Regulatory Toxicology and Pharmacology xxx (2015) 1–11

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats

Linda J. Young [a, *], Peter Morfeld [b, c]

[a] Department of Statistics, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, USA
[b] Institute for Occupational Epidemiology and Risk Assessment of Evonik Industries, Essen, Germany
[c] Institute and Policlinic for Occupational Medicine, Environmental Medicine and Preventive Research, University of Cologne, Germany

ARTICLE INFO

Article history:
Received 2 May 2015
Received in revised form
2 December 2015
Accepted 10 December 2015
Available online xxx

Keywords:
Decamethylcyclopentasiloxane D5
Adenocarcinoma
Rat experiment
Statistical analysis
Probit model
Max poly-k trend test
Fisher exact test
Exact logistic regression

ABSTRACT

Decamethylcyclopentasiloxane (D5) is a cyclic siloxane used in the production of industrial and consumer products. Four groups of 60 Fischer-344 female rats were analyzed for uterine endometrial adenocarcinoma (inhalation study with exposure levels in ppm/number of observed cases: 0/0, 10/1, 40/0, and 160/5) by exact regression (logistic, Poisson), the max poly-3 trend test, and a random effects probit model. When comparing the 160 ppm group to controls after 24 months, the incidence of adenocarcinomas was elevated (borderline significant); it was significant when all exposure levels were included. Four sets of (historical) control groups were formed, with varying heterogeneity. The effect of D5 was either significant or borderline significant when comparing all control sets to the 160 ppm group. When considering all exposure groups using any of the analysis methods, a significant effect was observed when the high dose group was included in the analysis; the effect was not significant when the high dose group was not included. The evidence tends to support the conclusion that D5 at the highest dose level (160 ppm) results in an increased incidence of adenocarcinomas. However, it is important to verify any potential effect through a biological investigation.

© 2015 Published by Elsevier Inc.

1. Introduction

Decamethylcyclopentasiloxane (D5), a low-molecular-weight cyclic siloxane, is used in the production of industrial and consumer products with potential exposure to consumers, the general public and manufacturing workers.

As described in the companion paper, Jean et al. (2015) conducted a study to evaluate the possible long-term toxic and oncogenic effects of D5 on Fischer-344 rats. Four groups of 96 males and 96 females were formed, one for each of the D5 exposure levels: 0, 10, 40, and 160 ppm. All animals within a group were placed in a chamber, and the animals' positions within the chamber were rotated on a weekly basis. The rats were exposed for 6 h/day, 5 days/week, to D5 vapor by whole body inhalation route. Four subgroups of rats were formed within an exposure level: (A) 6 months of exposure, (B) 12 months of exposure, (C) 12 months of exposure followed by 12 months of recovery, and (D) 24 months of exposure. The 240 females in the 24-month exposure subgroup (subgroup D), which had 60 female rats at each of the four exposure levels, is the focus of this work.

During the D5 24-month exposure study, mortality, clinical signs, ophthalmoscopical changes, and body weights were recorded. Clinical measurements of hematology, clinical biochemistry, and urinalysis were performed after 3, 6, and 12 months, and hematology was performed after 14 months. Terminal macroscopic examination and organ collection was conducted on all animals. Organs were weighed at scheduled necropsies. Lungs, liver, kidneys, nasal cavities, gross lesions, and tissue masses were examined.

Unexpectedly, five cases of endometrial uterine adenocarcinoma occurred in the highest dose group. Only one additional case

* Corresponding author. Department of Statistics Institute of Food and Agricultural Sciences, 404 McCarty Hall C, P.O. Box 110339, University of Florida, Gainesville, FL 32611-0339, USA.
E-mail address: LJYoung@ufl.edu (L.J. Young).

http://dx.doi.org/10.1016/j.yrtph.2015.12.005
0273-2300/© 2015 Published by Elsevier Inc.

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

was documented. This adenocarcinoma occurred at a dose of 10 ppm (Table 1).

The incidence of endometrial adenocarcinomas within the highest dose group (160 ppm) was compared to that within the control group using the Fisher exact test. Consistent with the approach used by the National Toxicology Program (NTP), a one-sided Fisher exact test was used to assess whether the highest exposure level (160 ppm) group had an increase in the incidence of spontaneous adenocarcinomas compared to the controls. The *p*-value for this one-sided test was 0.0287, indicating a significant increase in spontaneous adenocarcinomas for this highest exposure group. In a review of the study protocol, the concern was raised that the stated hypotheses would indicate that a two-sided, not a one-sided, test was more appropriate. The two-sided Fisher exact test comparing the incidence of spontaneous adenocarcinomas in the highest exposure group to the control group produced a *p*-value of 0.0573. The one-sided test was significant and the two-sided test was not significant, but close to the 0.05 significance level. In either case, sufficient concerns were raised to lead researchers to examine the full D5 24-month study. A statistically significant increase in spontaneous endometrial adenocarcinomas was found, with the increase appearing to be due to the highest exposure level (160 ppm). The biological mechanism that could explain such an effect was explored (Dekant and Klaunig, 2014). When no apparent mechanism was identified, a re-evaluation of the statistical analysis was conducted. Given the large number of statistical tests conducted in the original analysis, could the observed result be a spurious one? The purpose of this work is to provide a full analysis of incidence of endometrial adenomas, adenocarcinomas and combined adenomas and adenocarcinomas based on data from the D5 24-month exposure study.

