Chapter 12
Recent Research Projects by the FDA's Pharmacology and Toxicology Statistics Team

Karl K. Lin, Matthew T. Jackson, Min Min, Mohammad Atiar Rahman, and Steven F. Thomson

Abstract In addition to regular review work, the Pharmacology and Toxicology Statistics Team in CDER/FDA is actively engaged in a number of research projects. In this chapter we summarize some of our recent investigations and findings.

We have conducted a simulation study (discussed in Sect. 12.2) to evaluate the increase in Type 2 error attributable to the adoption by some non-statistical scientists within the agency of more stringent decision criteria than those we have recommended for the determination of statistically significant carcinogenicity findings in long term rodent bioassays. In many cases, the probability of a Type 2 error is inflated by a factor of 1.5 or more.

A second simulation study (Sect. 12.3) has found that both the Type 1 and Type 2 error rates are highly sensitive to experimental design. In particular, designs using a dual vehicle control group are more powerful than designs using the same number of animals but a single vehicle control group, but this increase in power comes at the expense of a greatly inflated Type 1 error rate.

Since the column totals of the tables of permutations of animals to treatment groups cannot be presumed to be fixed, the exact methods used in the Cochran-Armitage test are not applicable to the poly-*k* test for trend. Section 12.4 presents simple examples showing all possible permutations of animals, and procedures for computing the probabilities of the individual permutations to obtain the exact *p*-values. Section 12.5 builds on this by proposing an exact ratio poly-*k* test method using samples of possible permutations of animals. The proposed ratio poly-*k*

The article reflects the views of the authors and should not be construed to represent FDA's views or policies.

K.K. Lin (✉) • M. Min • M.A. Rahman • S.F. Thomson
US Food and Drug Administration, Center for Drug Evaluation and Research, Office of Translational Sciences, Office of Biostatistics, Division of Biometrics 6, Silver Spring, MD, USA
e-mail: karl.lin@fda.hhs.gov

M.T. Jackson
Formerly of FDA/CDER/OTS/OB/DB6, Silver Spring, MD, USA

L. Zhang (ed.), *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*, Statistics for Biology and Health, DOI 10.1007/978-3-319-23558-5_12

test does not assume fixed column sums and uses the procedure in Bieler and Williams (Biometrics 49(3):793–801, 1993) to obtain the null variance estimate of the adjusted quantal tumor response estimate. Results of simulations show that the modified exact poly-3 method has similar sizes and levels of power compared to the method proposed in Mancuso et al. (Biometrics 58:403–412, 2002) that also uses samples of permutations but uses the binomial null variance estimate of the adjusted response rates and is based on the assumption of fixed column sums.

Bayesians attempt to model not only the statistical data generating process as in the frequentist statistics, but also to model knowledge about the parameters governing that process. Section 12.6 includes a short review of possible reasons for adopting a Bayesian approach, and examples of survival and carcinogenicity analyses.

Keywords Bayesian methods in nonclinical biostatistics • Carcinogenicity studies • Consumer's risk • Exact poly-3 trend tests • Experimental designs • Finite dimensional logistic model • Finite dimensional proportional • Hazards model • Multiplicity adjustment • Nonparametric Bayesian analysis • Permutational distribution • Producer's risk

12.1 Introduction

It is required by law that the sponsor of a new drug that is intended for chronic use by patients for certain indications conduct carcinogenicity studies in animals to assess the carcinogenic potential of the drug. These studies are reviewed independently within CDER/FDA. These independent reviews are conducted by interdisciplinary groups. The statistical component of such a review includes an assessment of the design and conduct of the study, and a complete reanalysis of all statistical data. This work is performed by members of the Pharmacology and Toxicology (Pharm/Tox) Statistics Team. However, the decision of whether the drug should be considered a potential carcinogen is based on more than just statistical evidence, and as such the statistical review comprises just one part of the FDA internal decision process.

In 2001, the FDA released a draft document entitled "Guidance for Industry; Statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals" (US Food and Drug Administration—Center for Drug Evaluation and Research 2001). This guidance was issued in the Federal Register (Tuesday, May 8 2001, Vol. 66 No. 89) and was reviewed by the public over a 90 day comment period in 2001. Sixteen comments were received from drug companies, professional organizations of the pharmaceutical industry, and individual experts from the U.S., Europe, and Japan; these are available in FDA Docket No. 01D0194. Great efforts are being made within the FDA to finalize this document.

The draft guidance describes the general process and methods used by the Pharm/Tox team in their reviews. These methods are also used widely by drug

companies in the U.S. and abroad. In addition, the team also draws on results from other published research.

In addition to writing reviews, the statistical Pharm/Tox team contributes to the development of FDA policy and conducts statistical research. Indeed, research published by members of the team (Lin 1998; Lin and Ali 2006; Lin and Rahman 1998; Lin et al. 2010; Rahman and Lin 2008, 2009, 2010; Rahman and Tiwari 2012) is often incorporated into the team's statistical reviews. More generally, in their efforts to keep abreast of new advancements in this area and to improve the quality of their reviews, members of the team have conducted regulatory research studies in collaborations with experts within and outside the agency.

The purposes of this chapter are twofold. One is to provide updates of results of some research studies that have been presented or published previously (Sects. 12.4 and 12.6). The other is to share with the professional community the results of some research studies that have not been presented or published (Sects. 12.2, 12.3 and 12.5).

In this chapter, results of five recent research projects by members of the Pharm/Tox Statistics Team are presented. Sections 12.2 and 12.3 describe two simulation studies investigating the effect on Type 1 and Type 2 error of varying the decision rules (Sect. 12.2) and experimental design (Sect. 12.3). Sections 12.4 and 12.5 discuss the development of exact methods for the poly-k test for a dose response relationship. Finally, Sect. 12.6 is a general discussion of the use of Bayesian methods in reviews of carcinogenicity studies.

12.2 An Evaluation of the Alternative Statistical Decision Rules for the Interpretation of Study Results

12.2.1 Introduction

It is specifically recommended in the draft guidance document (US Food and Drug Administration—Center for Drug Evaluation and Research 2001) that when evaluating the carcinogenic potential of a drug:

1. Trend tests, which have been extensively studied in the literature (Lin 1995, 1997, 1998, 2000a,b; Lin and Ali 1994, 2006; Lin and Rahman 1998; Rahman and Lin 2008, 2009), should be the primary tests used.
2. Pairwise tests may be used in lieu of trend tests, but only in those rare cases where they are deemed more appropriate than the trend tests.

The reason for preferring the trend test to the pairwise test is that, under most circumstances, the trend test will be more powerful than the pairwise test (for any given significance level). Note that the above guidance document recommends that only one test, either the trend test or the pairwise test, is to be used to conclude a statistically significant carcinogenic effect.

In the context of carcinogenicity studies (and safety studies in general), the Type 1 error rate is primarily a measure of the producer's risk; if a drug is withheld from market due to an incorrect finding of a carcinogenic effect (or even if its usage is merely curtailed), then it is the producer of the drug who faces the greatest loss. Conversely, the Type 2 error rate is primarily a measure of the consumer's risk, as it is the consumer who stands to suffer in the event that a truly carcinogenic drug is brought to market without the carcinogenic effect being reported. Proceeding from this philosophical stance (and the similar position of Center for Drug Evaluation and Research 2005), the draft guidance (US Food and Drug Administration—Center for Drug Evaluation and Research 2001) recommends a goal of maximizing power while keeping the overall (study-wise) false positive rate at approximately 10 %. In order to achieve this goal, a collection of significance thresholds are recommended. These thresholds, presented in Table 12.1, are grounded in Lin and Rahman (1998) and Rahman and Lin (2008) (for the trend test) and Haseman (1983, 1984) (for the pairwise test).

However, this goal has not been universally accepted. There is a desire on the part of some non-statistical scientists within the agency to restrict positive findings to those where there is statistical evidence of *both* a positive dose response relationship *and* an increased incidence in the high dose group compared to the control group. In other words, a joint test is desired. This is not an intrinsically unreasonable position. Nonetheless, every test needs significance thresholds, and since the only significance thresholds included in US Food and Drug Administration—Center for Drug Evaluation and Research (2001) are for single tests, it is natural (but incorrect!) for non-statistical scientists to construct a joint test using these thresholds. We will refer to this decision rule as the *joint test* rule. See Table 12.2.

We are very concerned about the ramifications of the use of this rule. While the trend and pairwise test are clearly not independent, their association is far from perfect. Accordingly, the requirement that both tests yield individually statistically significant results necessarily results in a more conservative test than either the trend test or the pairwise test alone (at the same significance thresholds). The purpose of this section is to present the results of our simulation study showing a serious consequence of the adoption of this rule: a huge inflation of the false negative rate (i.e., the consumer's risk) for the final interpretation of the carcinogenicity potential of a new drug.

12.2.2 Design of Simulation Study

The objective of this study is to conduct a simulation study to evaluate the inflation of the false negative rate resulting from the joint test (compared with the trend test alone).

We modeled survival and tumor data using Weibull distributions (see Eqs. (12.1) and (12.2)). The values of the parameters A, B, C, and D, were taken from the landmark National Toxicology Program (NTP) study by Dinse (1985) (see

Table 12.1 Recommended significance levels for the trend test or the pairwise comparisons (US Food and Drug Administration—Center for Drug Evaluation and Research 2001: Lines 1093–1094 on page 30)

	Tests for positive trend	Control-high pairwise comparisons (one-tailed)
Standard 2-year studies with 2 species and 2 sexes	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively
Alternative ICH studies (one 2-year study in one species and one short- or medium- term study, two sexes)	Common and rare tumors are tested at 0.01 and 0.05 significance levels respectively	Under development and not yet available

Note: Following Haseman (1983), a tumor is classified as a rare tumor if it has a background rate of less than 1 %, and is classified as a common tumor otherwise

Table 12.2 The joint test rule (not recommended!)

	Trend test	Pairwise test
Rare tumor	0.025	0.050
Common tumor	0.005	0.010

Note: The joint test is positive for a particular tumor endpoint only if both the pairwise test between the high dose and control groups and the dose response (trend) test are significant at the levels indicated above

Tables 12.3 and 12.4). Values of these parameters were chosen to vary four different factors, ultimately resulting in 36 different sets of simulation conditions.

The factors used in the NTP study were defined as follows:

1. Low or high tumor background rate: The prevalence rate at 2 years in the control group is 5 % (low) or 20 % (high).
2. Tumors appear early or late: The prevalence rate of the control group at 1.5 years is 50 % (appearing early) or 10 % (appearing late) of the prevalence rate at 2 years.
3. No dose effect, a small dose effect, or a large dose effect on tumor prevalence: The prevalence of the high dose group at 2 years minus the prevalence of the control group at 2 years is 0 % (no effect), or 10 % (small effect), or 20 % (large effect).
4. No dose effect, a small dose effect, or a large dose effect on mortality: The expected proportion of animals alive in the high dose group at 2 years is 70 % (no effect), 40 % (small effect), or 10 % (large effect). The expected proportion of animals alive in the control group at 2 years is taken as 70 %.

However, there are important differences between the NTP design described above and the design used in our simulation study. Whereas the NTP study simulated three treatment groups with doses $x = 0$, $x = 1$, and $x = 2$ (called the *control*, *low*, and *high* dose groups), our study used four treatment groups (with doses $x = 0$,

Table 12.3 Data generation parameters for the Weibull models for time to tumor onset (Dinse 1985)

Simulation conditions	Model description				Weibull parameters			
	Background tumor rate		Tumor appearance[a]	Dose effect[b,c]	*A*	*B*	*C*	*D*
1, 13, 25	Low	0.05	Early	None	17	2	6.78×10^{-6}	0
2, 14, 26	Low	0.05	Early	Small	17	2	6.78×10^{-6}	7.36×10^{-6}
3, 15, 27	Low	0.05	Early	Large	17	2	6.78×10^{-6}	1.561×10^{-5}
4, 16, 28	Low	0.05	Late	None	56	3	4.65×10^{-7}	0
5, 17, 29	Low	0.05	Late	Small	56	3	4.65×10^{-7}	5.025×10^{-7}
6, 18, 30	Low	0.05	Late	Large	56	3	4.65×10^{-7}	1.0675×10^{-6}
7, 19, 31	High	0.20	Early	None	21	2	3.24×10^{-5}	0
8, 20, 32	High	0.20	Early	Small	21	2	3.24×10^{-5}	9.7×10^{-6}
9, 21, 33	High	0.20	Early	Large	21	2	3.24×10^{-5}	2.09×10^{-5}
10, 22, 34	High	0.20	Late	None	57	3	2.15×10^{-6}	0
11, 23, 35	High	0.20	Late	Small	57	3	2.15×10^{-6}	6.45×10^{-7}
12, 24, 36	High	0.20	Late	Large	57	3	2.15×10^{-6}	1.383×10^{-6}

Notes on factors used in the simulation by Dinse (1985):
[a]Tumors appear early or late: The prevalence rate of the control group at 1.5 years is 50 % (appearing early) or 10 % (appearing late) of the prevalence rate at 2 years
[b]No effect, a small effect, or a large effect on tumor prevalence: The prevalence of the high dose group ($x = 2$) at 2 years minus the prevalence of the control group at 2 years is 0 % (none effect), or 10 % (small effect), or 20 % (large effect)
[c]It is also be noted that for our study, the percentage differences corresponding to those in note b are 0, 15, and 28 % (for the high dose group with $x = 3$)

Table 12.4 Data generation parameters for the Weibull models for time to death (Dinse 1985)

Simulation conditions	Drug effect on death[a,b]	Weibull parameters			
		A	*B*	*C*	*D*
1–12	None	0	4	3.05×10^{-9}	0
13–24	Small	0	4	3.05×10^{-9}	2.390×10^{-9}
25–36	Large	0	4	3.05×10^{-9}	8.325×10^{-9}

Notes on factors used in the simulation by Dinse (1985):
[a]No effect, a small effect, or a large effect on mortality: The expected proportion of animals alive in the high dose group ($x = 2$) at 2 years is 70 % (none), 40 % (smalleffect), or 10 % (large effect). The expected proportion of animals alive in the control group at 2 years is taken as 70 %
[b]It is also be noted that for our study, the survival probabilities corresponding to those in note d are 70, 30, and 4 % (for the high dose group with $x = 3$)

$x = 1$, $x = 2$, and $x = 3$, called the *control*, *low*, *mid*, and *high* dose groups respectively). Since the values of the parameters *A*, *B*, *C*, and *D* used were the same in the two studies (see Tables 12.3 and 12.4), the characterizations of the effect of the dose level on tumorigenesis and mortality, factors 3 and 4, apply to the dose

level $x = 2$, i.e., to the mid dose level. To recast these descriptions in terms of the effect at the $x = 3$ (high dose) level, factors 3 and 4 become factors 3′ and 4′:

3′ No dose effect, a small dose effect, or a large dose effect on tumor prevalence: The prevalence of the high dose group at 2 years minus the prevalence of the control group at 2 years is 0 % (no effect), or approximately 15 % (small effect), or approximately 28 % (large effect).

4′ No dose effect, a small dose effect, or a large dose effect on mortality: The expected proportion of animals alive in the high dose group at 2 years is 70 % (no effect), 30 % (small effect), or 4 % (large effect). The expected proportion of animals alive in the control group at 2 years is taken as 70 %.

These differences can be expected to have the following effects on the Type 2 error rates for our study (relative to the NTP study):

- The higher tumorigenesis rates in the high dose groups should help to reduce the false negative rates (or to increase the levels of power) of statistical tests.
- On the other hand, higher levels of mortality will reduce the effective sample size and thus tend to increase the false negative rates (or to decrease the levels of power).[1]

In our study, tumor data were generated for 4 treatment groups with equally spaced increasing doses (i.e., $x = 0$, $x = 1$, $x = 2$, and $x = 3$). There were 50 animals per group. The study duration was 2 years (104 weeks), and all animals surviving after 104 weeks were terminally sacrificed. All tumors were assumed to be incidental.

