R and R^* can produce Type I error rates below $2/(k + 1)$, unless ties are broken randomly (which we did not do).

The propensity for ties varies with two factors. Tumor rates are ratios of tumor counts and sample sizes, where the latter can be adjusted for survival effects or not. Differences among tumor rates can arise from differences among numerators, denominators, or both. The probability of observing tied tumor counts is lowest for a tumor rate of 0.5 and increases as the tumor rate approaches 0 or 1. Also, exact matches are more likely among unadjusted rates, where denominators are always integers, than among survival-adjusted rates, where denominators are typically not integers. Thus, as predicted, our simulation study obtained Type I error rates nearly identical to $2/(k +1)$ when the range-based decision rule was based on poly-3 survival-adjusted tumor rates (R^*) and the true tumor rate (π) was nearest one-half. See the bottom portion of Table 2, where the simulated Type I error rates are close to 66.7% for $k = 2$, 33.3% for $k = 5$, and 18.2% for $k = 10$. The false positive rate decreased from the predicted value of $2/(k + 1)$ when the range-based decision rule used unadjusted tumor rates (R) or when the granularity in incidence rates increased for rarer tumors, either of which tended to create a greater number of ties among the observed tumor rates.

3.3. Type I error rates for proposed test

The proposed test performed very well. In contrast to the range-based rules, Q maintained the nominal 5% level in all situations. The worst (i.e., highest) Type I error rate out of 96 null scenarios was 5.8% (Table 3). On the other hand, the most conservative results were obtained for $\pi = 0.01$ and $k = 1$, where the false positive rate varied from 2.2% to 2.4%. Our simulation studies show that, unless the tumor is extremely rare and only one historical control group is available, the Type I error rates for the proposed test are very reasonable.

We also note that neither the shape of the incidence curve nor the introduction of extra variability had any noticeable impact on the Type I error rates for the proposed test. Even though Q uses a poly-3 survival adjustment [1], originally derived under the assumption of a Weibull tumor incidence model with a shape parameter of 3, the false positive rates for shapes $\gamma_1 = 1.5$ and $\gamma_1 = 6$ were essentially the same as for $\gamma_1 = 3$. Similarly, the Type I error rates did not seem to be affected by increasing the heterogeneity of the tumor onset times and death times among the historical control groups (Table 3).

3.4. Power

The power of the proposed test to detect a difference between π^c and π^h is summarized in Table 4. As expected, Q gained power as the number of historical control groups increased. Except for the case of $k = 1$, the power to detect a small difference between the mean tumor rates (i.e., $ln(\pi^c) = 1.25 \times ln(\pi^h)$) varied from 23% to 44%, and the power to detect a large difference (i.e., $ln(\pi^c) = 1.5 \times ln(\pi^h)$) varied from 77% to 99%. As with the Type I error rates, the power of Q did not depend on the shape of the incidence curve (γ_1) or whether any extra variability was present. Thus, the proposed method, which accounts for survival differences between controls, is robust to variations in the shape of the incidence curves.

We do not present powers for the range-based rules because their Type I error rates can be much too high or much too low, depending on the number of historical studies involved. As an interesting special case, however, we examined rejection rates for $k = 39$, where the predicted Type I error rate for a range-based rule in the absence of tied tumor rates, $2/(k+1)$, equals the nominal 0.05 level. In the null case, the rejection rates for R^* were right on target (4.9% to 5.5%) for tumor rates of 0.15 and 0.30, but were lower (2.4% to 3.5%) for tumor rates of 0.01 and 0.05 (as expected with the higher number of ties produced by lower tumor rates). For the higher tumor rates (0.15 and 0.30), where both Q and R^* operated at the nominal 5% level, Q had 17% to 24% greater power than R^* to detect small differences in tumor rates (e.g., 44.1% for Q versus 35.7% for R^*) and 2% to 5% greater power to detect large differences (even though there was not much room for improvement, as the powers were all above 90%). Thus, even when selecting the "best" value of k with respect to the Type I error rate of R^*, the proposed test was more powerful than the range-based decision rule.