1.1. Initial analyses

Endometrial adenomas/adenocarcinomas were believed to be relatively rare based on a review of the NTP data on Fischer-344 rats (Haseman et al., 1998). As indicated in Table 1, no adenoma was observed in any of the animals in the D5 study. Five of the 60 Fischer rats receiving the highest dose (160 ppm) had uterine endometrial adenocarcinomas; only one adenocarcinoma was observed in all of the other dose groups. Smith et al. (2005) and Casella et al. (2012) considered various analyses. Since then, a correction was made in the data because a spontaneous adenoma recorded for an animal in the 10 ppm exposure group was later determined to be a focal glandular hyperplasia (Regan, 2014). As in Jean et al. (2015), we use the revised data and begin by summarizing the earlier analyses. Although the incidences of adenomas and adenocarcinomas were analyzed separately as well as combined in the original analyses, no adenomas were observed in this study. Thus, only the adenocarcinomas will be considered here. Further, all analyses are fully described so that this paper can stand alone in the presentation of the statistical analyses.

1.2. Fisher exact tests and Poisson regression

Fisher's exact test was originally developed to test whether the proportion of observations with a positive response is the same for each of two treatments, given the number of observations for each treatment and the total number of positive responses. Under the null hypothesis that the proportions of positive responses are equal for two treatment groups, the test statistic is a random observation from a hypergeometric distribution, and the *p*-value of the test can be determined exactly; that is, computation of the *p*-value does not depend on asymptotic (large sample) theory.

When considering all dose levels, the *p*-value associated with the test of no differences in the proportions with adenocarcinomas, the test statistic is 0.0166 (see Table 2). However, when the highest dose level (160 ppm) is not included in the analysis (leaving the control, 10 ppm, and 40 ppm dose groups), the proportions expressing adenocarcinomas are not significantly different. Further, as noted earlier, the incidence of adenocarcinomas is close to significant ($p = 0.0573$) when using a two-sided test and significant ($p = 0.287$) using a one-sided test.

When the incidence is low, slight changes in the data can result in a substantial change in the *p*-value associated with a test. For example, the 12-month exposure/12-month recovery control group in the D5 study had one adenocarcinoma, instead of the 0 tumors observed in the 24-month exposure control group. If there had been one adenocarcinoma observed in the study control group, as in the 12-month exposure/12-month control group, the *p*-value for the initial one-sided Fisher's exact test comparing the control and highest dose groups would have been 0.1034; the two-sided Fisher's exact test would have had a *p*-value of 0.2068; and the test for the full study would have had a *p*-value of 0.0708. Similarly, if four and not five rats in the 160 ppm dose group had been observed with adenocarcinomas but no rats in the control group had one, the Fisher's exact tests would have *p*-values of 0.0594 and 0.1187, respectively, for the one-sided and two-sided tests comparing the highest exposure group to the control group and 0.0586 for the test comparing all exposure groups.

The power of the Fisher exact test is lower than others because it does not incorporate any supplementary information that may be available. For example, it ignores survival time differences between groups. By accounting for each individual rat's “time at risk”, Poisson regression can be used to estimate and compare the rate of occurrences, *i.e.*, the number of counted cases per observation time (Cameron and Trivedi, 2005; Rothman et al. 2008). Exact Poisson regression analysis, which is based on a permutation approach, provides valid results for small samples (Hirji, 2006; Rothman et al. 2008).

Because not all of the animals survived the full period of the study, the time each was at risk to develop an adenoma or an adenocarcinoma varied. To adjust for these differences in exposure time, survival time was incorporated in an exact Poisson regression comparing the highest dose and the control groups resulted in a two-sided p-value of 0.0657. Thus, incorporating survival time into the analysis did not increase the statistical evidence of an effect.

1.3. Peto test

The Peto test (*see* Heimann and Heuhaus, 2001) considers not only the incidence of adenocarcinomas, but also mortality due to

Table 1
Adenoma/adenocarcinomas observed in D5 24-month exposure study.

Dose	Total animals	Study length unit	Mean life time unit	Adenomas	Adenocarcinomas
0	60	738	708.42	0	0
10	60	738	709.35	0	1
40	60	738	705.65	0	0
160	60	738	722.80	0	5

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Table 2
Results from Fisher's exact test for the D5 24-month exposure study.

Test	*p*-value for tests of equality of proportions with adenocarcinomas
All dose levels	0.0166
Excluding highest dose level (160 ppm)	>0.9999
Control vs. 160 ppm dose group	0.0573

adenocarcinomas. Thus, in addition to knowing whether or not a rat had an adenoma/adenocarcinoma, the time to death and the grade of each lesion (incidental, fatal or mortality independent) are incorporated in the analysis. For lesions with a combination of grades, the observed, expected and variance terms were calculated for the incidental, fatal, and mortality independent grades separately and then combined to produce an overall test statistic and associated *p*-value. Based on the Peto test, the effect of exposure level was significant at the 5% level (Jean et al. 2015).