The tumor detection time (T_0) (measured in weeks) and the time to natural death (T_1) of an animal receiving dose level x were modeled by four parameter Weibull distributions:

$$S(t,x) = P[T_i > t | X = x] = \begin{cases} e^{-(C+Dx)(t-A)^B} & \text{if } t > A \\ 1 & \text{if } t \le A \end{cases} \tag{12.1}$$

where A is the location parameter, B is the shape parameter, C is the baseline scale parameter, and D is the dose effect parameter. Tables 12.3 and 12.4 list the sets of values for these parameters used in Dinse (1985).

The prevalence function for incidental tumors equals the cumulative function of time to tumor onset, i.e.,

$$P(t|x) = \Pr[T_0 \le t | X = x] = 1 - S(t,x). \tag{12.2}$$

Each of the 36 simulation conditions described in Tables 12.3 and 12.4 was simulated 10,000 times. For each simulation, 200 animals were generated; each animal was assigned to a dose group (50 animals per group) and had a tumor

[1] Although this tendency is not absolute. See the discussion in footnote 8 on page 313.

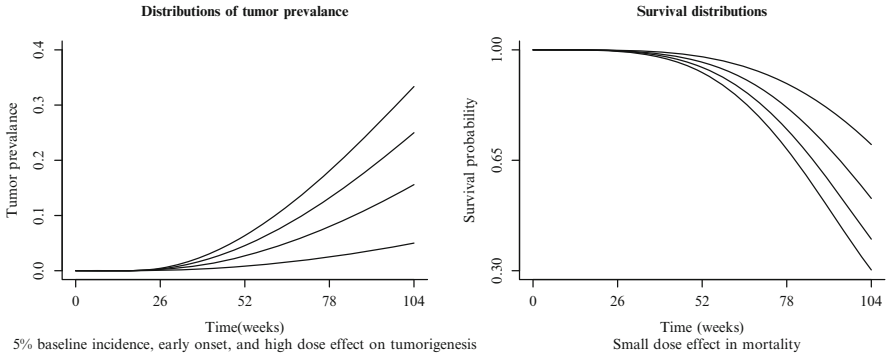

Fig. 12.1 Sample tumor prevalence and mortality curves

onset time (T_0) and death time (T_1) simulated using Eq. (12.1). The actual time of death (T) for each animal was defined as the minimum of T_1 and 104 weeks, i.e., $T = \min\{T_1, 104\}$. The animal developed the tumor (i.e., became a tumor bearing animal (TBA)) only if the time to tumor onset did not exceed the time to death. The actual tumor detection time was assumed to be the time of death T. Animals in the same dose group were equally likely to develop the tumor in their life times. It was assumed that tumors were developed independently of each other. The first panel of Fig. 12.1 graphically represents the Weibull models used to generate the tumor prevalence data when the background tumor rate is low, the dose effect on tumor prevalence is large, and the tumor appears early (the model used in simulation conditions 3, 15, and 27). The second panel graphically represents the Weibull models used to generate the survival data when the dose effect on mortality is small (simulation conditions 13–24). The age-adjusted Peto method to test for a dose response relationship (Peto et al. 1980) and the age adjusted Fisher exact test for pairwise differences in tumor incidence, (in each case using the NTP partition of time intervals[2]), were applied to calculate p-values.

Three rules for determining if a test of the drug effect on development of a given tumor type was statistically significant were applied to the simulated data. They were:

1. Requiring a statistically significant result in the trend test alone. This is the rule recommended in US Food and Drug Administration—Center for Drug Evaluation and Research (2001).
2. Requiring statistically significant results both in the trend test and in any of the three pairwise comparison tests (control versus low, control versus medium, control versus high).

[2]The NTP partition divides the 104 week study into the following subintervals: 0–52 weeks, 53–78 weeks, 79–92 weeks, 93–104 weeks, and terminal sacrifice.

3. Requiring statistically significant results both in the trend test and in the control versus high group pairwise comparison test. This is the joint test rule.

In each case, it was assumed that the tests were being conducted as part of a standard two-species study. The rules for rare tumor types were used when the incidence rate in the control group were below 1%; otherwise the rules for common tumors types were used.

After simulating and analyzing tumor data 10,000 times for each of the 36 sets of simulation conditions, the Type 1 and Type 2 error rates were estimated.

12.2.3 Results of the Simulation Study

Since we are simultaneously considering both models where the null hypothesis is true (so that there is no genuine dose effect on tumor incidence) and models where it is false (where there is a genuine dose effect), we need terminology that can apply equally well to both of these cases. For any given set of simulation conditions, the *retention rate* is the probability of retaining the null hypothesis. If the null hypothesis is true, then this rate is the probability of a *true negative*, and is 1 − thefalsepositiverate (Type I error). If the null hypothesis is false, then the retention rate is the probability of a *false negative* or Type 2 error. In this case, it is 1 − power. Correspondingly, the *rejection rate* is 1 − theretentionrate, and is the probability that the null hypothesis is rejected. It is either the false positive rate (if the null hypothesis is true) or the level of power (if the alternative hypothesis is true). The results (retention rates and percent changes of retention rates) of the simulation study are presented in Table 12.5.

Results of the evaluation of Type 1 error patterns in the study conducted and reported in Dinse (1985) show that the Peto test without continuity correction and with the partition of time intervals of the study duration proposed by NTP (see Footnote 2) yields attained false positive rates close to the nominal levels (0.05 and 0.01) used in the test. That means that the test is a good one that is neither conservative nor anti-conservative.

The evaluation of Type 1 error patterns found by this simulation study is done by using the rates at which the null hypothesis was rejected under those simulation conditions for which there was no dose effect on tumor rate (simulation conditions 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34 in Table 12.5). The tumor types were classified as rare or common based on the incidence rate of the concurrent control. The results of this simulation study show a very interesting pattern in levels of attained Type 1 error. The attained levels of Type 1 error under various simulation conditions divided into two groups. The division was by the factor of background rate, either 20 or 5 %. The attained Type1 levels of the first group were around 0.005. The attained Type 1 error rates for the second group were around 0.015. The observed results and pattern of the attained Type 1 errors make sense. For the simulated conditions with 20 % background rate, probably almost all of the 10,000

Table 12.5 Estimated retention rates under three decision rules

Simulation condition	Simulation condition properties				Retention probabilities			% Change in retention rate	
	Dose effect on mortality	Tumor appearance time	Dose effect on tumor prevalence	Tumor background rate	Trend test only	Trend test and H/C	Trend test and any pairwise	Trend test and H/C	Trend test and any pairwise
1	No	Early	No	0.05	0.984	0.9934	0.9919	0.9553	0.8028
2	No	Early	Small	0.05	0.6283	0.7084	0.6957	12.75	10.73
3	No	Early	Large	0.05	0.1313	0.178	0.1595	35.57	21.48
4	No	Late	No	0.05	0.9827	0.9927	0.9915	1.018	0.8955
5	No	Late	Small	0.05	0.6314	0.7208	0.7076	14.16	12.07
6	No	Late	Large	0.05	0.1408	0.2018	0.1811	43.32	28.62
7	No	Early	No	0.2	0.9953	0.9979	0.9974	0.2612	0.211
8	No	Early	Small	0.2	0.8377	0.8805	0.8715	5.109	4.035
9	No	Early	Large	0.2	0.3424	0.427	0.398	24.71	16.24
10	No	Late	No	0.2	0.9952	0.9972	0.9972	0.201	0.201
11	No	Late	Small	0.2	0.8399	0.8869	0.8772	5.596	4.441
12	No	Late	Large	0.2	0.3754	0.4864	0.4565	29.57	21.6
13	Small	Early	No	0.05	0.9855	0.9985	0.9978	1.319	1.248
14	Small	Early	Small	0.05	0.6967	0.8465	0.8324	21.5	19.48
15	Small	Early	Large	0.05	0.2152	0.4112	0.3574	91.08	66.08
16	Small	Late	No	0.05	0.9819	0.9991	0.9977	1.752	1.609
17	Small	Late	Small	0.05	0.722	0.9161	0.8903	26.88	23.31
18	Small	Late	Large	0.05	0.2682	0.6794	0.6021	153.3	124.5
19	Small	Early	No	0.2	0.9948	0.9996	0.9995	0.4825	0.4725
20	Small	Early	Small	0.2	0.8753	0.9694	0.9606	10.75	9.745

21	Small	Early	Large	0.2	0.4649	0.7564	0.711	62.7	52.94
22	Small	Late	No	0.2	0.9961	0.9999	0.9996	0.3815	0.3514
23	Small	Late	Small	0.2	0.8935	0.9939	0.9885	11.24	10.63
24	Small	Late	Large	0.2	0.538	0.9455	0.9095	75.74	69.05
25	Large	Early	No	0.05	0.9856	0.9994	0.9989	1.4	1.349
26	Large	Early	Small	0.05	0.8381	0.9587	0.948	14.39	13.11
27	Large	Early	Large	0.05	0.5358	0.8133	0.7796	51.79	45.5
28	Large	Late	No	0.05	0.9828	1	1	1.75	1.75
29	Large	Late	Small	0.05	0.8675	0.996	0.9886	14.81	13.96
30	Large	Late	Large	0.05	0.6447	0.9807	0.9428	52.12	46.24
31	Large	Early	No	0.2	0.994	1	1	0.6036	0.6036
32	Large	Early	Small	0.2	0.9414	0.9994	0.9985	6.161	6.065
33	Large	Early	Large	0.2	0.7445	0.9823	0.97	31.94	30.29
34	Large	Late	No	0.2	0.9956	1	1	0.4419	0.4419
35	Large	Late	Small	0.2	0.9585	1	0.9999	4.33	4.319
36	Large	Late	Large	0.2	0.835	0.9998	0.9989	19.74	19.63

Note: The estimated retention rates of the simulation conditions where the null hypothesis is true, (conditions 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34), are the probabilities of not committing a Type 1 error. For the remaining simulation conditions, the rates are the Type 2 error rates

generated datasets (each dataset containing tumor and survival data of four treatment groups of 50 animals each group) will have a tumor rate of equal to or greater than 1 % (the definition of a common tumor) in the control group. The attained levels of Type 1 error rates under various simulated conditions in this group are close to the nominal significance levels for common tumors.[3]

The attained Type 1 error rates for the other group were between the nominal levels of significance of 0.005 (for the trend test for common tumors) and 0.025 (for the trend test for rare tumors) and not around 0.005. The reason for this phenomenon is that, though the background rate in the simulated conditions for this group was 5 % that is considered as a rate for a common tumor, some of the 10,000 generated datasets had tumor rates less than 1 % in the control group. For this subset of the 10,000 datasets, the nominal level of 0.025 was used in the trend test. See Sect. 12.3.4.3 for a more detailed discussion of this factor.

As mentioned previously, the main objective of our study is the evaluation of the Type 2 error rate under various conditions. As was expected, the Type 2 error (or false negative) rates resulting from the joint test decision rule are higher than those from the procedure recommended in the guidance document of using trend test alone. This is due to the fact that in statistical theory the false positive rate (measuring the producer's risk in the regulatory review of toxicology studies) and the false negative rate (measuring the consumer's risk) run in the opposite direction; use of the joint test decision rule will cut down the former rate only at the expense of inflating the latter rate.

The estimated false negative rates resulting from the extensive simulation study under the three decision rules listed in Sect. 12.2.2 are shown in Table 12.5. The last two columns of the table show the percentage changes in the retention rates of decision rules (2) and (3) respectively, compared to those of (1). For those simulation conditions where the null hypothesis is true (1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34), these values measure the percentage change in the probability of not committing a Type 1 error. For the remaining simulation conditions, these values measure the inflation of the Type 2 error rate attributable to the adoption of the more stringent rules.

The magnitude of the inflation of false negative rate resulting from the joint test decision rule of requiring statistically significant results in both the trend test and the C-H (High versus Control) pairwise comparison test depends on all the four factors,

[3] It becomes more complicated in the evaluation of conservativeness or anti-conservativeness of the joint test (simultaneous combination of the trend test and the pairwise comparison test) under the agency practice. This is so because the two tests are not independent since the pairwise comparison tests used a subset (a half) of the data used in the trend test. Theoretically, if the trend test and the pairwise comparison test are actually independent and are tested at 0.005 and 0.01 levels of significance, respectively for the effect of a common tumor type, then the nominal level of significance of the joint tests should be $0.005 \times 0.010 = 0.00005$. Some of the levels of attained Type I error of the joint tests are larger than 0.00005 due to the dependence of the two tests that were applied simultaneously. To evaluate this nominal rate directly would require estimation of the association between the two tests. Results of the simulation study, as expected, show that the attained levels of type I error (1-retention probability under the simulation conditions in which there is no drug effect on tumor prevalence) are smaller than those of the trend test alone.

namely, drug effect on mortality, background tumor rate, time of tumor appearance, and drug effect on tumor incidence considered in the simulation that are also listed in the notes at the bottom of Tables 12.3 and 12.4.

12.2.4 Discussion

Results of the simulation study show that the factor of the effect of the dose on tumor prevalence rates has the largest impact on the inflation of the false negative rate when both the trend test and the C-H pairwise comparison tests are required to be statistically significant simultaneously in order to conclude that the effect is statistically significant. The inflations are most serious in the situations in which the dose has a large effect on tumor prevalence. The inflation can be as high as 153.3 %. The actual Type 2 error rate can be more than double.

The above finding is the most alarming result among those from our simulation study. When the dose of a new test drug has large effects on tumor prevalence (up to 28 % difference in incidence rates between the high dose and the control groups), it is a clear indication that the drug is carcinogenic. Exactly in these most important situations the joint test decision rule causes the most serious inflations of the false negative error rate (or the most serious reductions in statistical power to detect the true carcinogenic effect). The net result of this alarming finding is that using the levels of significance recommended for the trend test alone and the pairwise test alone in the joint test decision rule can multiply the probability of failure to detect a true carcinogenic effect by a factor up to two or even more, compared with the procedure based on the result of the trend test alone.

It is true that the results in Table 12.5 show that, for the situations in which the dose has a small effect (up to 15 % difference in incidence rates between the high dose and the control groups) on tumor prevalence, the increases of false negative rates caused by the joint test decision rule are not much more than those from using the trend test alone (increases can be up to 27 %). However, this observation does not imply that the joint test decision rule is justified. The reason is that standard carcinogenicity studies use small group sizes as a surrogate for a large population with low tumor incidence. There is very little power (i.e., the false negative rates are close to 100 %) for a statistical test using a small level of significance such as 0.005 to detect any true carcinogenicity effect because of low tumor incidence rates. In those situations there will be little room for the further increase in the false negative rate no matter how many additional tests are put on top of the original trend test.

It might be argued that the large inflations in false negative rates in the use of the joint test over the use of the trend test alone could be due to the large dose effect on death (only 4 % animals alive at 2 years introduced by the additional treatment group with $x = 3$. The argument may sound valid since as mentioned previously, the decrease of percentage of animals alive at 2 years increases the false negative rates. We are aware of the small number of alive animals at 2 years in the simulation

condition of large dose effect on death caused by using the built-in Weibull model described in Dinse (1985) and by including the additional group with $x = 3$ in the our study.

However, as mentioned previously, our main interest in this study is to evaluate the percentages of inflation in false negative rates attributable to the use of the joint test compared with the trend test alone. The false negative rates in the joint test and in the trend test alone are certainly also of our interest. They are not the main interest. So the issue of excessive mortality under the simulation condition of a large dose effect on death should not be a major issue in this study since it has similar impacts on false negative rates in both the joint test and the trend test alone. Furthermore, it is seen from Table 12.5 that the largest inflations (63–153 %) in the false negative rate happened under the conditions in which the dose effect on death is small (30 % alive animals at 2 years) rather than under the condition in which the effect is large (4 % alive animals at 2 years).

The extremely large false negative rates in the above simulated situations caused by the nature (low cancer rates and small group sample sizes) of a carcinogenicity experiment, reinforce the important arguments that it is necessary to allow an overall (for a compound across studies in two species and two sexes) false positive rate of about 10 % to raise the power (or to reduce the false negative rate) of an individual statistical test. This important finding of the simulation study clearly supports our big concern about failing to detect carcinogenic effects in the use of the joint test decision rule in determining the statistical significance of the carcinogenicity of a new drug. Again, the producer's risk using a trend alone is known at the level of significance used (0.5 % for a common tumor and 2.5 % for a rare tumor in a two-species study) and is small in relation to the consumer's risk that can be 100 or 200 times the level of the known producer's risk. The levels of significance recommended in the guidance for industry document were developed with the consideration of those situations in which the carcinogenicity experiment has great limitations. Trying to cut down only the producer's risk (false positive rate in toxicology studies) beyond that which safeguards against the huge consumer's risk (false negative rates in toxicology studies) is not consistent with the FDA mission as a regulatory agency that has the duty to protect the health and well-being of the American general public.