4. ANALYSIS OF BENZOPHENONE DATA

Benzophenone is an aryl ketone, produced in large quantities in the United States, with widespread occupational and consumer exposures through its use as a fragrance enhancer, flavor additive, photoinitiator, and ultraviolet curing agent [8]. It is also used in manufacturing pharmaceuticals, insecticides, and agricultural chemicals, as well as being an additive in plastics and adhesives. Short-term animal studies suggested that the liver and kidneys were the target organs, but toxicity also was observed in the hematopoietic system.

The NTP conducted a 2-year study of male and female $B6C3F_1$ mice and F 344/N rats exposed to benzophenone. Our example focuses on mononuclear cell leukemia (MCL) in female rats. Groups of size 50 received doses of 0, 312, 625, or 1250 ppm of benzophenone in their diet throughout the study. The numbers of female rats that developed MCL were 19, 25, 30, and 29, respectively, with poly-3 survival-adjusted tumor rates of 42.3%, 51.5%, 61.3%, and 59.6%. With respect to female rats, the NTP concluded that there was equivocal evidence of carcinogenic activity of benzophenone, based in part on marginally increased incidences of MCL and histiocytic sarcoma.

Several factors contributed to the uncertainty in the NTP decision. The NTP's trend test gave a p-value of 0.058 for the current experimental data, which is not statistically significant at the usual 0.05 level, though the pairwise comparison of the control and mid-dose groups was marginally significant (p = 0.048). On the other hand, accounting for historical control data supported the notion of an increasing trend in MCL rates with dose. The NTP examined 6 contemporary feed studies and found lower MCL rates among the untreated female rats (see Table B4b of the NTP report [8]). The unadjusted tumor rates ranged from 12% to 35% in these 6 historical control groups, which suggests the spontaneous MCL rate might be lower than the 38% rate observed in the current control group. Similarly, the corresponding poly-3 survival-adjusted rates were 42.3% in the current study and 12.7% to 35.6% in the historical studies. A lower MCL rate among controls would produce a more significant p-value for an increasing dose-related trend, especially in view of the relatively high tumor rates in the treated groups (i.e., poly-3 rates of 51.5%, 61.3%, and 59.6%). This argument is valid if we believe the current and historical controls have the same mean tumor rate, but otherwise we cannot necessarily draw that conclusion.

The NTP observed that the MCL rate in the current control group fell outside the historical range and, for that and other reasons, declared an equivocal result with respect to the possible carcinogenicity of benzophenone in female rats. As a range-based decision rule can reject too often when there are only k = 6 historical control groups, we applied the proposed test to these same data. Our formal test, which operates at the proper level, gave a 2-sided significance value of p = 0.010, which supports the NTP's informal observation that the current control group differs from the historical control groups with respect to MCL.

5. DISCUSSION

Although various statistical methods for incorporating historical control information have been proposed over the past few decades, none have gained widespread use by scientists in the field or by regulatory agencies. For example, many early procedures assumed a beta-binomial model [9], which allowed for extra-binomial variation among the control groups. Other related procedures involved generalized binomial [10] or logistic-normal [11] models. An important problem with these approaches, however, is that they do not adjust for survival, which can introduce bias when mortality differs across groups, as all animals are not at equal risk for developing a tumor. Alternative methods account for survival but make assumptions about tumor lethality [12]. Several Bayesian procedures adjust for survival and avoid lethality assumptions [13] but require investigators to specify prior distributions and hyperparameters. For these and other reasons, none of these methods have been adopted for routine use in practice.

Recently, the Technical Reports Review Subcommittee of the NTP Board of Scientific Counselors, which included two statisticians, decided against endorsing any of the current statistical methods and instead recommended developing a new procedure to address the important problem of incorporating historical control data in the analysis of a current study (http://ntp.niehs.nih.gov/files/TRRSMins0905.pdf). Consequently, Peddada et al [5] developed a simple trend test that incorporates historical control data, adjusts for survival, and makes no assumptions about tumor lethality or parametric distributions. Further discussions with NTP toxicologists and pathologists revealed the need for an additional test for comparing current and historical control groups, which was the motivation for this article.