1.4. Max poly-k trend test

Instead of assuming the cause of death is known as in the Peto test, the poly-k test simply assumes that the time of death is known (Bailar and Portier, 1988). In general, this test is useful when there is a control group and G treated groups, where the ith group receives treatment dose z_i, $i = 0, 1, \ldots, G$. Let π_i be the probability an animal in group i develops a tumor. Consider the null and alternative hypotheses $H_0 : \pi_1 = \pi_2 = \cdots = \pi_G$ versus $H_1 : \pi_i = a + b \times d_i$ for some $b > 0$, where it is assumed there are G doses with $d_1 < d_2 < \cdots < d_G$. That is, the assumption is that the incidence rate is an increasing linear function of the dose. If the incidence rate is not linear, but instead increases only for the highest dose level as suggested by Fisher's exact test, this test will not be as powerful. The following notation will be used.

n_i = number of animals at risk in the ith group
d_i = number of animals with the response of interest in the ith group
e_i = expected number with response of interest in the ith group

For the Cochran–Armitage test (Armitage 1955), it is assumed that all animals are at equal risk of developing a tumor over the duration of the study. If the risk varies over the duration of the study, time strata can be formed so that the risk is equal within a stratum. For the poly-k test, the following modified proportion is tested for trend:

$$r_i = \frac{d_i}{n_i^*}$$

where

$$n_i^* = \sum_{j=1}^{n_i} w_{ij},$$

given the weight assigned at age of death to the jth animal in the ith group is

$$w_{ij} = \begin{cases} 1, & \text{if the } j\text{th animal in the } i\text{th group dies with a tumor} \\ (t_{ij}/t_{\max})^k, & \text{otherwise} \end{cases}$$

and t_{ij} is the survival time and t_{max} is the maximum survival time. The weights w_{ij} reflect the observation that tumors often appear at the rate of a third-to fifth-order polynomial in time (see Portier et al., 1986). Bailar and Portier (1988) recommended the use of $k = 3$, resulting in the poly-3 trend test, which has become standard and is used here. Noting that n_i^* is a random variable, Bieler and Williams (1993) recommended an adjustment to the variance, which is also adopted here.

The isotonic modification of the poly-3 test uses the same poly-3 correction, but the alternative hypothesis is

$$H_2 : \pi_1 \le \pi_2 \le \cdots \le \pi_G$$

with at least one strict inequality (Peddada et al. 2005). This test is more appropriate than the poly-3 test when the response is nonlinear with respect to dose. For example, if increasing exposure levels does not affect the response below a certain threshold exposure level, but does have an effect above that threshold value, the response is nonlinear, and this test is more powerful.

Peddada et al. (2007) proposed combining the two test statistics, W_{BW} from the linear trend test and W_{ISO} from the isotonic trend test, to test H_0 versus (H_1 or H_2). The overall test statistic for the max poly-k trend test is then

$$M_1 = \max\{W_{BW}, W_{ISO}\} \quad (1)$$

Because $M_1 > 0$ and large positive values represent deviations from H_0, the one-sided *p*-value is.

$$p = \Pr_{H_0}(M_1 \ge M_{obs}).$$

To find the one-sided *p*-value, Peddada et al. (2007) considered the asymptotic distribution of the test statistic and Casella et al. (2012) used bootstrapping. Both approaches are considered here. Ten thousand simulations of the asymptotic distribution and 10,000 bootstrap samples were drawn for each test. Because the bootstrap distribution was anticipated to be skewed, the quantiles of the empirical bootstrap distribution were used to establish the *p*-value of each test.

The results including all dose levels and then all but the highest dose levels are presented in Table 3. Note that with the highest dose level included, the incidences of adenocarcinomas are highly significantly different, but neither are close to being significant when the highest dose level is not included. Further, although the conclusions are the same when using either the bootstrap approach or the asymptotic distribution of the test statistic, the *p*-values differ when rounding to the first significant digit. This might warrant further study.

1.5. Probit analysis

An alternative to the previous analyses is to consider a probit analysis of the study where the model reflects the design as in

Table 3
Poly-k max trend test using all dose levels and all but the highest level.

Tumor type	*p*-values all dose levels		*p*-values excluding highest dose level	
	Bootstrap	Asymptotic	Bootstrap	Asymptotic
Adenocarcinoma	0.00172	0.00077	0.30214	0.37891

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Casella et al. (2012). Let z_{ik} be the 0/1 (no tumor/tumor) response of rat k receiving dose i. Then

$$z_{ik} \sim \text{Bernoulli}\ (p_{ik})$$

and

$$\text{probit}(p_{ik}) = \mu + \beta w_i + \gamma t_{ik} \quad (2)$$

where

p_{ik} = probability of a tumor for rat k in dose group i
μ = overall mean
β = treatment effect
t_{ik} = the cube of the proportion of time that rat k in dose group i was alive in the study
γ = time effect
w_i = 0, 1, 4, and 16 for dose group i = 0, 10, 40 and 160, respectively

As in the poly-3 trend test, the use of the cube of the time that a rat is alive in the study reflects the fact that tumors often appear at the rate of a third-to fifth-order polynomial in time.