As mentioned previously, the decision rules (levels of significance) recommended in US Food and Drug Administration—Center for Drug Evaluation and Research (2001) are for trend tests alone and for pairwise comparisons alone, and not for the joint test. To meet the desire of some non-statistical scientists within the agency to require statistically significant results for both the trend test and the C-H pairwise comparison simultaneously to conclude that the effect on the development of a given tumor/organ combination as statistically significant, and still to consider the special nature of standard carcinogenicity studies (i.e., using small group sizes as a surrogate of a large population with low a tumor incidence endpoint), we have conducted additional studies and proposed new sets of significance levels for a joint test along with some updates of the previously recommended ones. These are presented in Table 12.6. We have found that the use of these new levels keeps the overall false positive rate (for the joint test) to approximately 10 % again for a compound across studies in two species and two sexes.

Table 12.6 Recommended decision rules (levels of significance) for controlling the overall false positive rates for various statistical tests performed and submission types

Submission type		Tumor type	Decision rule			
			Trend test alone	Pairwise test alone	Joint test	
					Trend test	Pairwise test
Standard 2 year study with two sexes and two species		Common	0.005	0.01	0.005	0.05
		Rare	0.025	0.05	0.025	0.10
Alternative ICH Studies (One 2-year study in one species and one short- or medium-term alternative study, two sexes)	Two-year study	Common	0.005	0.01	0.005	0.05
		Rare	0.025	0.05	0.025	0.10
	Short- or medium-term alternative study	Common	0.05	0.05	0.05	0.05
		Rare	0.05	0.05	0.05	0.05
Standard 2 year studies with two sexes and one species		Common	0.01	0.025	0.01	0.05
		Rare	0.05	0.10	0.05	0.10

12.3 The Relationship Between Experimental Design and Error Rates

In this section, we describe the results of a second simulation study. The aim of the simulation study discussed in Sect. 12.2 was to compare decision rules, evaluating the impact on error rates of the adoption of the joint test rule (which is more conservative than the trend test rule recommended in US Food and Drug Administration—Center for Drug Evaluation and Research (2001)—see Tables 12.1 and 12.2). By contrast, this second study, which was conducted independently, compares the effects of the use of different experimental designs on the error rates, all under the same decision rule. The decision rule used in this study is the joint test rule (Table 12.2) since, despite the absence of any theoretical justification for its use, this rule is currently used by non-statistical scientists within the agency as the basis for labeling and other regulatory decisions.[4]

We first consider the nature of the various hypotheses under consideration, and the associated error rates (Sect. 12.3.1). This provides us with the terminology to express our motivation (Sect. 12.3.2). We then describe in detail the four designs that have been compared (Sect. 12.3.3), and the simulation models used to test these designs (Sect. 12.3.4). The results of the simulation are discussed in Sects. 12.3.5 (power) and 12.3.6 (Type 1 error). We conclude with a brief discussion (Sect. 12.3.7).

12.3.1 Endpoints, Hypotheses, and Error Rates

The task of analyzing data from long term rodent bioassays is complicated by a severe multiplicity problem. But it is not quite the case that we are merely faced with a multitude of equally important tumor endpoints. Rather, we are faced with a hierarchy of hypotheses.

- At the lowest level, we have individual tumor types, and some selected tumor combinations, associated with null hypotheses of the form:

  > Administration of the test article is not associated with an increase in the incidence rate of *malignant astrocytomas of the brain* in *female rats*.

  We call such hypotheses the *local null hypotheses*.
- The next level of the hierarchy is the *experiment* level. A standard study includes four experiments: on male mice, on female mice, on male rats, and on female rats. Each of these experiments is analyzed independently, leading to four *global null hypotheses* of the form:

[4]See, for instance, the discussion of osteosarcomas and osteomas in female mice in Center for Drug Evaluation and Research (2103).

> There is no organ–tumor pair, or reasonable combination of organ-tumor pairs, for which administration of the test article is positively associated with tumorigenesis in *male mice*.

Note that some studies consist of just two experiments in a single species (or, very rarely, in two species and a single sex).

- The highest level of the hierarchy of hypotheses is the *study* level. There is a single study-wise null hypothesis:

> For none of the experiments conducted is the corresponding global null hypothesis false.

For any given local null hypothesis, the probability of rejecting that null hypothesis is called either the *local false positive rate* (LFPR) or the *local power*, depending on whether the null hypothesis is in fact true. If all the local null hypotheses in a given experiment are true, then the *global false positive rate* (GFPR) for that experiment is the probability of rejecting the global null hypothesis, and can be estimated from the various estimates for the LFPRs for the endpoints under consideration.[5] The goal of the multiplicity adjustments in US Food and Drug Administration—Center for Drug Evaluation and Research (2001) is to maintain the study-wise false positive rate at about 10 %. Since most studies consist of four independent experiments, we consider our target level for false positives to be a GFPR of approximately 2.5 %.[6]

The calculation of a GFPR from the LFPR depends on the relationship between the local and global null hypotheses. We capture this relationship with the notion of a tumor *spectrum*: If $\mathcal{T}$ is the parameter space for tumor types, then a spectrum is a function $S : \mathcal{T} \to \mathbb{N}$; $S(t)$ is the number of independent tumor types being tested with parameter value t. In our case, $\mathcal{T}$ is one dimensional: under the global null hypothesis we assume that each tumor can be characterized by its background prevalence rate.[7]

In our simulations, we generate estimates for the power and LFPR for three different classes of tumor:

1. *Rare* tumors have a background prevalence rate (i.e., the lifetime incidence rate among those animals who do not die from causes unrelated to the particular tumor type before the end of the study, typically at 104 weeks) of 0.5 %.

[5]This calculation assumes that all endpoints are independent. This assumption is not strictly true, especially when considering combinations of endpoints. However, it is reasonable to assume that the endpoints are close enough to being independent that the resulting estimate of the GFPR is accurate enough for our purposes.

[6]In fact, if four independent experiments are to have a combined study-wise false positive rate of 10 %, then it suffices for them individually to have GFPRs of $1 - (1 - 0.1)^{1/4} = 0.026$. However, since it is not practical to calibrate the GFPR so precisely, there is no practical distinction between target GFPRs of 0.025 and 0.026.

[7]More sophisticated models might treat $\mathcal{T}$ as higher dimensional. For example, in the simulation study in Sect. 12.2, the parameter space $\mathcal{T}$ is two dimensional, with the two dimensions representing the background prevalence rate and the tumor onset time. (Although Eq. (12.1) has *three* independent parameters (not counting the dose response parameter D), the parameters A and B are not varied independently—see Table 12.3.)

2. *Common* tumors have a background prevalence rate of 2 %.
3. *Very common* tumors have a background prevalence rate of 10 %.

A tumor spectrum for us therefore consists of a triple $\langle n_1, n_2, n_3 \rangle$, indicating that the global null hypothesis is the conjunction of $n_1 + n_2 + n_3$ local null hypotheses, and asserts the absence of a treatment effect on tumorigenicity for n_1 rare, n_2 common, and n_3 very common independent tumor endpoints.

Given such a spectrum, and under any given set of conditions, the GFPR is easy to calculate from the LFPR estimates for the three tumor types under those conditions:

$$\text{GFPR} = 1 - \prod_{i=1}^{3} (1 - F_i)^{n_i} \tag{12.3}$$

where F_i is the estimated LFPR for the i-th class of tumors. Since our desired false positive rates are phrased in terms of the study-wise false positive rate (which we want to keep to a level of approximately 10 %), we are more concerned with the GFPR than the LFPR.

Global power is slightly harder to calculate, since it is a function of a specific global alternate hypothesis. It is unclear what a realistic global alternative hypothesis might look like, except that a global alternative hypothesis is likely to be the conjunction of a very small number of local alternative hypotheses with a large number of local null hypotheses. Accordingly, we focus our attention on the local power.

In summary then, the two quantities that we most wish to estimate are the local power and the GFPR.

12.3.2 Motivation

For any given experimental design, there is a clear and well understood trade-off between the Type 1 rate (the false positive rate) and the Type 2 error rate (1 minus the power): by adjusting the rejection region for a test, usually by manipulating the significance thresholds, the test can be made more or less conservative. A more conservative test has a lower false positive rate, but only at the expense of a higher Type 2 error rate (i.e., lower power), while a more liberal test lowers the Type 2 error rate at the cost of raising the Type 1 error rate. Finding an appropriate balance of Type 1 and Type 2 errors is an important part of the statistical design for an experiment, and requires a consideration of the relative costs of the two types of error. It is generally acknowledged (see Center for Drug Evaluation and Research 2005) that for safety studies this balance should prioritize Type 2 error control.

However, this trade-off applies only to a fixed experimental design; by adjusting the design, it may be possible to simultaneously improve both Type 1 *and* Type 2

error rates.[8] Beyond this general principle, there is a particular reason to suspect that adjusting the design might affect error rates for carcinogenicity studies. It has been shown (Lin and Rahman 1998; Rahman and Lin 2008) that, using the trend test alone and the significance thresholds in Table 12.1, the study-wise false positive rate for rodent carcinogenicity studies is approximately 10 %. However, under this decision rule, the nominal false positive rate for a single rare tumor type is 2.5 %. Given that each study includes dozens of rare endpoints, the tests must be strongly over-conservative for rare tumor types[9]; the decision rules in US Food and Drug Administration—Center for Drug Evaluation and Research (2001) rely heavily on this over-conservativeness in order to keep the GFPR to an acceptable level. But this sort of over-conservativeness is exactly the sort of phenomenon that one would expect to be quite sensitive to changes in study design.

12.3.3 Designs Compared

To get a sense of the designs currently in use, we conducted a brief investigation of 32 recent submissions, and drew the following general conclusions:

- While most designs use a single vehicle control group, a substantial proportion do use two duplicate vehicle control groups.
- A large majority of designs use three treated groups.
- The total number of animals used can vary considerably, but is typically between 250 and 300.
- The "traditional" design of four equal groups of 50 is still in use, but is not common; most designs use larger samples of animals.

Bearing these observations in mind, we compare four designs, outlined in Table 12.7. Three of these designs (D1–2 and D4) utilize the same number of animals (260), so that any effects due to differences in the disposition of the animals will not be obscured by differences due to overall sample size.

[8]It is a familiar result for asymptotic tests that simply increasing the sample size improves power while maintaining the Type 1 error rate at the nominal level. (This principle also applies to rare event data except that exact tests are frequently over-conservative, meaning that increases in the sample size can actually increase the Type 1 error rate even while keeping the rate below the nominal level, and that power can sometimes decrease as the sample size increases—see Chernick and Liu 2002). However, the inclusion of large numbers of extra animals is an inelegant (and expensive) way to shift the ROC curve; we are interested in modifications to the experimental design that leave the overall number of animals unchanged.

[9]The observation that the trend tests is strongly over-conservative for rare tumors is not at odds with the finding in Dinse (1985) that the trend test is not over-conservative for tumors with a background prevalence rate of 5 or 20 %. As the expected number of tumors increases, one expects exact tests to converge to the asymptotic tests, and the LFPRs to converge to the nominal value of α.

Table 12.7 Experimental designs considered

	Number of animals per group				
Design number	Control	Low	Mid	High	Total
D1	65	65	65	65	260
D2	104	52	52	52	260
D3	50	50	50	50	200
D4	60	50	100	50	260

The first two designs, D1 and D2, are representative of designs currently in use. Design D1 uses four equal groups of 65 animals whereas design D2 uses a larger control group (104 animals) and three equal dose groups (52 animals each). This is equivalent to a design with five equal groups, comprising two identical vehicle control groups (which, since they are identical, may be safely combined) and three treated groups.

The third design tested (D3) is the "traditional" 200 animal design. Although D3 uses fewer animals than the other designs (but is otherwise similar to D1), it has been included to enable comparison with the many simulation studies and investigations which use this design, such as that described in Sect. 12.2, and in Dinse (1985), Portier and Hoel (1983), Lin and Rahman (1998), and Rahman and Lin (2008)

In light of the investigation (Jackson 2015) of the possible benefits of unbalanced designs (where the animals are not allocated equally to the various dose groups), we have also included an unbalanced design for comparison. This design (D4) follows the suggestions of Portier and Hoel (1983):

> ... we feel that a design with 50 to 60 of the experimental animals at control ($d_0 = 0$), 40 to 60 of the animals at the MTD ($d_3 = 1$) and the remaining animals allocated as one-third to a group given a dose of 10–30 % MTD ($d_1 = 0.25$ seems best) and two-thirds to a group given a dose of 50 % MTD ($d_2 = 0.5$). No less than 150 experimental animals should be used, and more than 300 animals is generally wasteful. An acceptable number of animals would be 200.

Accordingly, 60 animals have been allocated to the control group and 50 to the high dose group, with the remaining 150 animals allocated 2:1 to the mid and low dose groups.

12.3.4 Statistical Methodology

12.3.4.1 Simulation Schema

We have conducted two separate simulation studies. The first study was designed to compare the (local) power of the four designs to detect genuine increases in tumorigenicity for the three tumor types (rare, common, and very common) described in Sect. 12.3.1. In each case, about fifty different effect sizes (measured as the odds ratio for tumor incidence between a high dose and control animals at 104 weeks) were tested 1000 times. While 1000 simulations are not adequate to

accurately estimate the power for a particular effect size (we can expect a margin of error in the estimate of approximately 3 %), we may still form an accurate impression of the general shape of the power curves.

The second simulation study was aimed at evaluating false positive (Type 1 error) rates. The immediate focus was on the LFPR, the rate at which individual organ-tumor endpoints for which there is no genuine effect are falsely found to be targets of a carcinogenic effect. Because local false positives are very rare, and because imprecision in the estimate of the local false positive rate is amplified when computing the global false positive rate, each simulation scenario has been repeated at least 250,000 times. The resulting estimates are amalgamated to compute the GFPR by appealing to independence and applying Eq. (12.3) to three different tumor spectra.

For both of the simulation studies, data were simulated using a competing risks model. The two competing hazards were tumorigenesis and death due to a non-tumor cause.

- Since these simulations are intended to evaluate power and GFPRs under fairly optimal circumstances, only one toxicity model has been considered: the hazard function for non-tumor death has the form $h_M(t) = \lambda t(\mu x + 1)$, where x is the dose and t is the time. The parameters λ and μ are chosen so that the probabilities of a control animal and a high dose animal experiencing non-tumor death before the scheduled termination date are 0.4 and 0.7 respectively.
- Tumor onset time is modeled according to the poly-3 assumptions. This means that for any given animal, the probability of tumorigenesis before time t has the form $P[T \leq t] = \lambda t^3$ where the parameter λ is a measure of the animal's tumor risk, and so depends on the dose x, the background prevalence rate (i.e., the tumor incidence rate when $x = 0$), and the dose effect on tumorigenesis to be simulated. (In the case of the LFPR simulations, it is assumed that there is no dose effect on tumorigenesis, and λ therefore depends on the background prevalence rate alone).

Although these simulations were devised independently of those in Sect. 12.2, the resulting models are in practice quite similar. Tumor onset times modeled by this approach are very similar to those of the "early onset" models, although non-tumor mortality times tend to be earlier than those simulated in Sect. 12.2. The effect of this difference is likely to be a small reduction in power (and LFPRs) in the present model compared with those used in Sect. 12.2.

12.3.4.2 Decision Rule

As noted above, we are initially concerned with estimating local power and LFPRs. Accordingly, under each scenario, we simulate data for a single 24 month experiment (*male mice*, for example), and a single tumor endpoint (*cortical cell carcinoma of the adrenal gland*, for example). Each set of simulated data includes a death time for each animal and information about whether the animal developed a

tumor. From these data, two poly-3 tests (see Sect. 12.4 and Bailer and Portier 1988; Bieler and Williams 1993; US Food and Drug Administration—Center for Drug Evaluation and Research 2001) are conducted: a trend test across all groups, and a pairwise test between the control and high dose groups. As we are using the joint test rule (discussed at length in Sect. 12.2.1), the null hypothesis of no tumorigenic effect is rejected only when *both* the trend and pairwise tests yield individually significant results, at the levels indicated in Table 12.2.