In summary, when evaluating a chronic bioassay, toxicologists and pathologists routinely assess the relevance of historical studies to the current study by comparing tumor rates in the current control group to the range of control tumor rates from contemporary studies performed under similar conditions. Current and historical control groups are often informally labeled dissimilar if the current control rate falls outside the historical range; see Keenan et al [2] for a discussion. One natural concern is that this type of approach might be conservative and have low power when the historical range becomes too wide, which can occur if *k* is large or if there is an outlier among the tumor rates. There has been a recent push for creating a global historical control database [3], where a large value of *k* could lead to a range-based procedure that could be extremely conservative and have little power. A less appreciated, though possibly more disturbing, concern is that such a range-based process could be highly anti-conservative with huge Type I error rates when *k* is small. As an alternative to this type of range-based approach, we provide a simple procedure that controls Type I errors, while adjusting for survival effects, accounting for extra variability among historical control groups, and avoiding tumor lethality assumptions.

We emphasize that an important feature of the proposed test is that it works well with a small number of studies and in fact can be applied with only one historical study, whereas range-based decision rules would perform poorly for small *k* and would not even be defined for $k = 1$. For example, the NTP recently switched from using Fischer rats to using Sprague Dawley rats in its 2-year rodent cancer bioassay. Initially the new historical control database will not contain enough Sprague Dawley rat studies to construct a reasonable range of tumor rates, but the proposed method will be readily applicable. This issue is widespread, as many small labs conducting rodent cancer bioassays do not have extensive historical control databases. Some have even discussed the construction of a global database to deal with these types of situations [3]. Our proposed approach provides a simple solution to this problem.

Finally, although the bootstrap procedure for approximating the null distribution of the proposed test statistic in extreme cases (i.e., $k < 5$ or $\hat{\pi}^h \leq 0.005$) does not account for between-group variability, the Type I error rate is maintained. The validity of the bootstrap methodology of Peddada et al [7] is not surprising for $\hat{\pi}^h \leq 0.005$ because the between-group variability in this situation is expected to be trivial. In the remaining extreme cases, where $\hat{\pi}^h > 0.005$ but $k < 5$, the proposed test still operates at the nominal level, perhaps because the test statistic accounts for between-group variability, even though the bootstrap procedure does not. Our extensive simulations indicate that our test maintains the correct Type I error rate, despite the bootstrap not adjusting for between-group variability.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES045007 and Z01-ES101744). The authors thank Grace Kissling, Joseph Haseman, and the reviewers for their helpful comments.

References

1. 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]
2. Keenan C, Elmore S, Francke-Carroll S, Kemp R, Kerlin R, Peddada S, Pletcher J, Rinke M, Schmidt S, Taylor I, Wolf D. Best practices for use of historical control data of proliferative rodent lesions. Toxicologic Pathology. 2009; 37:679–693. [PubMed: 19454599]
3. Keenan C, Elmore S, Francke-Carroll S, Kerlin R, Peddada S, Pletcher J, Rinke M, Schmidt S, Taylor I, Wolf D. Potential for a global historical control database for proliferative rodent lesions. Toxicologic Pathology. 2009; 37:677–678. [PubMed: 19638441]
4. Elmore S, Peddada S. Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. Toxicologic Pathology. 2009; 37:672–676. [PubMed: 19516052]
5. Peddada S, Dinse G, Kissling G. Incorporating historical control data when comparing tumor incidence rates. Journal of the American Statistical Association. 2007; 102:1212–1220. [PubMed: 20396669]
6. Bieler G, Williams R. Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. Biometrics. 1993; 49:793–801. [PubMed: 8241374]
7. Peddada SD, Prescott K, Conaway M. Tests for order restrictions in binary data. Biometrics. 2001; 57:1219–1227. [PubMed: 11764263]
8. National Toxicology Program. Technical Report Series No. 533, NIH Publication No. 05-4469. 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 Carcino-genesis Studies of Benzophenone (CAS No. 119-61-9) in $F344/N$ Rats and $B6C3F_1$ Mice (Feed Studies).
9. Tarone R. The use of historical control information in testing for a trend in proportions. Biometrics. 1982; 38:214–220.
10. 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.
11. 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.
12. Ibrahim J, Ryan L. Use of historical controls in time-adjusted trend tests for carcinogenicity. Biometrics. 1996; 52:1478–1485. [PubMed: 8962464]
13. Dunson D, Dinse G. Bayesian incidence analysis of animal tumorigenicity data. Applied Statistics. 2001; 50:125–141.

Table 1

Simulation values for the baseline incidence scale parameter (ψ_1) and the extra variation parameter (τ) by tumor rate (π^h), shape of the incidence curve (γ_1), and number of historical control groups (k).