If the incidence of tumors does not change with dose level, the value of β would be 0 in model (2). As with the poly-3 trend test, if the response increases in a nonlinear manner as suggested by each of the preceding analyses, then the power associated with this test of a linear effect will be less than if a linear trend was present.

A Bayesian analysis was conducted (Lee, 2003). Normal priors with a mean of 0 and a variance of 4 were used for μ, β, and γ. The Gibbs sampler had a burn-in of 10,000 iterations, a thinning rate of 300, and 10,000 recorded samples (Gelfand et al. 1990).

Because the posterior distribution of β is skewed, a credible interval was used to assess whether 0 is a plausible value for the treatment effect β. The 95% credible interval on β does not include 0 (see Table 4). Thus, when considering all dose levels, the effect of dose is significant when considering the incidence of adenocarcinomas. However, if the highest dose of 160 ppm D5 is not included in the analysis, the treatment effect is not significant.

The prior for β has a mean of 0 and a variance of 4, placing 95% of the prior probability to βs between −4 and 4. This represents a relatively flat prior, which is appropriate because no earlier studies of the long-term effects of D5 had been conducted so the data should dominate in the analysis. To ensure that the inferences with respect to β are not sensitive to the choice of prior, a more informative prior normal distribution with a mean of 0 and a variance of 1 was also considered. Some shrinkage of the point estimates of the βs that were not significantly different from 0 was observed. However, the point estimate of β for incidence of adenocarcinoma was 0.096 for standard deviations of 2 and 1. Most importantly, the inferences concerning the significance of β did not change in any case.

1.6. Concurrent and historical controls

Based on the initial findings, researchers began working to identify a biological mechanism to explain the increased incidence of uterine endometrial adenocarcinomas at high levels of chronic D5 exposure, but only slight responses, if any, in various *in vitro* and *in vivo* test systems have been observed. This naturally led to the question as to whether the finding could be spurious. At terminal sampling, a complete necropsy was performed on all animals (see the companion paper Jean et al. 2015 for details). The subsequent measurements made on each animal resulted in a comparison of the treatment groups for each of 129 response variables, making it highly likely that some results would be declared significant if no effort is made to control the false discovery rate. The incidence of endometrial adenocarcinomas in the uterus of F344 rats was believed to be low, based on the review of the NTP data on F344 rats (Haseman et al., 1998). Thus, large numbers of observations are required for precise estimation. To address this concern, additional concurrent and historical controls were used in a re-analysis of the data (Tarone, 1982; Haseman, 1995; EPA, 2005).

For the Fischer 344 rat, uterine endometrial adenomas and adenocarcinomas have been considered to be rare, typically less than 1% at 24 months of age (Maekawa et al., 1983; Haseman et al., 1998) and increasing to 8–12% at 30 months (Nyska et al., 1994). However, more recently published information on control groups challenge the rareness of these tumors in Fischer 344 rats. Nyska et al. (1994) observed significant differences in tumor marker incidences between two groups of unexposed female Fischer-344 rats associated with studies conducted in the same laboratory, under similar conditions, but separated by three years in time. Kuroiwa et al. (2013) reported the incidence of uterine adenocarcinomas in control groups of Fischer 344 rats from chronic bioassays from studies in 1990–1999, 2000–2004 and 2005–2009 to be 3.3%, 12.0%, and 13.5%, respectively. The incidence of uterine adenocarcinomas has been observed to be greater in F344/CrlBr rats (mean of 8% based on Charles River data, 1990 compilation) than in the F344/N rats (0.7% in NTP feeding studies reported by Haseman et al., 1998). Dinse et al., 2010, confirmed the incidence of uterine adenocarcinoma was 0.22% in Fischer 344/N rats. Thus, based on the recent literature, the spontaneous tumor incidences are different for different sub-strains of Fischer 344 rats, and the incidences in sub-strains can change with time. In selecting the control groups for the re-analysis, the studies used the same sub-strain and were conducted over a narrow span of time.

As noted in the section “Initial Analyses,” in parallel of the subgroup (D) experiencing a 24-month exposure protocol another subgroup (C) experienced another 24-month exposure protocol, consisting of 12 months of exposure followed by12 months of recovery period. For the non-exposed control animals in the latter subgroup, this corresponds to 12 months of air-exposure followed by 12 months of non-exposure in the recovery phase. Thus, the only difference in the two subgroups is the second 12 months in which one subgroup continued to be placed in the test chamber (with no exposure) according to the study protocol and the other remained in their home cage. Obviously the conditions for the control animals of subgroup C come very close to the conditions for the control animals of subgroup D; closer than is possible for any other control group not being part of the same study in the same laboratory at the same time as is the case here.

One may argue that the differences in handling in the second

Table 4
Assessment of the effect of dose, β, relative to the incidence of adenocarcinomas from Probit model.