12.3.4.3 Misclassification

The use of the observed incidence rate in the control group to classify a tumor as rare or common is potentially problematic. There is clearly a substantial likelihood that common tumors (with a background prevalence rate of 2 %) will be misclassified as rare, and judged against the "wrong" significance thresholds. Given the difference in the significance thresholds between those used for rare and for common tumors, it is to be expected that this misclassification effect could have an appreciable liberalizing effect on the decision rules used. Furthermore, this liberalizing effect will be amplified by the fact that misclassification is positively associated with low p-values.[10] This effect was noted, discussed, and even quantified (albeit for different decision rules and simulation scenarios than those used here) in Westfall and Soper (1998).

The probability of misclassification is dependent on both the background prevalence rate of the tumor and the number of animals in the control group. Since D2 has more than 100 animals in the control group, an experiment using this design will treat a particular tumor endpoint as rare if there is no more than one tumor bearing control animal; the other designs will consider an endpoint to be rare only if no tumor bearing control animals are found at all. The effects of this difference are seen in Fig. 12.2.

Under more traditional circumstances, given an exact test procedure and a fixed set of significance standards, one would expect the false positive rate to increase asymptotically to the nominal level as the expected event count increased. However, given the misclassification effect in this context, we expect something different; for tumors with a 2 % background prevalence rate (and for those with a 5 % background rate in the simulation study in Sect. 12.2), the LFPR can be anticipated to converge to a value below the nominal significance level for rare tumors, but *above* the nominal significance level for common tumors. For tumors with a 10 % background prevalence rate, by contrast, we can expect the LFPR to be somewhat closer to the nominal value for common tumors.[11]

[10] For this reason, the converse effect is of much less concern; misclassification of rare tumors as common is positively associated with large p-values so the cases where misclassification occurs are unlikely to be significant at even the rare tumor thresholds.

[11] The actual nominal value of the joint test is hard to evaluate. See footnote 3 on page 306.

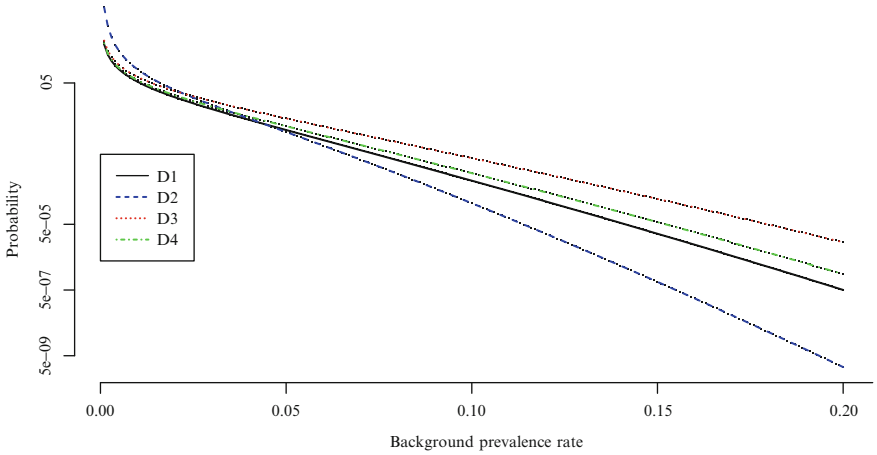

Fig. 12.2 Probability of classifying a tumor type as rare

It is uncertain whether the differential effect of misclassification on the four designs should be viewed as intrinsic to the designs, or an additional, unequal source of noise. However, given the paucity of relevant historical control data (Rahman and Lin 2009), and that a two tiered decision rule is in use, there seems to be little alternative to this method for now.[12] Accordingly, this is the most commonly used method for classifying tumors as rare or common, and we have elected to treat it as an intrinsic feature of the statistical design.

Nonetheless, it should also be remembered that statistical analysis is only one stage in the FDA review process, and that pharmacology and toxicology reviewers are free to exercise their professional prerogative and overturn the empirical determination. This is especially likely for the rarest and commonest tumors. More generally though, it is apparent that this misclassification effect must be taken into account when designing, conducting, and interpreting any simulations to evaluate carcinogenicity studies.

12.3.5 Power

The results of the power simulations are shown in Fig. 12.3.

Designs D1 and D2 are clearly more powerful than D3 and D4. For very common tumors, there is little difference between the two, but for both rare and common tumors, design D2 appears to be appreciably more powerful than D1. For rare

[12] Although it is to be hoped that in the longer term the use of the SEND data standard (Clinical Data Interchange Standards Consortium (CDISC) 2011) will enable the more efficient construction of large historical control databases.

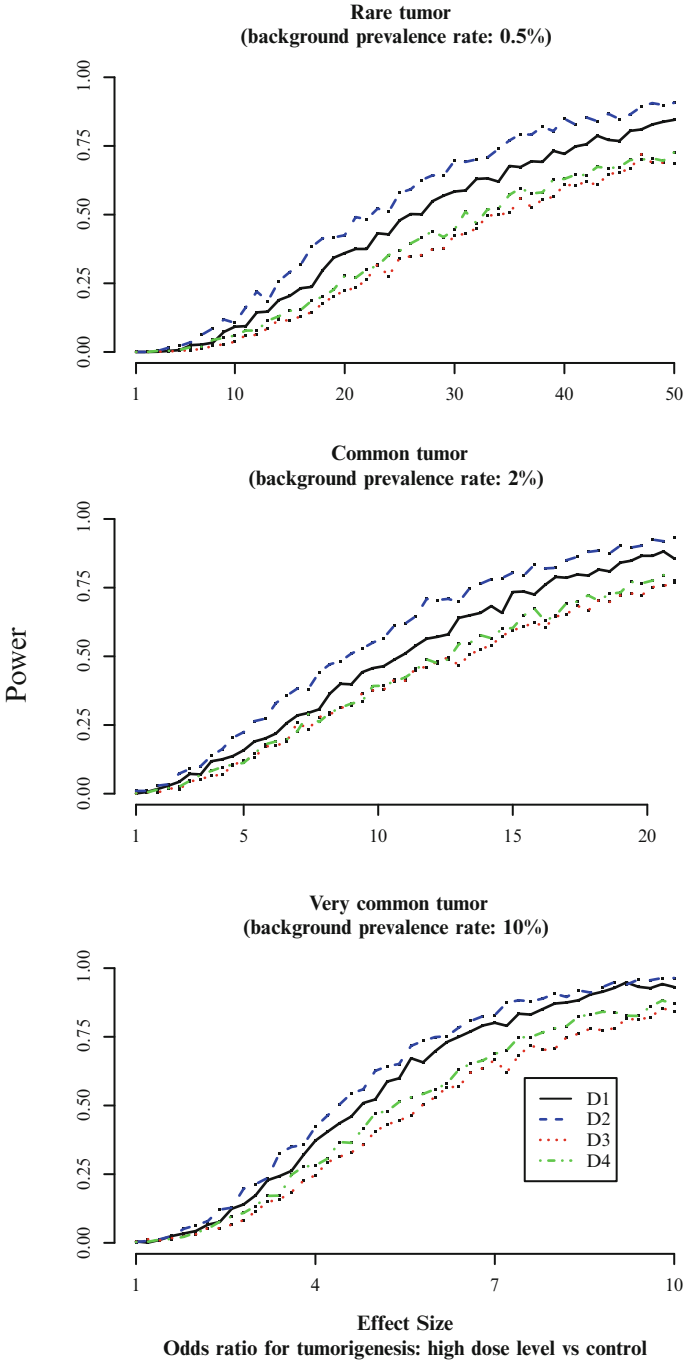

Fig. 12.3 Estimated power

tumors and an effect size of 30 (corresponding to a risk difference (RD) between the control group and the high dose group of 12.6 %), D1 and D2 have approximately 60 and 70 % power respectively. For common tumors and an effect size of 10 (RD=14.9%), D1 and D2 have approximately 45 and 55 % power respectively. More generally, Fig. 12.3 suggests that design D2 delivers about 10 % more power than D1 across a fairly wide range of scenarios of interest.

Since it uses the fewest animals, it is not surprising that D3 is the least powerful of the four designs. Direct comparison of D3 and D1 (which are similar except for the fact that D1 uses 30 % more animals in each group) shows the benefit in power that an increased sample size can bring.

That said, it is striking that the design D4, with 260 animals, is barely more powerful than D3. As we have seen from our comparison of D1 with D3, adding animals in such a way that the groups remain equal in size does increase the power of the design (absent any sawtooth effects). Furthermore, adding animals unequally to the groups can improve the power even more (see Jackson 2015). However, in the case of design D4, the extra animals (compared with D3) were not added with the goal of improving power. Indeed, the notion of optimality which D4 is intended to satisfy is quite different from our narrow goal of maximizing power while keeping the GFPR to approximately 2.5 % (Portier and Hoel 1983):

> For our purposes, an optimal experimental design is a design that minimizes the mean-squared error of the maximum likelihood estimate of the virtually safe dose from the Armitage-Doll multistage model and maintains a high power for the detection of increased carcinogenic response.

In addition, the intended maintenance of "a high power for the detection of increased carcinogenic response" was predicated on a decision rule using the trend test alone, with a significance threshold of 0.05—a much more liberal testing regime even than that recommended in US Food and Drug Administration—Center for Drug Evaluation and Research (2001), let alone than the more conservative joint test rule used here.

It is worth noting that the unbalanced approach of Design D4 is almost the anthesis of that proposed in Jackson (2015); in the latter, power is maximized by concentrating animals in the control and high dose groups, whereas in D4 they are concentrated in the intermediate groups.

12.3.6 False Positive Rate

12.3.6.1 The Local False Positive Rate

For each design, at least 250,000 simulations were conducted to estimate the rate at which the local null hypothesis is rejected when the tumor hazard is unchanged across dose groups. The resulting estimates of the LFPR (with 95 % confidence intervals) for each of the four designs and three tumor types (rare, common, and very common) are shown in Table 12.8.

Table 12.8 Local false positive rates (%) with 95 % confidence intervals

	Background prevalence rate					
Design	0.5 % (rare)		2 % (common)		10 % (very common)	
D1	0.0024	(0.0009,0.0052)	0.2264	(0.2078, 0.2450)	0.3000	(0.2786, 0.3214)
D2	0.0424	(0.0343,0.0505)	0.5928	(0.5627, 0.6229)	0.3792	(0.3551, 0.4033)
D3	0.0012	(0.0002,0.0035)	0.1008	(0.0884, 0.1132)	0.4528	(0.4265, 0.4791)
D4	0.0011	(0.0000,0.0024)	0.1048	(0.0926, 0.1169)	0.2070	(0.1899, 0.2241)

Table 12.9 Probability of misclassification

	Lifetime tumor incidence rate		
Design	0.5 % (rare)	2 % (common)	10 % (very common)
D1	0.278	0.269	0.001
D2	0.096	0.382	<0.001
D3	0.222	0.364	0.005
D4	0.260	0.298	0.002

Generally speaking, we can expect two aspects of design to affect the LFPRs. As the sample size increases, the expected number of tumors also increases, which in turn means that exact tests will behave more like asymptotic tests. In particular, we can expect the LFPRs to converge to the nominal α-level as the sample size increases. Since the tests are exact, this means that the LFPRs will tend to increase (with sample size), and since we know that the tests are strongly over-conservative for design D3 (see Sect. 12.3.2) we know that there is considerable room for growth from the levels associated with this design.

The second effect will tend to act in the opposite direction. The determination of significance thresholds depends on the number of tumor bearing animals (TBAs) in the control group: for a design with fewer than one hundred control animals, the tumor type will be considered rare just in the case that no control animals develop the tumor. (see Sect. 12.3.4.3). Thus, increasing the number of control animals (while keeping this number below 100) will increase the likelihood of a tumor type being classified as common. Since the significance thresholds for common tumors are lower than for rare tumors, this effect will make designs with more animals *more* conservative. (This reasoning does not apply to D2 which has more than 100 control animals. Since for this design, at last two tumor bearing control animals must be found in order for a tumor type to be considered common, D2 is more likely than the other designs to classify tumors as rare.) See Table 12.9 to see tumor misclassification rates for the different designs.[13]

[13]For computational reasons, these calculations use the lifetime tumor incidence rate rather than the background prevalence rate used elsewhere in this chapter.

12.3.6.1.1 Comparison of D1 and D2

The most striking feature of Table 12.8 is the difference between D1 and D2. As discussed in Sect. 12.3.3, both of these designs are in regular use, and are treated similarly for analysis purposes. However, D2 is prone to far higher false positive rates than D1, raising doubts about whether it is reasonable to treat the two designs interchangeably.

For rare tumor types, the LFPR rate for D1 is so low as to be almost negligible (approximately 1 in 40,000). Even if there are 200 such endpoints, their combined false positive rate (12.3) will be under 0.5 %. As a result, genuinely rare tumor endpoints do not contribute much to the GFPR under this design. The LFPR for D2 is about 17 higher than that for D1. While still small in absolute terms, and still strongly over-conservative, this rate is high enough that it would only take a relatively small number of rare endpoints to generate an unacceptably high combined false positive rate. In other words, despite being over-conservative, the LFPR for D2 is not over-conservative enough to meet our goals for the GFPR. See Sect. 12.3.6.2 for a more detailed discussion.

Compounding this problem, the LFPR for common tumors under D2 is exceedingly high; almost 0.6 %. This level is actually above the nominal rate for the trend test for common tumors (0.5 %), and so must be at least partially attributable to the misclassification effect noted above. The LFPR for common tumors under D1 is much lower (about 40 % of the rate under D2), but still far higher than that for rare tumors.

The LFPRs for very common tumors are much closer to each other than the rates for less common tumors (although D1 still has a higher LFPR than D1). This confirms the idea that D2's very high LFPR for common tumors is largely due to misclassification, as this misclassification effect would be expected to be less pronounced for the very common tumors (see Table 12.9).

12.3.6.1.2 Comparison of D1 and D3

For both rare and common tumors, D1 has a considerably higher LFPR than D3; about twice as high in both cases. This difference is to be expected, although it is still unsettling that an increase of just 30 % in the sample size (from D3) can result in a doubling of the LFPRs. As the background rate of tumors increases, the difference in over-conservativeness of the designs becomes less influential than D3's greater tendency toward misclassification as rare, so that D3 actually exhibits a higher LFPR than D1 for very common tumors (see Fig. 12.2).

12.3.6.1.3 Comparison of D3 and D4

The designs D3 and D4 have comparable LFPRs for rare and common tumors, although D4 has a noticeably lower LFPR for very common tumors than D3.

This is somewhat surprising since D4 was not proposed with the specific goal of constraining the Type 1 error rate. Overall, we can say that of the four designs considered, D4 has the lowest false positive rates.

12.3.6.2 The Global False Positive Rate

As discussed in Sect. 12.3.1, a tumor spectrum must be selected before the GFPR may be estimated. However, the selection of a realistic tumor spectrum is difficult. The endpoints under consideration are the *potential* tumor types which a pathologist may report if they are found. This list is not fixed from study to study. For example, one pathologist might group tumors found in the cervix or uterus together, whereas another might report these separately. Similarly, one pathologist might group hibernomas (lipomas of brown adipose tissue) with classic (white adipose tissue) lipomas under the general heading *lipoma—adipose tissue*, whereas another might not. Nonetheless, these variations are relatively minor, and the requirement that pathologists conduct a "complete necropsy," the widely accepted suggestions of Bergman et al. (2003), and the forthcoming adoption of the SEND data standard (Clinical Data Interchange Standards Consortium (CDISC) 2011) all ensure a reasonable level of uniformity.