			Values of τ for:		
π^h	γ_1	Value of ψ_1	k=2	k=5	k=10
0.01	1.5	9.00×10^{-5}	0.27	0.31	0.52
	3.0	8.00×10^{-7}	0.29	0.31	0.54
	6.0	6.50×10^{-11}	0.33	0.32	0.57
0.05	1.5	4.70×10^{-4}	0.25	0.27	0.37
	3.0	4.20×10^{-6}	0.26	0.28	0.38
	6.0	3.25×10^{-10}	0.29	0.28	0.39
0.15	1.5	1.50×10^{-3}	0.22	0.22	0.26
	3.0	1.34×10^{-5}	0.23	0.23	0.27
	6.0	1.04×10^{-9}	0.25	0.24	0.28
0.30	1.5	3.30×10^{-3}	0.20	0.19	0.21
	3.0	2.97×10^{-5}	0.21	0.20	0.22
	6.0	2.32×10^{-9}	0.22	0.21	0.23

Table 2

Type I error rates for the unadjusted (R) and survival-adjusted (R*) range-based tests by tumor rate (π), shape of the incidence curve (γ_1), number of historical control groups (k), and absence/presence of extra variation. All tests were based on a sample size of n=50 per group, with 70% of the animals (on average) surviving to the terminal sacrifice.

		No extra variation[b]			Extra variation[b]		
π[a]	γ_1	k=2	k=5	k=10	k=2	k=5	k=10
		Unadjusted range-based test (R)					
0.01	1.5	27.9	9.0	4.4	30.2	9.6	4.6
	3.0	27.9	9.0	4.3	30.1	9.4	4.8
	6.0	28.6	8.8	5.1	31.0	9.7	4.6
0.05	1.5	48.7	20.9	11.0	49.6	22.0	11.5
	3.0	48.4	21.0	11.0	49.6	21.7	11.7
	6.0	49.2	20.8	11.2	50.0	21.5	11.4
0.15	1.5	55.9	26.0	13.1	56.1	25.8	13.6
	3.0	56.1	26.3	13.1	56.3	26.1	13.4
	6.0	56.1	26.2	13.0	56.1	26.5	13.6
0.30	1.5	57.7	26.5	14.0	58.1	26.7	14.1
	3.0	58.1	27.2	14.0	58.7	27.1	14.1
	6.0	58.7	26.7	13.8	58.5	26.9	14.5
		Poly-3 survival-adjusted range-based test (R*)					
0.01	1.5	36.9	16.4	9.1	40.4	16.8	8.8
	3.0	37.0	16.4	9.1	39.9	16.7	9.2
	6.0	38.0	16.3	9.9	41.4	17.3	9.1
0.05	1.5	65.6	31.5	17.1	66.2	32.6	18.4
	3.0	65.8	31.8	17.4	66.1	32.3	18.5
	6.0	65.9	31.9	17.2	66.2	32.4	17.8
0.15	1.5	67.0	34.0	18.2	66.3	33.7	18.2
	3.0	67.0	34.0	17.9	66.6	33.8	18.1
	6.0	66.7	33.8	17.9	66.6	33.8	18.4
0.30	1.5	65.9	33.2	18.2	66.1	32.6	17.8
	3.0	66.2	32.9	17.9	66.5	33.3	17.7
	6.0	66.5	32.9	17.5	66.8	33.2	18.3

[a] The mean tumor rate among the historical control groups (π^h) equals the mean tumor rate in the current control group (π^c) under the null hypothesis (H_0: $\pi^h = \pi^c = \pi$).

[b] The variance of the tumor rates among the historical control groups in the presence of extra variation is 20% larger than in the absence of extra variation.

Table 3

Type I error rates for the proposed test (Q) by tumor rate (π), shape of the incidence curve (γ_1), number of historical control groups (k), and absence/presence of extra variation. All tests were performed at the nominal 5% level, with a sample size of n=50 per group and with 70% of the animals (on average) surviving to the terminal sacrifice.