Tumor type	Dose levels (ppm)	Mean	Standard deviation	Percentile of posterior distribution				
				2.5	5	50	95	97.5
Adeno-carcinoma	0, 10, 40, 160	0.096	0.035	0.034	0.043	0.094	0.156	0.171
	0, 10, 40	−0.359	0.405	−1.325	−1.112	−0.299	0.198	0.281

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

twelve months is not important and that the two control groups could then be combined for analysis. However, this slight difference and the fact that, consistent with the design, the two groups were treated as two separate groups throughout the study result in the groups being contemporary historical control data (OECD, 2012, Sections 394–395). Consequently, when they are included in an analysis, any additional heterogeneity that may be associated with the inclusion of that control group is explicitly accounted for in the analysis, just as it is with the historical controls.

Great care should be exercised when selecting historical control groups for inclusion in the analysis of a study. "When historical control data are used, the discussion should address several issues that affect comparability of historical and concurrent control data, such as genetic drift in the laboratory strains, differences in pathology examination at different times and in different laboratories (*e.g.*, in criteria for evaluating lesions; variations in the techniques for the preparation or reading of tissue samples among laboratories), and comparability of animals from different suppliers. The most relevant historical data come from the same laboratory and the same supplier and are gathered within 2 or 3 years one way or the other of the study under review; other data should be used only with extreme caution." (EPA, 2005, p. 48).

In addition to the 12-month exposure/12-month recovery control group in the D5 study, historical control groups from three additional studies, each of 24-month duration, were considered: Octamethylcyclotetrasiloxane (D4) (Lee, 2004), Hexamethyldisiloxane (HMDS) (Dotti et al. 2005), and Polydimethylsiloxane (PDMS) (Mertens, 2003). Each of the studies had a 12-month exposure/12-month recovery and a 24-month exposure control group, both of which were considered as a potential historical control. All studies used the Fischer-344 rats of the same sub-strain from the same source (see Table 5). All rats were received from the source within 28 months of the start of the D5 study, and all but the HMDS study controls were received within 13 months of study initiation. Further, the rats were all of a similar age. The studies were all conducted by Dow Corning. However, the testing facilities varied within and among studies. This source of variation must be accounted for in the analysis.

The control groups were comprised of 20–65 animals that had little to no incidences of adenomas/adenocarcinomas (see Table 6). To further assess the appropriateness of including these historical control groups, the mortality and a set of markers for tumor incidence were compared among the control groups and specifically between the D5 24-month exposure and each of the other control groups. The markers for tumor incidence were spleen mononuclear cell leukemia, pituitary *P. Distalis* adenoma, thyroid gland c-cell adenoma, uterine stromal polyps, and mammary gland fibroadenoma.

1.7. Control group mortality

From Table 7, the survival rates among control groups are not significantly different ($p = 0.3498$ using Fisher's exact test). Yet, this test does not consider the time that each rat survives, only whether or not it is alive at the end of the study. The product-limit estimate of the survival function considers the additional information of survival time and the fact that a rat that is alive at study's end has an unknown survival time. The estimated functions were compared using the log-rank test and the Wilcoxon test and found to be highly significantly different (see Table 8). The plots of the estimated survival functions are displayed in Fig. 1.

Although the finding that at least one of the survival functions differs from the others is important, primary interest lies in assessing the difference, if any, between the survival function of the D5 24-month exposure control group and the survival function of each potential historical control. Thus, the survival functions of the other seven control groups were compared to the D5 24-month exposure control group (see Table 9). Recall that this analysis compares survival functions and thus considers all information on time of death. For animals that survived, it is known that they have survived until the study's end, but it is unknown how much longer they could have survived. This helps explain the result that both HMDS groups separate so clearly from the others, given that the proportions of rats surviving until the end of the study for the HDMS groups are similar to the proportions for the PDMS control groups. Consider the control groups from these two studies. The earliest death in the HMDS group was on May 20, 1999, just a little more than 3 months prior to the study's termination at the end of August, 1999. In contrast, the earliest death in the PDMS groups was on February 14, 2000, a little more than 9 months prior to the study's conclusion about November 20, 2000. These comparatively earlier deaths for some PDMS rats relative to the HMDS rats contributed to the projections that the PDMS groups would tend to have a shorter survival time than the HMDS groups. It may also be important that the HMDS study was conducted more than a year earlier than any of the other studies.

In summary, both HMDS control groups have significantly longer survival times than the D5 24-month exposure control group. The other control groups do not have significantly different survival times than the D5 24-month control group. We note that a non-significant test result does not establish equivalence (Hauck and Anderson, 1984).

1.8. Control group tumor incidence

The following tumors represent a selected group of neoplastic lesions that are commonly observed in control animals in chronic bioassays: pituitary *P. Distalis* adenoma, thyroid gland c-cell adenoma, uterine stromal polyps, and mammary gland fibroadenoma. Their use here affords a generalized selection of well recognized neoplastic lesions for comparison among the individual chronic bioassay studies of interest. If a set of controls differs significantly from each other, and more specifically from the D5 24-month exposure control group, then it may not be an appropriate

Table 5
Characteristics of the control groups.