However, the data from individual studies only reference tumor types which were actually found in those studies, so that many rare tumor endpoints are not mentioned at all in study reports. The data presented in the compilations by Charles River Laboratories (Giknis and Clifford 2004, 2005) are helpful in this regard. On the basis of the data in these tabulations, we have chosen three tumor spectra (shown in Table 12.10) which seem to form a plausible range for a typical tumor spectrum.[14]

For two of the three spectra, the GFPR for D1 is close to the target level of 2.5 %, and even for S3, it is "only" twice the desired level. However, the GFPRs for D2 consistently exceed the target GFPR by a very wide margin. The GFPR for D3 is close to the target, confirming the results of Sect. 12.2. The conservativeness of D4 carries over to the calculation of the GFPR which is actually below the target level for the thinner spectra S1 and S2, and close to 2.5 % for S3.

Table 12.10 Estimated global false positive rates

Spectrum number	Endpoints				Global FPR (%)			
	0.5 %	2 %	10 %	Total	D1	D2	D3	D4
S1	66	6	3	75	2.39	7.23	2.03	1.31
S2	100	6	4	110	2.76	8.91	2.51	1.56
S3	140	15	5	160	5.13	15.43	3.87	2.73

[14]Insofar as we know what is typical. These spectra should only be taken to represent a range of plausible scenarios, and not assumed to be in any way definitive.

Some care must be taken when interpreting these results. The LFPR for a tumor type with a true background prevalence rate of, say, 0.1 % is likely to be somewhat lower than that for a tumor with a background prevalence rate of 0.5 % (our rare tumor exemplar). This is unlikely to affect the estimates of the GFPRs for D1, D3, and D4, since the LFPR for rare tumors is so low for these designs that the GFPR is largely insensitive to the number of rare endpoints in the spectrum. This is not true of D2, though. A spectrum of 100 independent tumor types, half with a background prevalence rate of 0.1 % and half with a rate of 0.5 % might yield an appreciably lower GFPR rate under D2 than a spectrum of 100 independent tumor types each with a rate of 0.5 %. However, this effect can only explain a small part of the difference in GFPR between D1 and D2. For example, even for S2, with just 10 non-rare endpoints, the 6 common and 4 very common endpoints between them contribute over 40 % of the GFPR: if *all* the rare endpoints were disregarded for D2, the GFPR for S2 would still be 3.9 %; above the 2.5 % target and higher than the GFPR of any of the other designs.

12.3.7 Discussion

12.3.7.1 Comparison of Currently Used Designs

There are clearly substantial differences between designs D1 and D2. Figure 12.3 shows that D2 has appreciably more power than D1 over a range of meaningful scenarios—in many cases close to 10 % more—although the difference is slight for very common tumors. But as demonstrated in Tables 12.8 and 12.10, D2 also suffers from a considerably worse false positive rate than D1. Taken together, these two observations lead to the conclusion that D2 is *more liberal* than D1.

It is reasonable to consider phasing out the use of D2 altogether. The duplicate control design was originally introduced to test for the effects of extra-binomial within-study variability between the two concurrent control groups (Baldrick and Reeve 2007; Haseman et al. 1986; US Food and Drug Administration—Center for Drug Evaluation and Research 2001); any significant differences between the control groups were an indication of a failure of experimental conduct, such as when two groups of cages are subject to different environmental conditions. However, such comparisons between control groups are clearly underpowered to detect such environmental effects, especially given how hard it is to detect even moderately strong *treatment* effects.

If D2 is to be retained as an acceptable design, the fact that it is more liberal than D1 needs to be taken into account. At the very least, it seems inappropriate to continue to use the same significance thresholds for designs D1 and D2.

Designs D1 and D3 are very similar, differing only in the total number of animals used; both designs distribute animals equally among the single vehicle and three dose groups. Differences between these designs' statistical properties are therefore entirely attributable to differences in sample size. D1 is appreciably more powerful

than D3 across a wide range of scenarios, but also suffers from slightly higher false positive rates. In the case of D3, the GFPR tends to be well below the target rate of 2.5 %, confirming the findings of Sect. 12.2 that there is room to use more liberal significance thresholds (see Table 12.6), and thereby increase power. However, this does not appear to be the case for the more widely used D1 design.

In general, the differences between the three designs' statistical properties are not negligible, meaning that our ability to draw conclusions about one the behavior of design from the study of the other is limited. This is especially problematic since a great deal of our understanding of rodent carcinogenicity studies has its foundations in studies of design D3, but this design is fading from popularity.

12.3.7.2 The Design D4

The unbalanced design D4 was not optimized to maximize power using the sort of decision rule currently in place, and so it is not surprising that it is substantially less powerful than the other 260 animal designs. However, it is striking that the distribution of animals does yield a very low false positive rate (comparable to D3, although better for very common tumors). Compared with the traditional 200 animal design, then, the addition of 60 animals to D4 yields a real but modest benefit in both power and Type 1 error. It is reasonable to think that these extra animals instead provide considerable added information about the virtually safe dose (that being the intent behind this design), but testing this notion is outside the scope of this study.

12.3.7.3 The Balance Between Type 1 and Type 2 Error

The goal of achieving satisfactory power whilst keeping the GFPR to approximately 2.5 % is difficult to achieve, even aside from the ambiguity over what exactly constitutes "satisfactory power".

Inspection of Fig. 12.3 shows that for a tumor with a background prevalence rate of 0.5 %, design D1 delivers at least 50 % power when the effect size is above 27 (RiskDifference(RD) $=$ 11.4%), and 75 % when the effect size is above 42 (RD $=$ 16.9%). For a tumor with a background prevalence rate of 2 %, the power is above 50 % when the effect size is above 11 (RD $=$ 16.3%) and above 75 % when the effect size is above about 16 (RD $=$ 22.6%).

But even if we conclude that these levels of power are adequate, we are faced with the fact that the GFPR for D1 is generally somewhat above the target rate of 2.5 %. Lowering the significance thresholds to further limit the GFPR could only be done at the expense of the power, and so seems unwise, given that these studies are essentially safety studies.

In Jackson (2015), an alternative designs which delivers considerably more power than even D2 (the most powerful design considered here) while simultaneously lowering the GFPR rate to a level similar to or below that associated with D1 is investigated.

12.4 The Exact Poly-*k* Test

12.4.1 Introduction

The poly-*k* method is a mortality adjusted trend test for tumor incidence. This method was originally suggested in Bailer and Portier (1988), and was improved in Bieler and Williams (1993).

As some tumors may have long latency periods, animals with shorter life spans may face a disproportionately reduced risk of tumor onset. The poly-*k* method suggests correcting this problem by adjusting the number n_i of animals at risk in the *i*th dose group to compensate for early deaths. Operationally, the *j*th animal in the *i*th dose group gets a score $w_{ij} \leq 1$; this score is 1 if the animal lives for the full study period (T), or develops the tumor type being tested before dying. Conversely, if this animal dies at the time $t_{ij} < T$ before the end of the study without developing the tumor being tested, it gets a score of

$$w_{ij} = \left(\frac{t_{ij}}{T}\right)^k < 1.$$

The adjusted group size for Group *i* is then defined as

$$n_i^* = \sum_i w_{ij}.$$

As an interpretation, an animal with score $w_{ij} = 1$ can be considered as a whole animal, while an animal with score $w_{ij} < 1$ can be considered as a partial animal. The adjusted group size n_i^* is equal to n_i (the original group size) if all the animals in the group either survive until the end of the study or develops at least one tumor of the type being tested; otherwise the adjusted group size is less than n_i, except for some marginal cases due to rounding. These adjusted group sizes are then used to perform trend and pairwise (between treated groups and the control group) tests of tumor incidence rates using the Cochran-Armitage test procedure (Armitage 1955).

One critical point to consider when using the poly-*k* test is the choice of the appropriate value of *k*, which depends on the tumor incidence pattern with the increased dose. For long term 104 week standard rat and mouse studies, a value of k = 3 is suggested in the literature.[15] In this case we refer to the procedure as the *poly-3 test*. It should be noted that the assumption for Cochran-Armitage test is that the marginal total n_i is fixed. However, in this case n_i^* is a random variable. As a result the calculation of the variance of the test statistic needs to be modified. An estimate of this variance, using the delta method and the weighted least squares technique is suggested in Bieler and Williams (1993).

[15] Portier et al. (1986) recommends $k = 3$, although other values have been investigated (Gebregziabher and Hoel 2009; Moon et al. 2003). However, as noted in Gebregziabher and Hoel (2009), it appears that the tests are largely insensitive to the choice of *k*.

It may be noted that unlike the methods suggested in Peto et al. (1980), the poly-k analysis is independent of tumor context of observation information (i.e. if the tumor was observed on incidental or fatal context), which is a major advantage of this method over the Peto method.

12.4.2 The Exact Poly-k Method

The outcome of the experiment, for a specific tumor endpoint, can typically be summarized by a results table such as Table 12.11. Replacing the number of animals in each cell by the corresponding adjusted group sizes, we get a new results table, Table 12.12.

A simple-minded exact poly-k test can now be conducted performing the Cochran-Armitage test using the data in Table 12.12. However, in order to use the exact Cochran-Armitage test, the row and column totals for all permuted configurations of the observed table must be fixed. Since calculation of the n_i^* terms (the column totals) depends on the survival pattern of the animals, these terms cannot be assumed to be fixed, and the naïve use of the Cochran-Armitage test is not correct. For an appropriate exact test the adjusted column totals must be recalculated for every permutation of all animals.

We illustrate this method by considering a simple example:

12.4.2.1 Illustrative Example

Consider an experiment with two dose groups and five animals per group, continued up to 104 weeks. Suppose that the observed data for a specific tumor type are as

Table 12.11 Results table for single endpoint without survival adjustments

Group number	0	1	⋯	i	⋯	r
Dose level	$d_0 = 0$	d_1	⋯	d_i	⋯	d_r
Original group size	n_0	n_1	⋯	n_i	⋯	n_r
TBAs	x_0	x_1	⋯	x_i	⋯	x_r
Non tumor bearing animals (NTBAs)	$(n_0 - x_0)$	$(n_1 - x_1)$	⋯	$(n_i - x_i)$	⋯	$(n_r - x_r)$

Table 12.12 Results table for single endpoint with survival adjustments

Group number	0	1	⋯	i	⋯	r
Dose level	$d_0 = 0$	d_1	⋯	d_i	⋯	d_r
Adjusted group size	n_0^*	n_1^*	⋯	n_i^*	⋯	n_r^*
TBAs	x_0	x_1	⋯	x_i	⋯	x_r
Adjusted NTBAs	$(n_0^* - x_0)$	$(n_1^* - x_1)$	⋯	$(n_i^* - x_i)$	⋯	$(n_r^* - x_r)$

Table 12.13 Raw output for example

Animal number	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Dose level	0	0	0	0	0	1	1	1	1	1
Time of death (weeks)	80	104	104	96	104	74	98	104	50	104
Tumor code	0	0	0	0	0	1	0	0	0	1
Poly-3 weight	0.46	1.00	1.00	0.79	1.00	1.00	0.84	1.00	0.11	1.00

Table 12.14 Example of results table without survival adjustments

Group number	0	1	
Dose level	0	1	Total
Original group size	5	5	10
TBAs	0	2	2
Non tumor bearing animals	5	3	8

Table 12.15 Example of results table with survival adjustments

Group number	0	1	
Dose level	0	1	Total
Adjusted group size (rounded)	4	4	8
TBAs	0	2	2
Adjusted non tumor bearing animals	4	2	6

shown in Table 12.13. Calculating the adjusted group sizes, and rounding so we may use discrete tests, we have

$$n_0^* = \text{Round}\,(0.46 + 1.00 + 1.00 + 0.79 + 1.00) = \text{Round}(4.25) = 4$$
$$n_1^* = \text{Round}\,(1.00 + 0.84 + 1.00 + 0.11 + 1.00) = \text{Round}(3.95) = 4.$$

Tables 12.14 and 12.15 summarize this data for analysis in the styles of Tables 12.11 and 12.12. The table p values for the asymptotic Cochran-Armitage tests are $p = 0.06681$ for the unadjusted data and $p = 0.06332$ for the adjusted data.

12.4.2.2 Exact Method

An exact methodology for the poly-k test is possible. In the following we will describe this method.

The approach is combinatorial; we consider permutations of the animals among the dose group. Each permutation generates a summary table of the form of Table 12.15 (although many permutations can generate similar tables). For each table, we calculate the test statistic T defined by:

$$T = \sum_{i=0}^{r} x_i d_i \tag{12.4}$$

Table 12.16 Output after a typical permutation

Animal number	A1	A3	A7	A9	A10	A2	A4	A5	A6	A8
Dose level	0	0	0	0	0	1	1	1	1	1
Time of death (weeks)	80	104	98	50	104	104	96	104	74	104
Tumor code	0	0	0	0	1	0	0	0	1	0
Poly-3 weight	0.46	1.00	0.84	0.11	1.00	1.00	0.79	1.00	0.36	1.00

Table 12.17 Summary table for permuted data

Group number	0	1	
Dose level	0	1	Total
Adjusted group size (rounded)	3	4	7
TBAs	1	1	2
Adjusted non tumor bearing animals	2	3	5

The table p-value of a given outcome with test statistic $T = t$ can be calculated using the hypergeometric distribution.

We define an equivalence relation on permutations, saying that two permutations p_1 and p_2 are equivalent ($p_1 \sim p_2$) if they both allocate the same number of TBAs to each dose group. It is clear that the test statistic T respects this equivalence relation, but it is possible to have $T_{p_1} = T_{p_2}$ without $p_1 \sim p_2$ (Table 12.16).

For this permutation, we have

$$n_0^* = \text{Round}\,(0.46 + 1.00 + 0.84 + 0.11 + 1.00) = \text{Round}(3.40) = 3$$
$$n_1^* = \text{Round}\,(1.00 + 0.7865 + 1.00 + 0.3602 + 1.00) = \text{Round}(4.14) = 4.$$

The resulting summary table is displayed in Table 12.17:

We now calculate the probability that a randomly selected permutation induces a particular table of permuted values (such as Table 12.17). The random variables x_0 and x_1 can be assumed to be drawn from a random hypergeometric distribution $g(n_0^*, n_1^*)$. The probability of a given hypergeometric distribution's realization depends on the random parameters n_0^* and n_1^*, and so can be denoted $\Pr\left[g(n_0^*, n_1^*)\right]$. The probability of observing a particular table is therefore:

$$\Pr\left[x_0, x_1, n_0^*, n_1^*, g(n_0^*, n_1^*)\right] = \Pr\left[g(n_0^*, n_1^*)\right] \cdot \Pr\left[x_0, x_1, n_0^*, n_1^* \middle| g(n_0^*, n_1^*)\right] \quad (12.5)$$

where

$$\Pr\left[x_0, x_1, n_0^*, n_1^* \middle| g(n_0^*, n_1^*)\right] = \frac{\binom{n_0^*}{x_0}\binom{n_1^*}{x_1}}{\binom{n^*}{x}} \Big/ C\,(x_0, x_1) \quad (12.6)$$

with $C(x_0, x_1)$ as the total number of ways in which x TBAs and $n - x$ NTBAs can be arranged into groups of size n_0 and n_1 such that exactly x_0 of the TBAs are in Group 0. This quantity is given by:

$$C(x_0, x_1) = \binom{x}{x_0}\binom{n-x}{n_0-x_0}\binom{x-x_0}{x_1}\binom{(n-x)-(n_0-x_0)}{n_1-x_1}. \tag{12.7}$$

Note that $C(x_0, x_1)$ is equal to the total number of ways in which n number of animals can be arranged so that x_0 of the tumor bearing animals are in Group 0, and the remaining x_1 are in Group 1. For $r + 1$ treatment groups the general formula for $C(x_0, x_1, \ldots, x_r)$ is

$$C(\mathbf{x}) = \prod_{i=0}^{r-1} \binom{x - \sum_{k<i} x_k}{x_i}\binom{(n-x) - \sum_{k<i}(n_k - x_k)}{n_i - x_i}. \tag{12.8}$$

with $x = x_0 + x_1 + \ldots + x_r$, $x_{-1} = -x_0$, and $n_{-1} = -n_0$.