		No extra variation[b]				Extra variation[b]		
π[a]	**γ_1**	**k=1**	**k=2**	**k=5**	**k=10**	**k=2**	**k=5**	**k=10**
0.01	1.5	2.3	5.5	4.6	5.0	5.5	4.4	4.5
	3.0	2.2	5.5	4.5	5.0	5.5	4.4	5.1
	6.0	2.4	5.5	4.3	5.7	5.8	4.1	4.8
0.05	1.5	5.4	5.1	3.9	4.1	5.1	4.4	4.6
	3.0	5.3	5.0	3.7	4.3	5.1	4.3	4.7
	6.0	5.3	5.0	3.8	4.2	5.0	4.1	4.4
0.15	1.5	4.8	5.1	4.8	4.8	4.9	4.8	4.8
	3.0	4.7	5.0	4.7	4.9	5.1	4.7	4.8
	6.0	4.4	4.9	4.9	4.8	4.8	4.9	4.9
0.30	1.5	5.3	5.0	5.1	5.0	4.8	5.0	5.0
	3.0	5.2	5.0	5.2	4.8	4.9	5.1	5.2
	6.0	5.2	5.0	5.1	5.1	5.0	5.0	5.3

[a] The mean tumor rate among the historical control groups (π^h) equals the mean tumor rate in the current control group (π^c) under the null hypothesis (H_0: $\pi^h = \pi^c = \pi$).

[b] The variance of the tumor rates among the historical control groups in the presence of extra variation is 20% larger than in the absence of extra variation.

Table 4

Power of the proposed test (Q) to detect small and large differences[a] between mean tumor rates for historical (π^h) and current (π^c) controls by tumor rate (π^h), shape of the incidence curve (γ_1), number of historical control groups (k), and absence/presence of extra variation. All tests were performed at the nominal 5% level, with a sample size of n=50 per group and with 70% of the animals (on average) surviving to the terminal sacrifice.

		No extra variation[b]				**Extra variation[b]**		
π^h	**γ_1**	**k=1**	**k=2**	**k=5**	**k=10**	**k=2**	**k=5**	**k=10**
		Small difference between π^c and π^h						
0.01	1.5	18.1	24.0	26.4	33.1	25.3	30.0	38.7
	3.0	18.3	24.1	26.9	33.3	25.5	29.7	37.7
	6.0	18.0	23.0	25.9	32.9	24.1	27.8	35.5
0.05	1.5	7.3	25.6	35.6	39.5	27.7	39.1	43.9
	3.0	7.1	25.6	35.8	40.4	27.7	39.0	43.8
	6.0	7.1	25.4	35.7	40.3	27.3	38.1	43.2
0.15	1.5	9.9	27.0	36.7	39.8	29.0	38.7	43.1
	3.0	10.0	27.6	36.9	40.1	29.4	39.4	42.9
	6.0	10.3	27.6	36.6	40.4	28.9	38.9	42.2
0.30	1.5	17.5	24.2	31.7	34.5	25.5	33.0	36.0
	3.0	17.8	24.6	32.2	35.3	26.4	33.5	36.7
	6.0	18.4	25.6	32.7	35.6	26.4	33.8	36.6
		Large difference between π^c and π^h						
0.01	1.5	55.5	78.0	87.7	91.9	81.9	91.9	96.1
	3.0	55.8	78.1	87.9	91.5	81.4	91.6	96.0
	6.0	54.7	76.9	86.7	91.3	79.6	90.2	94.8
0.05	1.5	24.3	88.0	95.9	97.4	90.8	97.6	98.8
	3.0	22.5	86.1	95.2	96.7	88.7	96.8	98.2
	6.0	23.7	86.3	95.3	96.9	88.7	96.7	98.0
0.15	1.5	69.4	86.5	95.1	96.7	89.2	96.7	97.9
	3.0	71.2	87.2	95.5	96.7	89.4	96.7	98.1
	6.0	72.7	87.4	95.6	97.2	88.9	96.6	97.9
0.30	1.5	70.5	80.0	91.2	93.8	82.6	92.7	95.0
	3.0	71.4	81.3	92.1	94.5	83.1	93.3	95.0

		No extra variation[b]				Extra variation[b]		
π^h	**γ_1**	**k=1**	**k=2**	**k=5**	**k=10**	**k=2**	**k=5**	**k=10**
	6.0	74.1	82.7	93.1	94.7	83.0	93.8	95.3

[a] Small (large) differences are values of π^c that are 25% (50%) larger than π^h on the log scale.

[b] The variance of the tumor rates among the historical control groups in the presence of extra variation is 20% larger than in the absence of extra variation.