Control group	Source	Strain/sub-strain	Date of animal receipt	Age at start of study
D4 12 exposure/12 recovery	Charles River	F344/CrlBR	1/12/1999	7–8 weeks
D4 24-month exposure	Charles River	F344/CrlBR	1/12/1999	7–8 weeks
D5 12 exposure/12 recovery	Charles River	F344/CrlBR	12/1/1999	6 weeks
D5 24-month exposure	Charles River	F344/CrlBR	12/1/1999	6 weeks
HMDS 12 exposure/12 recovery	Charles River	F344/CrlBR	8/20/1997	6 weeks
HMDS 24-month exposure	Charles River	F344/CrlBR	8/20/1997	6 weeks
PDMS 12 exposure/12 recovery	Charles River	F344/CrlBR	11/3/1998	7–8 weeks
PDMS 24-month exposure	Charles River	F344/CrlBR	11/3/1998	7–8 weeks

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Table 6
Incidence of adenomas/adenocarcinomas in the control groups.

Control group	Number of animals	Adenoma	Adenocarcinoma	Combined
D4 12 exposure/12 recovery	20	0	0	0
D4 24-month exposure	59	0	0	0
D5 12 exposure/12 recovery	20	0	1	1
D5 24-month exposure	60	0	0	0
HMDS 12 exposure/12 recovery	20	0	1	1
HMDS 24-month exposure	65	1	1	2
PDMS 12 exposure/12 recovery	20	0	0	0
PDMS 24-month exposure	60	2	0	2

Table 7
Survival rates within control groups.

Treatment	Total	Deaths	Survivals	Percent survival
D4 12 exposure/12 recovery	20	6	14	70.00
D4 24-month exposure	60	17	43	71.67
D5 12 exposure/12 recovery	20	6	14	70.00
D5 24-month exposure	60	14	46	76.67
HMDS 12 exposure/12 recovery	20	3	17	85.00
HMDS 24-month exposure	65	11	54	83.08
PDMS 12 exposure/12 recovery	20	3	17	85.00
PDMS 24-month exposure	60	8	52	86.67
Totals	325	68	257	79.08

Table 8
Results from testing equality of 8 survival functions.

Test of equality over treatments			
Test	Chi-square	DF	Pr > Chi-Square
Log-Rank	31.5471	7	<0.0001
Wilcoxon	30.9812	7	<0.0001

control group for assessing the effect of D5. The numbers of rats that are positive and negative for these markers within each control group are displayed in Tables 10–14. Note that, in some cases, a measurement(s) could not be made on an animal so some data are missing, resulting in the total number of observations being less than the number of rats within the group. It is assumed that these data are missing at random.

Fisher's exact test was used first to compare all control groups for each of the markers of tumor incidence (Table 15) and then to do pairwise comparisons with the D5 24-month exposure control group (Table 16). Although some indication of differences among groups was found for spleen mononuclear cell leukemia, pituitary *P. Distalis* adenoma, and mammary gland fibroadenoma, in each case, the key comparison is with the D5 24-month exposure control group. When making these comparisons, the only concerns for differences are raised for the *P. Distalis* adenoma, which tended to be higher for the D4 24-month exposure, PDMS 24-month exposure, and the PDMS 12-month exposure/12-month recovery control groups.

1.9. Summary of control group findings

Some indications of differences have been found between the D5 24-month exposure control group and the D4 24-month exposure control group and each of the PDMS control groups. However, as the number of tests increases, the probability of a false positive, also called a false discovery, increases. The Benjamini-Hochberg procedure (BH step-up procedure) (Benjamini and Hochberg, 1995) controls the rate of false positives at a specified level, say α. To conduct the procedure for m hypothesis tests, the p-values for the tests are ordered from smallest to largest: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Note: The parentheses are used to denote the

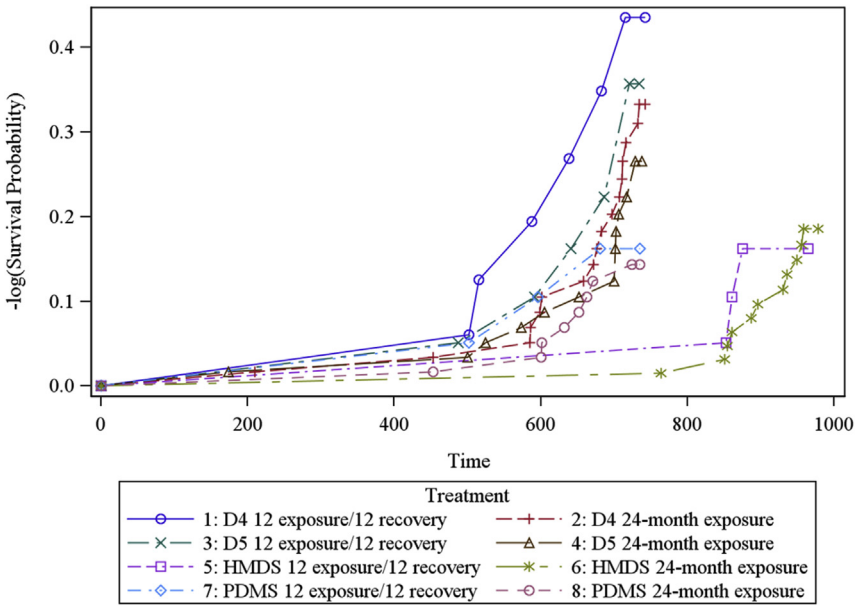

Fig. 1. Plot of negative log of the estimated survival function from seven control groups.