For example, for Table 12.14, with $x_{-1} = 0$, $x_0 = 0$, $x_1 = 2$, $n_{-1} = -5$, $n_0 = 5$, and $n_1 = 5$, we have $x = 0 + 2 = 2$, and $n = 5 + 5 = 10$, and

$$C(0, 2) = \binom{2}{0}\binom{10-2}{5-0}\binom{(2-0)}{2}\binom{(10-2)-(5-0)}{5-2} = 56.$$

The complete list of these 56 arrangements of 10 animals is given in Table 12.18. For each permutation p in this list, the test statistic T_p is equal to $0 \times 0 + 2 \times 12$. Furthermore, since there are just two groups, this is an exhaustive list of all permutations satisfying $T_p = 2$. The probability that $T = 2$ is therefore calculated by adding the probabilities for each of these 56 permutations, using Eq. (12.5).

As a simple choice, the $\Pr[g(n_0^*, n_1^*)]$ can be taken to be equal for each $g(n_0^*, n_1^*)$, and can be determined from the identity

$$\begin{aligned} &\sum \Pr\left[x_0, x_1, n_0^*, n_1^*, g(n_0^*, n_1^*)\right] \\ &\quad = \sum \Pr\left[g(n_0^*, n_1^*)\right] \Pr\left[x_0, x_1, n_0^*, n_1^* | g(n_0^*, n_1^*)\right] = 1 \end{aligned} \tag{12.9}$$

so that

$$\Pr\left[g(n_0^*, n_1^*)\right] = \frac{1}{\sum \Pr\left[x_0, x_1, n_0^*, n_1^* | g(n_0^*, n_1^*)\right]}. \tag{12.10}$$

In this respect, the factor $\Pr\left[g(n_0^*, n_1^*)\right]$ can also be considered as a normalizing factor. Therefore, operationally we first calculate $\sum \Pr\left[x_0, x_1, n_0^*, n_1^* | g(n_0^*, n_1^*)\right]$ and then normalize using $\Pr\left[g(n_0^*, n_1^*)\right]$.

It may be noted here that for the use of the hyper geometric distribution, although we round the n_j^*s to their nearest values, it is possible to use the ceiling or floor functions instead. This is a matter of individual discretion. One's choice will have

Table 12.18 All possible arrangements of 10 animals at risk with 0 TBAs in Group 0 and 2 TBAs in Group 1

Permutation	Animal numbers										n_0^*	n_1^*	n_0^*	n_1^*
number	Group 0					Group 1					Exact		Rounded	
1	1	2	3	4	5	6	7	8	9	10	4.24	3.95	4	4
2	1	2	3	4	9	5	6	7	8	10	3.35	4.84	3	5
3	1	2	3	4	8	5	6	7	9	10	4.24	3.95	4	4
4	1	2	3	8	9	4	5	6	7	10	3.57	4.62	4	5
5	1	2	3	4	7	5	6	8	9	10	4.08	4.11	4	4
6	1	2	3	7	8	4	5	6	9	10	4.29	3.9	4	4
7	1	2	3	7	9	4	5	6	8	10	3.4	4.79	3	5
8	1	2	7	8	9	3	4	5	6	10	3.4	4.79	3	5
9	1	2	3	5	7	4	6	8	9	10	4.29	3.9	4	4
10	1	2	3	5	8	4	6	7	9	10	4.46	3.73	4	4
11	1	2	3	5	9	4	6	7	8	10	3.57	4.62	4	5
12	1	2	5	8	9	3	4	6	7	10	3.57	4.62	4	5
13	1	2	5	7	8	3	4	6	9	10	4.29	3.9	4	4
14	1	2	5	7	9	3	4	6	8	10	3.4	4.79	3	5
15	1	5	7	8	9	2	3	4	6	10	3.4	4.79	3	5
16	1	2	4	5	9	3	6	7	8	10	3.35	4.84	3	5
17	1	2	4	5	8	3	6	7	9	10	4.24	3.95	4	4
18	1	2	4	5	7	3	6	8	9	10	4.08	4.11	4	4
19	1	2	4	7	9	3	5	6	8	10	3.19	5	3	5
20	1	2	4	7	8	3	5	6	9	10	4.08	4.11	4	4
21	1	2	4	8	9	3	5	6	7	10	3.35	4.84	3	5
22	1	4	7	8	9	2	3	5	6	10	3.19	5	3	5
23	1	4	5	8	9	2	3	6	7	10	3.35	4.84	3	5
24	1	4	5	7	8	2	3	6	9	10	4.08	4.11	4	4
25	1	4	5	7	9	2	3	6	8	10	3.19	5	3	5
26	4	5	7	8	9	1	2	3	6	10	3.73	4.46	4	4
27	1	3	4	8	9	2	5	6	7	10	3.35	4.84	3	5
28	1	3	4	7	8	2	5	6	9	10	4.08	4.11	4	4
29	1	3	4	7	9	2	5	6	8	10	3.19	5	3	5
30	1	3	4	5	7	2	6	8	9	10	4.08	4.11	4	4
31	1	3	4	5	8	2	6	7	9	10	4.24	3.95	4	4
32	1	3	4	5	9	2	6	7	8	10	3.35	4.84	3	5
33	1	3	5	7	9	2	4	6	8	10	3.4	4.79	3	5
34	1	3	5	7	8	2	4	6	9	10	4.29	3.9	4	4
35	1	3	5	8	9	2	4	6	7	10	3.57	4.62	4	5
36	1	3	7	8	9	2	4	5	6	10	3.4	4.79	3	5
37	3	5	7	8	9	1	2	4	6	10	3.95	4.24	4	4
38	3	4	7	8	9	1	2	5	6	10	3.73	4.46	4	4
39	3	4	5	8	9	1	2	6	7	10	3.9	4.29	4	4
40	3	4	5	7	8	1	2	6	9	10	4.62	3.57	5	4

(continued)

Table 12.18 (continued)

Permutation	Animal numbers										n_0^*	n_1^*	n_0^*	n_1^*
number	Group 0					Group 1					Exact		Rounded	
41	3	4	5	7	9	1	2	6	8	10	3.73	4.46	4	4
42	2	3	7	8	9	1	4	5	6	10	3.95	4.24	4	4
43	2	3	5	8	9	1	4	6	7	10	4.11	4.08	4	4
44	2	3	5	7	8	1	4	6	9	10	4.84	3.35	5	3
45	2	3	5	7	9	1	4	6	8	10	3.95	4.24	4	4
46	2	3	4	5	9	1	6	7	8	10	3.9	4.29	4	4
47	2	3	4	5	8	1	6	7	9	10	4.79	3.4	5	3
48	2	3	4	5	7	1	6	8	9	10	4.62	3.57	5	4
49	2	3	4	7	9	1	5	6	8	10	3.73	4.46	4	4
50	2	3	4	7	8	1	5	6	9	10	4.62	3.57	5	4
51	2	3	4	8	9	1	5	6	7	10	3.9	4.29	4	4
52	2	4	5	7	9	1	3	6	8	10	3.73	4.46	4	4
53	2	4	5	7	8	1	3	6	9	10	4.62	3.57	5	4
54	2	4	5	8	9	1	3	6	7	10	3.9	4.29	4	4
55	2	4	7	8	9	1	3	5	6	10	3.73	4.46	4	4
56	2	5	7	8	9	1	3	4	6	10	3.95	4.24	4	4

some effects on the calculation of the p-value if the sample size is very small (as would have been the case with our example, with five animals per group). However for moderately large sample size this choice will have minimal effect. It should be noted that, as discussed in Sect. 12.3.3, the regular carcinogenicity studies have 50–70 animals per group.

12.4.3 Second Example

We further illustrate the use and properties of our method with another example:

12.4.3.1 Framework of Second Example

12.4.3.1.1 Design

Two dose groups with dose levels 0 and 1. Twelve animals, with animal numbers $1, \ldots, 12$ are randomly divided into two treatment groups, G0 and G1, of six animals each. Terminal sacrifice is at Week 104. The example data are presented in Table 12.19. The observed value of the test statistic T is $1 \times 0 + 4 \times 1 = 4$.

As described in Sect. 12.4.2.1, we can calculate the distribution of T; the distribution is shown in Table 12.20 Since the observed value of T is 4, the p-value

Table 12.19 Observed data from second example

Animal number	1	2	3	4	5	6	7	8	9	10	11	12
Dose group	G1	G0	G1	G0	G1	G1	G1	G1	G0	G0	G0	G0
Tumor code	0	0	1	0	1	0	1	1	0	0	1	0
Survival time (Week)	78	104	55	104	104	104	65	85	104	50	45	104

Table 12.20 Distribution of test statistic for second example

	PDF	
T	Poly-3 adjusted test	Unadjusted Cochran-Armitage test
0	0.00973	0.00758
1	0.11932	0.11364
2	0.37095	0.37879
3	0.37095	0.37879
4	0.11932	0.11364
5	0.00973	0.00758

of the test is $\Pr[T \geq 4]$. Using the exact poly-3 test, we get a p-value of 0.12905. For the unadjusted Cochran-Armitage test, the p-value is $0.11364+0.00758 = 0.12122$. For the asymptotic one-tailed test (using StatXact), it is 0.05124.

It should be noted that this method is based on an extensive computational procedure requiring the evaluation of all possible permutations of the animals to $r + 1$ dose groups. This computational complexity is a big challenge for the application of the proposed method in the data analysis of real studies. However, some modifications of commercially available software for the calculations of the probabilities of the hypergeometric distribution may facilitate these calculations.

12.5 Modified Exact Poly-3 Method

Since the exact poly-3 method described in Sects. 12.4.2.1 and 12.4.3 has severe computational limitations when we have group sizes of 50 or larger, the alternative and more practical way is to use the permutation sample to estimate p-values. A survival-adjusted exact randomization trend test procedure (Mancuso et al. 2002) has been proposed to use the permutation sample to estimate the p-values. The test is carried out by using PROC STRATIFY with fixed row and column sums assumptions. In order to reduce biases caused by the assumptions of fixed column and row sums using PROC STRATIFY and binomial null variance estimate from Mancuso et al. (2002), we are proposing a modified exact poly-3 trend test that can be regarded as an exact version of the poly-3 test (Bieler and Williams 1993).

12.5.1 Motivating Problem

As discussed earlier (see Sect. 12.4.2, and especially Table 12.13), animal survival time is not a fixed quantity. The adjusted quantal response estimates, $p_i^* = x_i/n_i^*$, are actually ratios of linear statistics. Hence, the numerators and denominators of these estimates are both subject to random variation.

12.5.2 Permutational Distribution for the Modified Poly-3 Test

12.5.2.1 The Modified Poly-3 Trend Test

The quantal response tests that focus on crude lifetime tumor incidence rates and make no adjustment for differences in survival experiences across dose groups are often biased, since they implicitly assume that all animals are at equal risk of developing a tumor over the course of study. As mentioned in Sect. 12.4.1, in order to address this issue, Bailer and Portier (1988) introduced a modification to the Cochran-Armitage test for trend that adjusts for differences in treatment lethality while requiring no assumptions regarding tumor lethality or changes to study design. This poly-3 trend test incorporates a weighting scheme that allows fractional information into the analysis for animals not at full risk for tumor development. This weighting scheme essentially modifies the denominators of the crude quantal response estimates of lifetime tumor incidence to more closely approximate the total number of animal years at risk in each experimental group.

Using this weighting scheme and the notation used previously, we define $p_i^* = x_i/n_i^*$ as the *adjusted quantal response estimate of lifetime tumor incidence in group i*; and

$$p^* = \frac{\sum_i x_i}{\sum_i n_i^*} \tag{12.11}$$

as the *experiment-wide adjusted quantal response estimate of lifetime tumor incidence*; and $q_i^* = 1 - p_i^*$ and $q^* = 1 - p^*$. As previously stated, binomial null variance estimates of p_i^* do not apply since they are based on the assumption that the number of animals at risk is fixed. However, by using a Taylor expansion, a pooled null variance estimate for p_i^* can be found:

$$\text{var}_0\left(p_i^*\right) \approx \left(\frac{n_i}{n_i^*}\right)^2 \cdot \frac{\sum_i \sum_j (r_{ij} - \bar{r}_i)^2}{n - (g+1)}. \tag{12.12}$$

where $n = \sum_i n_i$, $r_{ij} = x_i - p_i^* w_{ij}$, and $g + 1$ is the number of experimental groups in the study.

A computational formula for the modified Cochran-Armitage test statistic, which will be referred to as the *ratio test* proposed in Bieler and Williams (1993) and denoted by Z_r is given as follows:

$$Z_r = \frac{\sum_i a_i p_i^* d_i - \left(\sum_i a_i d_i\right)\left(\sum_i a_i p_i^*\right) \Big/ \sum_i a_i}{\sqrt{C\left(\sum_i a_i d_i^2 - \left(\sum_i a_i d_i\right)^2 \Big/ \sum_i a_i\right)}} \tag{12.13}$$

where

$$C = \sum_i \sum_j \frac{\left(r_{ij} - \bar{r}_{ij}\right)^2}{n - (g+1)} \quad a_i = \frac{n_i}{\left(n_i^*\right)^2}.$$

12.5.3 Permutational Distribution for the Modified Poly-3 Test

Exact methods are preferable for sparse data. Permutation tests consider all possible assignments of animals to dose groups as equally likely, while fixing the rest of the information obtained in the experiment. Under the null hypothesis of no treatment effect, this results in an exact conditional distribution of the test statistic when intercurrent mortality patterns are equal across groups, and it is asymptotically correct when the mortality patterns are unequal (Fairweather et al. 1998; Heimann and Neuhaus 1998). Given a data set, consider all, say M, possible allocations of animals to groups while keeping the observed data for each animal fixed. Corresponding to these M arrangements, we may obtain M values of the test statistic. The permutational distribution of the modified poly-3 test statistic results from assigning equal probability to each of these M values. Letting Z_r^* be the observed value, the p-value is the proportion of the M values that are at least as extreme as Z_r^*. By exhaustive enumeration, the computation for the p-value using the permutational distribution for the test is straightforward and efficient if the number of animals in the study is small. For data involving large numbers of subjects, the p-value associated with the permutational distribution of the test statistic may be approximated by a sample of the set of all permutations.

12.5.4 Simulations and Results

We conducted a Monte Carlo simulation study to evaluate the following tests: the exact version of the modified poly-3 test (Bieler and Williams 1993), the exact version of the poly-3 test (Bailer and Portier 1988) and PROC STRATIFY in two simulation designs (Dinse 1985; Portier et al. 1986). For each configuration, 10,000 simulated data sets were generated and tested by various methods at the nominal

significance level $\alpha = 0.05$. Additionally, these methods were tested against the significance thresholds described in Table 12.1 for a standard study (where, as in Sects. 12.2 and 12.3, rarity was determined by the incidence rate in the control group).

In conducting the exact versions of the modified or the not-modified poly-3 test using the permutational distribution, p-values were estimated from samples of 5000 permutations.

12.5.4.1 Monte Carlo Simulation Design 1

A typical bioassay design with four groups of 50 animals each and an experimental duration of 104 weeks is used in the study. The design is simulated to have a single terminal sacrifice at the end of the experiment, as in the customary long-term rodent bioassay. The dose levels used are (0,1,2,3) across groups. The three independent variables T_0 (time to tumor onset), T_2 (time from tumor onset until death from the tumor), and T_1 (time until death from a competing risk) are used to model animal tumorigenicity data. These variables are generated from the modified Weibull distributions used by Portier et al. (1986) and others in the literature (Ahn and Kodell 1995; Chang et al. 2000; Kodell and Ahn 1997; Kodell et al. 1994). The survival function for T_0 is

$$S(t) = \exp\left[-\delta_1 (1/104)^{\delta_2}\right].$$

with $\delta_2 \in \{1.5, 3, 6\}$ and δ_1 chosen so that the probability of tumor onset by the end of the study attains the desired rate. Since the study is concerned with rare events, tumor rates between 0.01 and 0.15 are used.

The survival function for T_1 is

$$Q(t) = \exp\left[-\phi\left(\gamma_1 t + \gamma_2 t^{\gamma_3}\right)\right] \tag{12.14}$$

with $\gamma_1 = 10^{-4}$, $\gamma_2 = 10^{-16}$ and $\gamma_3 = 7.425531$, and the value ϕ chosen such that the competing risks survival rate with respect to all causes of death except for the tumor of interest at 104 weeks is either 0.5 for all groups or (0.5, 0.4, 0.3, 0.2) across groups. The control survival rate chosen represents the one recently observed in the NTP studies for male Fischer 344 rats (Haseman et al. 1998), although it is somewhat below average for $B6C3F_1$ mice and F344/N female rats in the NTP feeding studies.