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Table 9
Comparison of mortality in control groups to D5 24-month exposure control group.

Test of equality with D5 24-month exposure control group						
Control group	Log−rank test			Wilcoxon test		
	Chi-square	DF	Pr>Chi-square	Chi-square	DF	Pr>Chi-square
D5 12 Exposure/12 Recovery	0.5572	1	0.4554	0.5678	1	0.4511
D4 24-Month Exposure	0.3337	1	0.5635	0.2932	1	0.5882
D4 12 Exposure/12 Recovery	1.2299	1	0.2674	1.3697	1	0.2419
HMDS 24-Month Exposure	17.0695	1	<0.0001	17.0040	1	<0.0001
HMDS 12 Exposure/12 Recovery	5.2677	1	0.0217	5.2400	1	0.0221
PDMS 24-Month Exposure	1.8324	1	0.1758	1.6969	1	0.1927
PDMS 12 Exposure/12 Recovery	0.4985	1	0.4802	0.4048	1	0.5246

Table 10
Observed incidences of spleen mononuclear cell leukemia for all control groups.

Control group	Positive	Negative	Percent positive
D4 12 exposure/12 recovery	6	14	30.0
D4 24-month exposure	14	45	23.7
D5 12 exposure/12 recovery	3[c]	17	15.0
D5 24-month exposure	13[c]	47	21.7
HMDS 12 exposure/12 recovery	5[c]	15	25.0
HMDS 24-month exposure	10[c]	55	15.4
PDMS 12 exposure/12 recovery	7[d]	13	35.0
PDMS 24-month exposure	12[d]	48	20.0

[c] Hemolymphoretic system used for “spleen”.
[d] Large granular lymphocyte leukemia (under “organ” systemic tumors).

Table 11
Observed incidences of pituitary *P. Distalis* adenoma for all control groups.

Control group	Positive	Negative	Percent positive
D4 12 exposure/12 recovery	5	15	25.0
D4 24-month exposure	27	31	46.6
D5 12 exposure/12 recovery	5	15	25.0
D5 24-month exposure	12	48	20.0
HMDS 12 exposure/12 recovery	7	13	35.0
HMDS 24-month exposure	18	46	28.1
PDMS 12 exposure/12 recovery	9	11	45.0
PDMS 24-month exposure	22	37	37.3

Table 12
Observed incidences of thyroid gland C-Cell adenoma for all control groups.

Control group	Positive	Negative	Percent positive
D4 12 exposure/12 recovery	3	17	15.0
D4 24-month exposure	4	55	6.8
D5 12 exposure/12 recovery	0	20	0.0
D5 24-month exposure	4	56	6.7
HMDS 12 exposure/12 recovery	1	19	5.0
HMDS 24-month exposure	2	63	3.1
PDMS 12 exposure/12 recovery	1	19	5.0
PDMS 24-month exposure	5	55	8.3

Table 13
Observed incidences of uterine stromal polyps for all control groups.

Control group	Positive	Negative	Percent positive
D4 12 exposure/12 recovery	1	19	5.0
D4 24-month exposure	11	48	18.6
D5 12 exposure/12 recovery	3	17	15.0
D5 24-month exposure	10	50	16.7
HMDS 12 exposure/12 recovery	6	14	30.0
HMDS 24-month exposure	17	48	26.2
PDMS 12 exposure/12 recovery	2	18	10.0
PDMS 24-month exposure	16	44	26.7

Table 14
Observed incidences of mammary gland fibroadenoma for all control groups.

Control group	Positive	Negative	Percent positive
D4 12 exposure/12 recovery	3[a]	17	15.0
D4 24-month exposure	13[a]	46	22.0
D5 12 exposure/12 recovery	2	17	10.5
D5 24-month exposure	8	52	13.3
HMDS 12 exposure/12 recovery	0	20	0.0
HMDS 24-month exposure	4	61	6.2
PDMS 12 exposure/12 recovery	1	19	5.0
PDMS 24-month exposure	2	57	3.4

[a] This value is composed of both diagnoses: Fibroadenoma + Fibroadenoma, multiple.

Table 15
Fisher's exact test comparing tumor incidence across all control groups.