For simplicity, the survival function for T_2 has the same form as $Q(t)$ with the same values of γ_1, γ_2 and γ_3.

12.5.4.2 Results of First Monte Carlo Simulation

The results of the first set of simulations are presented in Table 12.21 (common tumors) and Table 12.22 (rare tumors). In the following tables, Proposed refers to our proposed method and Mancuso refers to the method described in Mancuso et al. (2002). For trend test, the multiplicity adjustment decision rule (referred as Adjusted α in the following tables) is common and rare tumors are tested at 0:005 and 0:025 significance levels, respectively.

Table 12.21 Null hypothesis rejection rates for Monte Carlo simulation of modified poly-3 trend test and the poly-3 test described in Bailer and Portier (1988) and Mancuso et al. (2002)

	Tumor incidence				Decision rule			
					Adjusted α		$\alpha = 0.05$	
	Control	Low	Mid	High	Proposed method	Mancuso method	Proposed method	Mancuso method
Size	0.01	0.01	0.01	0.01	0.0314	0.0334	0.0631	0.0657
	0.02	0.02	0.02	0.02	0.0100	0.0100	0.0650	0.0637
	0.03	0.03	0.03	0.03	0.0083	0.0086	0.0644	0.0643
	0.04	0.04	0.04	0.04	0.0083	0.0080	0.0712	0.0704
	0.05	0.05	0.05	0.05	0.0080	0.0080	0.0741	0.0731
	0.075	0.075	0.075	0.075	0.0111	0.0106	0.0846	0.0838
	0.10	0.10	0.10	0.10	0.0123	0.0121	0.0875	0.0861
Power	0.01	0.02	0.06	0.12	0.86	0.94	0.89	0.96
	0.02	0.02	0.14	0.14	0.84	0.84	0.89	0.88
	0.01	0.03	0.14	0.12	0.69	0.70	0.78	0.78
	0.01	0.01	0.01	0.14	0.90	0.99	0.93	0.99
	0.01	0.12	0.10	0.14	0.44	0.45	0.59	0.59
	0.01	0.14	0.14	0.12	0.19	0.19	0.32	0.31

Table 12.22 Summary sizes of tests of rare tumors with spontaneous incidence rates 0–1 % in simulation design I

	Tumor incidence				Decision rule			
					Adjusted α		$\alpha = 0.05$	
	Control	Low	Mid	High	Proposed method	Mancuso method	Proposed method	Mancuso method
Size	0.001	0.001	0.001	0.001	0.0209	0.0335	0.0315	0.0442
	0.002	0.002	0.002	0.002	0.0307	0.0481	0.0473	0.0649
	0.003	0.003	0.003	0.003	0.0348	0.0541	0.0549	0.0716
	0.004	0.004	0.004	0.004	0.0360	0.0512	0.0569	0.0690
	0.005	0.005	0.005	0.005	0.0386	0.0500	0.0592	0.0690
	0.006	0.006	0.006	0.006	0.0372	0.0471	0.0598	0.0685
	0.0075	0.0075	0.0075	00.075	0.0356	0.0405	0.0601	0.0655
	0.009	0.009	0.009	0.009	0.0346	0.0376	0.0648	0.0676

These results in Table 12.21 indicate that there were inflations of Type 1 error under both adjusted α and 0.05 significance levels in all cases. Both the proposed Bieler and Williams (1993) and Mancuso's methods in Mancuso et al. (2002) have very similar power. The results in the Table 12.22 show that in the four treatment group experimental design with rare tumors with background incidence rates between 0 and 1 %, Type 1 errors were inflated in both the proposed and Mancuso's methods similar to the inflated Type 1 errors of the common tumors. But the proposed method has slightly more control over Type 1 error compared to Mancuso's method.

12.5.4.3 Second Set of Monte Carlo Simulations

We also conducted a second set of simulations to investigate the modified poly-3 trend test. For this set, thirty-six simulation models were obtained by varying the levels of the four factors described in Dinse (1985) and presented in Sect. 12.2.

In general we observed that:

1. By using multiplicity adjustment, sizes are all less than 0.041. An inflation of size still occurs in 8 out of 12 cases: seven in the range 0.0071–0.03 and one at 0.041 based on adjusted levels of significance.
2. When a small dose effect on tumor rate, a high background tumor rate and later tumor appearance were simulated, power appeared to be lower.
3. Some of the high background tumor rate cases have very low levels of power. In these cases, using just 5000 replicates to estimate the *p*-values was inadequate.

12.5.5 Test Results Using Some NDA Datasets

Three NDA submissions were randomly chosen from the set of submissions reviewed by the authors to compare the proposed method with PROC STRATIFY. The permutation sample sizes were all around 6000. There were about tenfold differences in *p*-values between the proposed and Mancuso's methods. We did further investigations on the significant cases to see how different permutation sample sizes would have an impact on *p*-value calculations for the trend test. Permutation samples of size of 1,000,000 were used. However, the *p*-values of samples of size of 1,000,000 were not much different from the p-values using the permutation samples size around 6000. The only changes were in the 3rd, 4th or 5th decimal places of the *p*-values. Since there are very large numbers of permutations (usually in the magnitude of 10^{100} permutations with for 50 animals per groups in a 4 group experimental design), computational limitations are a very real concern for all of these methods.

12.6 A Short Review of the General Bayesian Approaches to Possible Survival and Carcinogenicity Analyses

12.6.1 Points We Want to Make Before We Get Going

- Bayesian methods base conclusions solely on the posterior distribution of the parameters.
- It may be that an improper prior (i.e. a distribution that is not a proper density) is useful, provided it results in a valid posterior.
- Conclusions about a single parameter are based on the posterior distribution for that parameter found by integrating out all other parameters, including any nuisance parameters. This is a very general procedure, but may seem inflexible in comparison to the variety of frequentist methods.
- An approach to Bayesian hypothesis testing is consider the null hypothesis as a Bernoulli variable, and to use observed data to construct a posterior for the probability that the hypothesis is in fact true.

A central Bayesian result having far-reaching implications for both Bayesian and frequentist statistical analysis is the so-called *likelihood principle*:

> Unless results are based only on the data that were actually observed and not those that could have occurred, results can be improved.

12.6.2 Bayesianism and Nonclinical Biostatistics

Historically, except for those cases where researchers have been able to exploit conjugate priors (families of distributions for which if a prior is in the family, then any posterior will also be in the family) Bayesianism has been limited by the computational complexity of calculating posterior distributions, especially after repeated updating. However, as computational power has increased, more areas of statistics, including nonclinical biostatistics, have seen Bayesian methods become viable.

It is generally the case that as more data is collected, the influence of the initial prior diminishes, and the posterior becomes primarily a reflection of the observed empirical data. When working with large datasets, this is reassuring, and addresses the common criticism that Bayesian methods are founded on an unjustifiable choice of a prior. However, as has already been noted repeatedly in this chapter, one of the greatest challenges of reviewing rodent carcinogenicity data is the rarity of the events. So in contrast to the reassuring asymptotic case, Bayesian analyses of such data are more, rather than less, sensitive to the choice of a prior. For these smaller sample sizes, we need to use so-called *noninformative*, *vague*, or *objective* priors that do not dominate the data. The so-called *reference prior* method described in Bernardo (1979) and Bernardo and Smith (1994) is an automatic procedure for generating such priors.

In the analysis of most rodent carcinogenicity studies there are two primary goals:

1. To analyze the effect of the compound under study on survival.
2. To analyze the effect of the compound on the development of neoplasms.

In the typical frequentist testing of carcinogenicity hypotheses or survival the usual null hypothesis is that some set of parameters (typically slope parameters, such as D in the Weibull parameterization described in Sect. 12.2.2) are equal to zero. Testing this hypothesis in the manner described above (Sect. 12.6.1), we are interested in the posterior distribution of the random event that D is identically equal to 0.

The following proposed Bayesian analyses are intended to be illustrative only, and not prescriptive.

12.6.3 Notational Conventions for Examples

In the examples below, we adopt the following notational conventions:

There are I tumor types (as discussed in Sect. 12.3.6.2), J animals, and K dose groups. Animal j is a member of group $\kappa(j)$, and the total number of animals in group k is denoted n_k. Without loss of generality, let $k = 0$ denote the control group. The animals in group k are treated with dose d_k (so $d_0 = 0$). The maximum time in the study is denoted T, and the time at which animal j leaves the study (either through natural death or sacrifice) is denoted t_j.

12.6.4 Survival Analysis Example: Finite Dimensional Proportional Hazards Model

The probability of an animal surviving past time t is given by the survival function $S(t) = \Pr(T > t)$. Let $f(t)$ denote the density of T. The instantaneous hazard function is $h(t) = f(t)/S(t)$, and the cumulative hazard H is defined by:

$$H(t) = \int_0^t h(u)\mathrm{d}u.$$

The following identities follow immediately:

$$f(t) = h(t)S(t) \quad \ln(S(t)) = -H(t) \quad S(t) = \mathrm{e}^{-H(t)} \quad f(t) = \mathrm{e}^{-H(t)}.$$

The standard Cox regression form of the proportional hazards model for such survival models specifies the hazard function:

$$h(t|\mathbf{x}) = h_0(t)\mathrm{e}^{\mathbf{x}^\top \boldsymbol{\beta}}.$$

Then treatment effects can be investigated by assessing the differential effects of treatment in the $e^{\mathbf{x}^T\boldsymbol{\beta}}$ term. Among other possible specifications, this can reflect a trend over dose or individual dose effects.

Statistical inference on survival is based on proposing a probability model for $S(t)$ or one of its derivations. The probability model is defined so that hypotheses to be investigated are specified as parameters in the model. A frequentist analysis takes parameters as fixed and assesses the likelihood of the observed data. A Bayesian analysis starts by noting that parameters are not known, and assumes that a prior distribution is a natural measure of this lack of exact knowledge. Then the Bayesian analysis assesses the impact of the actual observed data on this prior.

Frequentist analysis of the Cox model uses asymptotics to analyze the linear predictor (and by extension the hazard ratio), but disregards the baseline hazard h_0.[16] By contrast, a Bayesian analysis requires priors on all parameters, including the baseline hazard. In this example, we consider a finite dimensional space of possible baseline hazard functions, namely the piecewise step functions; i.e., hazard functions of the form

$$h_0(t) = \sum_{m=0}^{M} \lambda_m \mathbf{I}_{(a_m, a_{m+1}]}(t). \tag{12.15}$$

Without loss of generality, we may assume $0 = a_0 < a_1 \ldots < a_M < a_{M+1} = T$.

In the formulation above, the baseline hazard is confounded with the specification of treatment effects, i.e., a multiplicative constant can be moved to either the baseline hazard or the term with covariates. The dose effect at level k is represented by the scalar β_k, interpreted as the log of the hazard ratio relative to the control group. Note that $\beta_0 = 0$. We thus have $K - 1$ unknown scalars, together with the unknown baseline hazard function $h_0(t)$. The model could be simplified further by assuming

$$\beta_k = \eta d_k + \mu \tag{12.16}$$

for $k > 0$ i.e., by assuming a simple linear trend in dose.

Given that $a_m < t \leq a_{m+1}$, the integrated cumulative hazard for an animal in group k may be written as:

$$H_0(t) = e^{\beta_k} \int_0^t h_0(u)\,\mathrm{d}u = e^{\beta_k} \left(\left(\sum_{n=0}^{m-1} \lambda_n (a_{n+1} - a_n) \right) + \lambda_m (t - a_m) \right)$$

and the likelihood for subject j can be written

$$L_j(\boldsymbol{\beta}) \propto \begin{cases} e^{-H_0(t_j)} & \text{if the } j^{\text{th}} \text{ subject is censored at time } t_j \\ \lambda_{\kappa(j)} e^{\beta_{\kappa(j)} - H_0(t_j)} & \text{if the } j^{\text{th}} \text{ subject fails at time } t_j. \end{cases} \tag{12.17}$$

[16] For this reason it is often called a *semiparametric* model.

Note that in the case that we use the model described in Eq. (12.16), the parameters in the likelihood function shown in Eq. (12.17) will be η and μ, rather than the parameter vector $\boldsymbol{\beta}$.

Because this looks like a sample of exponential inter-arrival times, we would expect the simple fail/not fail distributions to correspond to Poisson random variables. For subject j censored or failed at time t_j define γ_{jm} by

$$\gamma_{jm} = \begin{cases} \lambda_m \left(a_{m+1} - a_m\right) & \text{for } t_j > a_{m+1} \\ \lambda_m \left(t_j - a_m\right) & \text{for } a_m < t_j \leq a_{m+1} \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$S(t) = \mathrm{e}^{-H(t)} = \prod_{\{m|a_m \leq t_j\}}^{M} \exp\left(-\mathrm{e}^{\beta_{\kappa(j)}} \gamma_{jm}\right).$$

Furthermore, $(t_j - a_m)$ is constant with respect to the parameters, and hence can be incorporated in the likelihood for subjects who fail by multiplying λ_j by this difference. Thus for subject j, the likelihood can also be written as:

$$L_j(\boldsymbol{\beta}) \propto \begin{cases} \prod_{m=1}^{M} \exp\left(-\mathrm{e}^{\beta_{\kappa(j)}} \gamma_{jm}\right) & \text{if the } j^{\text{th}} \text{ subject is censored at time } t_j \\ \gamma_{jm} \mathrm{e}^{\beta_{\kappa(j)}} \prod_{m=1}^{M} \exp\left(-\mathrm{e}^{\beta_{\kappa(j)}} \gamma_{jm}\right) & \text{if the } j^{\text{th}} \text{ subject fails at time } t_j. \end{cases}$$

Although it looks messy, this is the likelihood of T independent Poisson random variables with mean $\mathrm{e}^{\beta_{\kappa(j)}} \gamma_{jm}$ where all responses are zero. This is only a computational convenience but allows easy estimation of the appropriate parameters using standard software (e.g., Lunn et al. 2000—see Sect. 12.6.8). Thus we need to specify an appropriate prior for the baseline hazard. Note that the baseline hazard is essentially the hazard of the control group. A gamma prior would be skewed to the right and would seem to be an appropriate choice. The standard 2 year study could be broken down into twelve 2 month periods. Sacrifice or accidental death could be treated as a reduction in the risk set, but not as a mortality event. In most circumstances we would might prefer a specification of an increasing hazard (again easily specified in WinBUGS or OpenBUGS (Lunn et al. 2000).

12.6.5 Carcinogenicity Example: Finite Dimensional Logistic Model

A logistic model is easy to implement in OpenBUGS or WinBUGS (Lunn et al. 2000). For this analysis we define mixed two-stage/three-stage hierarchical models for tests of trend and pairwise comparisons.

For testing trend, we define θ_{ij} to be the probability that tumor i is found in subject j, and we build the following model:

$$\text{logit}(\theta_{ij}) = \alpha_i + \beta_i d_{k(j)} + \gamma_i \ln(t_j) + \delta_j \tag{12.18}$$

where δ_j is the individual random subject effect.[17]

We assign model priors:

$$\begin{aligned}
\alpha_i &\sim \text{N}(\mu_\alpha, \sigma_\alpha^2) \\
\beta_i &\sim \pi_i \mathbf{I}_{[0]} + (1 - \pi_i)\text{N}(\mu_\beta, \sigma_\beta^2) \\
\pi_i &\sim \text{Beta}(\phi, \psi) \\
\gamma_i &\sim \text{N}(\mu_\gamma, \sigma_\gamma^2)
\end{aligned}$$

for $i = 1, \ldots, I$ and a random subject effect

$$\delta_j \sim \text{N}(\mu_\delta, \sigma_\delta^2)$$

for $j = 1, \ldots, J$. For computational convenience, we typically define $\mu_\alpha = \mu_\beta = \mu_\gamma = \mu_\delta = 0$ and $\sigma_\delta^2 = \sigma_\alpha^2 + \sigma_\beta^2 = \sigma_g^2 = \sigma_s^2 = 100$, $\sigma_\alpha^2, \sigma_\beta^2, \sigma_g^2 \sim$ InverseGamma$(1, 3)$.

The model for the pairwise comparison between group k and the control group is similar:

$$\text{logit}(\theta_{ij}) = \alpha_i + \beta_{ik} \mathbf{I}_{\{\kappa(j)=k\}} + \gamma_i \ln(t_j) + \delta_j. \tag{12.19}$$

Our priors have the form:

$$\begin{aligned}
\pi_{ik} &\sim \text{Beta}(\phi, \psi) \\
\beta_{ik} &\sim \pi_{ik} \mathbf{I}_{[0]} + (1 - \pi_{ik})\text{N}(\mu_{\beta_k}, \sigma_{\beta_k}^2)
\end{aligned}$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n_t$. Note that with this parameterization, for $k = 2, \ldots, K$, the β_{ik} terms represent the deviation of treatment effect from the controls. These should represent reasonably well dispersed priors on parameters.

[17] It is interesting to note that this model implies that the odds of tumorigenesis are proportional to $t_j^{\gamma_i}$, which (when the probability of tumorigenesis is low) is essentially equivalent to the poly-k assumption discussed elsewhere in this chapter.

12.6.6 Survival Analysis Example: Nonparametric Bayesian Analysis

Some applications involve increasing numbers of parameters or even infinite dimensional problems. Perhaps the knowledge about the parameter could follow a probability distribution *not* indexed by small set of parameters. For example, instead of something like a simple normal distribution indexed with a mean, μ and variance, σ^2, the family could be say one of the continuous location family distributions or possibly even the inclusive continuous probability distributions. In a simple misnomer such problems have come to be called "Bayesian Nonparametrics." The challenge is not the fact that there are no parameters, but rather that there are far too many. Since it seems to be quite difficult to specify priors with content in infinite dimensional space it seems more appropriate to work with objective priors that cover much of the parameter space.

One of many possible standard models for the survival function is to model the logarithm of the survival with a normal distribution, i.e. to specify that T_i follows a lognormal distribution. However, the typical Bayesian nonparametric model takes such a specification and uses it as a baseline function to be perturbed to "robustify" the model using a so-called Dependent Dirichlet Process (DDP) as the prior on this space of probability distributions. This function represents the prior using a so-called Dependent Dirichlet Process (DDP) as the prior on this space of probability distributions, which uses a mixture of normal distributions weighted by a Dirichlet process on the normal parameters. The prior is defined as a Dirichlet process where the baseline distribution models the linear parameters, where has the linear mean parameters has a normal distribution as prior and the variance parameters with a Gamma distribution. The prior of the precision parameter of the Dirichlet process is specified as a gamma distribution. The priors for the other hyperparameters in this function are conjugate distributions. Following the notation of Jara et al. (2014), we can write:

$$\begin{aligned} \ln(T_i) &= t_i|\mathbf{f}_{X_i} \sim \mathbf{f}_{X_i} \\ \mathbf{f}_{X_i} &= \int \mathrm{N}(X_i\boldsymbol{\beta}, \sigma^2) G(\mathrm{d}\boldsymbol{\beta}\mathrm{d}\sigma^2) \\ G|\alpha, G_0 &\sim DP(\alpha G_0) \end{aligned}$$

Typically distributions of the hyperparameters above can be specified as follows:

$$\begin{aligned} G_0 &= \mathrm{N}(\beta|\mu_b, s_b)\Gamma\left(\sigma^2|\frac{\tau_1}{2}, \frac{\tau_2}{2}\right) \\ \alpha|a_0, b_0 &\sim \mathrm{Gamma}(a_0, b_0) \\ \mu_b|m_0, s_0 &\sim \mathrm{N}(m_0, S_0) \\ s_b|\nu, \Psi &\sim \mathrm{InvWishart}(\nu, \Psi) \\ \tau_2|\tau_{s_1}, \tau_{s_2} &\sim \mathrm{Gamma}(\tau_{s_1}, \tau_{s_2}) \end{aligned}$$

See, for instance De Iorio et al. (2009). The parameterization used to compare doses can be captured by a dummy coding, as in the finite dimensional example (see Sect. 12.6.5).

12.6.7 *Carcinogenicity Example: Nonparametric Logistic Model*

A similar model to the one in Example 12.6.5 takes the baseline distribution as a logistic distribution. The nonparametric Bayesian approach treats an actual probability distribution as one of the parameters. This distribution is then sampled from an infinite dimensional space of possible distributions, which is both mathematically challenging and where, unlike most finite dimensional parameters, it is difficult to specify appropriate prior distributions. Thus one attempts to specify robust priors on the slope and treatment differences that have a small impact on the result. The baseline model follows a simple logit model for tests of trend and pairwise comparisons. For testing trend, we define p_{ijk} as the probability of tumor type i being found in subject j in treatment group k. That is, with $i = 1$ to n_t tumors and $j = 1$ to n_s animals, and dose d_k, leaving the experiment at time $t - j$ and subject effect δ_j:

$$\text{logit}(p_{ijjk}) = \alpha_i + \beta_i d_k + \gamma_i t_j + \delta_j \tag{12.20}$$

with assigned model priors:

$$\begin{aligned} \alpha_i &\sim \text{N}(\mu_{\alpha_i}, \sigma_\alpha^2) \\ \beta &\sim \text{N}(\mu_B, \sigma_\beta^2) \\ \gamma_i &\sim \text{N}(\mu_\beta, \sigma_g^2) \end{aligned}$$

But now, instead of directly specifying that the animal random effect δ_j, we specify the distribution as a Dirichlet process (DP) on the space of distributions.

$$\begin{aligned} \delta_i | G &\sim G \\ G | \alpha, G_0 &\sim \text{DP}(\alpha G_0) \\ G_0 &\sim \text{N}(\mu, \Sigma) \end{aligned}$$

Note that care seems to be needed to ensure that parameters are identified. Again, this is a simple application of the function in the `DPpackage` (Jara et al. 2014) in R (R Core Team 2012).

12.6.8 Software

Prior to the development of various Markov Chain Monte Carlo methods, actual software for doing a Bayesian analysis was largely limited to approximate solutions or even insisting on so-called conjugate priors. While many statisticians could see the philosophical advantages of Bayesian methodology, the lack of good methods of computing the posterior limited the application of these methods. All that has changed with the development of so called Markov Chain Monte Carlo methods, in their most simple form similar to so-called importance sampling.

For most problems in "classical" Bayesian analyses the user is faced with a plethora of choices in packages or programs. WinBUGS and its more recent descendant OpenBUGS (Lunn et al. 2000) are probably the oldest and most used general programs ("BUGS" stands for *Bayesian analysis Using Gibbs Sampling*). The various versions of the manuals have the warning in quite noticeable type "WARNING: MCMC can be dangerous." The point is that MCMC methods work well when they move around through the appropriate parameter space reaching near all feasible points. When they are stuck in a region of the parameter space and can not leave the MCMC methods can fail. WinBUGS includes several diagnostics for this type of behavior.

Several SAS procedures have options for a Bayesian analysis, usually with default, but quite reasonable priors. For more general use, PROC MCMC, provides a detailed analysis including extensive diagnostics on the MCMC Markov Chains, but, as with all very general procedures, requires careful coding of priors and likelihoods. It includes extensive, possibly nearly exhaustive, diagnostics for the MCMC behavior.

R users (R Core Team 2012) have a number of packages for Bayesian Analysis available to them. `LaplacesDemon` is a very general package. `MCMCpack` includes a relatively long list of functions for MCMC analysis. `BayesSurv` has a number of R functions for survival models. Last but certainly not least, in this short and by no means exhaustive list, `DPpackage` (Jara et al. 2014) is a very general collection of functions for Nonparametric Bayesian Analysis and is undoubtedly currently the easiest way to implement such models.

Acknowledgements The authors would like to acknowledge the support of Yi Tsong, the director of Division of Biometrics 6, in FDA/CDER/OTS/OB, while researching the work contained in this chapter.

References

Ahn H, Kodell R (1995) Estimation and testing of tumor incidence rates in experiments lacking cause-of-death data. Biom J 37:745–765

Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11(3): 375–386

Bailer AJ, Portier CJ (1988) Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics 44(2):417–431
Baldrick P, Reeve L (2007) Carcinogenicity evaluation: comparison of tumor data from dual control groups in the cd-1 mouse. Toxicol Pathol 35(4):562–575
Bergman CL, Adler RR, Morton DG, Regan KS, Yano BL (2003) Recommended tissue list for histopathologic examination in repeat-dose toxicity and carcinogenicity studies: a proposal of the society of toxicologic pathology (stp). Toxicol Pathol 31(2):252–253
Bernardo JM (1979) Reference posterior distributions for Bayesian inference. J R Stat Soc Ser B Methodol 41(2):113–147
Bernardo JM, Smith AFM (1994) Bayesian statistics. Wiley, Chichester
Bieler GS, Williams RL (1993) Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. Biometrics 49(3):793–801
Center for Drug Evaluation and Research (2005) Reviewer guidance: conducting a clinical safety review of a new product application and preparing a report on the review. United States Food and Drug Administration
Center for Drug Evaluation and Research (2103) Pharmacology review—NDA 205437 (otzela). Technical report, US Food and Drug Administration. http://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/205437Orig1s000PharmR.pdf
Chang J, Ahn H, Chen J (2000) On sequential closed testing procedures for a comparison of dose groups with a control. Commun Stat Theory Methods 29:941–956
Chernick MR, Liu CY (2002) The saw-toothed behavior of power versus sample size and software solutions. Am Stat 56(2):149–155
Clinical Data Interchange Standards Consortium (CDISC) (2011) Standard for exchange of nonclinical data implementation guide: nonclinical studies version 3.0
De Iorio M, Johnson WO, Müller P, Rosner GL (2009) Bayesian nonparametric nonproportional hazards survival modeling. Biometrics 65(3):762–771. doi:10.1111/j.1541-0420.2008.01166.x. http://dx.doi.org/10.1111/j.1541-0420.2008.01166.x
Dinse GE (1985) Testing for a trend in tumor prevalence rates: I. nonlethal tumors. Biometrics 41(3):751
Fairweather WR, Bhattacharyya A, Ceuppens PR, Heimann G, Hothorn LA, Kodell RL, Lin KK, Mager H, Middleton BJ, Slob W, Soper KA, Stallard N, Venture J, Wright J (1998) Biostatistical methodology in carcinogenicity studies. Drug Inf J 32:401–421
Gebregziabher M, Hoel D (2009) Applications of the poly-*k* statistical test to life-time cancer bioassay studies. Hum Ecol Risk Assess 15(5):858–875
Giknis MLA, Clifford CB (2004) Compilation of spontaneous neoplastic lesions and survival in Crl:CD® rats from control groups. Charles River Laboratories, Worcester
Giknis MLA, Clifford CB (2005) Spontaneous neoplastic lesions in the CrlCD-1(ICR) mouse in control groups from 18 month and 2 year studies. Charles River Laboratories, Worcester
Haseman J (1983) A reexamination of false-positive rates carcinogenesis studies. Fundam Appl Toxicol 3(4):334–343
Haseman J (1984) Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. Environ Health Perspect 58:385–392
Haseman J, Winbush J, O'Donnel M (1986) Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. Fundam Appl Toxicol 7:573–584
Haseman JK, Hailey JR, Morris RW (1998) Spontaneous neoplasm incidences in fischer 344 rats and b6c3f1 mice in two-year carcinogenicity studies: a national toxicology program update. Toxicol Pathol 26(3):428–441
Heimann G, Neuhaus G (1998) Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. Biometrics 54:168–184
Jackson MT (2015) Improving the power of long term rodent bioassays by adjusting the experimental design. Regul Toxicol Pharmacol 72(2):231–342. http://dx.doi.org/10.1016/j.yrtph.2015.04.011

Jara A, Hanson T, Quintana F, Mueller P, Rosner G (2014) Package dppackage. http://www.mat.puc.cl/~ajara

Kodell R, Ahn H (1997) An age-adjusted trend test for the tumor incidence rate for multiple-sacrifice experiments. Biometrics 53:1467–1474

Kodell R, Chen J, Moore G (1994) Comparing distributions of time to onset of disease in animal tumorigenicity experiments. Commun Stat Theory Methods 23:959–980

Lin KK (1995) A regulatory perspective on statistical methods for analyzing new drug carcinogenicity study data. Bio/Pharam Q 1(2):19–20

Lin KK (1997) Control of overall false positive rates in animal carcinogenicity studies of pharmaceuticals. Presentation, 1997 FDA Forum on Regulatory Science, Bethesda MD

Lin KK (1998) CDER/FDA formats for submission of animal carcinogenicity study data. Drug Inf J 32:43–52

Lin KK (2000a) Carcinogenicity studies of pharmaceuticals. In: Chow SC (ed) Encylopedia of biopharmaceutical statistics, 3rd edn. Encylopedia of biopharmaceutical statistics. CRC Press, Boca Raton, pp 88–103

Lin KK (2000b) Progress report on the guidance for industry for statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. J Biopharm Stat 10(4):481–501

Lin KK, Ali MW (1994) Statistical review and evaluation of animal carcinogenicity studies of pharmaceuticals. In: Buncher CR, Tsay JY (eds) Statistics in the pharmaceutical industry, 2nd edn. Marcel Dekker, New York

Lin KK, Ali MW (2006) Statistical review and evaluation of animal carcinogenicity studies of pharmaceuticals. In: Buncher CR, Tsay JY (eds) Statistics in the pharmaceutical industry, 3rd edn. Chapman & Hall, Boca Raton, pp 17–54

Lin KK, Rahman MA (1998) Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. J Biopharm Stat 8(1):1–15

Lin KK, Thomson SF, Rahman MA (2010) The design and statistical analysis of toxicology studies. In: Jagadeesh G, Murthy S, Gupta Y, Prakash A (eds) Biomedical research: from ideation to publications, 1st edn. Wolters Kluwer, New Delhi

Lunn D, Thomas A, Best N, Spiegelhalter D (2000) Winbugs—a bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337

Mancuso J, Ahn H, Chen J, Mancuso J (2002) Age-adjusted exact trend tests in the event of rare occurrences. Biometrics 58:403–412

Moon H, Ahn H, Kodell RL, Lee JJ (2003) Estimation of k for the poly-k test with application to animal carcinogenicity studies. Stat Med 22(16):2619–2636

Peto R, Pike MC, Day NE, Gray RG, Lee PN, Parish S, Peto J, Richards S, Wahrendorf J (1980) Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments. IARC Monogr Eval Carcinog Risk Chem Hum Suppl NIL (2 Suppl):311–426

Portier C, Hoel D (1983) Optimal design of the chronic animal bioassay. J Toxicol Environ Health 12(1):1–19

Portier C, Hedges J, Hoel D (1986) Age-specific models of mortality and tumor onset for historical control animals in the national toxicology programs carcinogenicity experiments. Cancer Res 46:4372–4378

R Core Team (2012) R: A language and environment for statistical computing. http://www.R-project.org/

Rahman MA, Lin KK (2008) A comparison of false positive rates of Peto and poly-3 methods for long-term carcinogenicity data analysis using multiple comparison adjustment method suggested by Lin and Rahman. J Biopharm Stat 18(5):949–958

Rahman MA, Lin KK (2009) Design and analysis of chronic carcinogenicity studies of pharmaceuticals in rodents. In: Peace KE (ed) Design and analysis of clinical trials with time-to-event endpoints. Chapman & Hall/CRC Biostatistics series. Taylor & Francis, Boca Raton

Rahman MA, Lin KK (2010) Statistics in pharmacology. In: Jagadeesh G, Murthy S, Gupta Y, Prakash A (eds) Biomedical research: from ideation to publications, 1st edn. Wolters Kluwer, New Delhi

Rahman MA, Tiwari RC (2012) Pairwise comparisons in the analysis of carcinogenicity data. Health 4:910–918

US Food and Drug Administration—Center for Drug Evaluation and Research (2001) Guidance for industry: statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. US Department of Health and Human Services, unfinalized – draft only

Westfall P, Soper K (1998) Weighted multiplicity adjustments for animal carcinogenicity tests. J Biopharm Stat 8(1):23–44