Tumor type	Test statistic	p-value
Spleen mononuclear cell leukemia	6.286E-07	0.6026
Pituitary *P. Distalis* adenoma	5.922E-09	0.0684
Thyroid gland C-cell adenoma	3.131E-05	0.5920
Uterine stromal polyps	1.339E-07	0.2525
Mammary gland fibroadenoma	3.862E-08	0.0180

ordered p-values whereas subscript i without parentheses (p_i) denotes the p-value corresponding to the ith hypothesis. Here one test for equality of survival rates and five for the equality of tumor incidence have been conducted in comparing each control group to the D5 24-month exposure control group. Thus, m is equal to six. The procedure is performed as follows:

1. For a given α, find the largest k such that $p_{(k)} \leq k\alpha/m = 0.05\ k/6 = 0.00833\ k$.
2. Reject all hypotheses (declare positive discoveries) for all $H_{(i)}$, $i = 1, 2, \ldots, k$.

Using the BH step-up procedure, the D4 24-month exposure control group is found to have a significantly higher incidence of pituitary *P. Distalis* adenoma than the D5 24-month exposure control group. Further, the HMDS 24-month exposure control group is found to have significantly higher survival rates than the D5 24-month exposure control group, but the differences in the incidences of pituitary *P. Distalis* adenoma between the D5 24-month exposure and either the PDMS 24-month or the PDMS 12-month exposure/12-month recover control groups are not significant. Further, the survival rates for the HMDS 12-month exposure/12-month recovery control group are not significantly different from those for the D5 24-month exposure control group. Thus, we found a significant heterogeneity in response between different groups of Fischer-344 control rats although all rats were received within 28 months of the start of the D5 study from the same

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Table 16
Comparison of Tumor Incidence in Control Groups to D5 24-Month Exposure Control Group: Fisher's Exact Test *p*-values.

Control group	Spleen Mononuclear cell leukemia	Pituitary *P. Distalis* adenoma	Thyroid gland C-Cell adenoma	Uterine stromal polyps	Mammary gland fibroadenoma
D5 12 Exposure/12 Recovery	0.7485	0.7533	0.5673	>0.9999	>0.9999
D4 24-month exposure	0.8294	0.0031	>0.9999	0.8140	0.2378
D4 12 exposure/12 recovery	0.5456	0.7533	0.3582	0.2750	>0.9999
HMDS 24-month exposure	0.4890	0.3042	0.4259	0.2768	0.2286
HMDS 12 exposure/12 recovery	0.7634	0.2257	>0.9999	0.2107	0.1906
PDMS 24-month exposure	>0.9999	0.0438	>0.9999	0.2677	0.0946
PDMS 12 exposure/12 recovery	0.2466	0.0401	>0.9999	0.7203	0.4375

supplier and although all rats were of the same sub-strain.

Based on these results, it was decided to conduct four sets of analyses, each with a different set of controls (see Table 17). The first is simply the concurrent controls, which have already been considered. The other sets are increasingly heterogeneous, allowing the impact of this heterogeneity to be explored. The second set includes both the D5 12-month exposure/12-month recovery and the concurrent control groups. The D5 12-month exposure/12-month recovery control group was in the same D5 study. They were treated as a group distinct from the 24-month exposure control group during the study. This and the small differences in handling discussed earlier make it necessary for them to be treated as contemporary historical controls. Consequently, this control group is not simply combined with the D5 24-month exposure control group for analysis. Instead, the potential additional heterogeneity due to the differences in handling is explicitly accounted for. The third set includes the two D5 study control groups as well as the control groups that were not significantly different from the D5 24-month exposure control group with respect to survival rate or the incidences of any of the five selected markers for tumors. The final set of control groups included all of those in set 3 as well as the control groups that had significantly different survival rates or tumor incidence of a marker tumor from the D5 24-month exposure control group. Thus, each set reflects increasing heterogeneity among the control groups.

A basic assumption in all scientific studies is that of exchangeability. Here the assumption implies that rats receiving the highest dose have the same susceptibility to adenocarcinomas as rats in the control group. The randomization of the rats to exposure levels in the D5 24-month exposure study provides a foundation for making the exchangeability assumption in that each rat was equally likely to have been in any exposure group. Further, for the full D5 study, rats were also randomized to subgroups. Thus, the assumption of exchangeability is also reasonable when including the D5 12-month exposure/12-month recovery group as a concurrent historical control. Because the other historical control groups were not subject to the randomization process for the study of interest, the question of whether the rats are truly exchangeable can be raised. To rely on exchangeability, one must rely on a biological argument. All control groups considered here were from the same source, strain, and sub-strain. They were also of similar ages at study onset. Yet, significant differences were found among them. This could be due to differences in laboratories in which the studies were conducted or to other study differences. Further, as discussed earlier, the significant differences could be due to drift in the strain over time. Thus, as control set 3 and then set 4 is added to the analysis, the question as to whether the assumption of exchangeability is met becomes an increasing concern.

1.10. Analyses incorporating historical controls

When adding historical controls, it is important to account for the additional heterogeneity introduced when incorporating them into the analyses. Three such analyses are considered in this paper. The first is Fisher's exact test. The second is an extension of the max poly-k trend test discussed earlier. The final one is a meta-analysis developed by Casella et al. (2012).

1.11. Fisher exact test

Fisher's exact test can be used to test whether the proportion of observations with a positive response is the same for each of t treatment groups, given the number of observations for each treatment and the total number of positive responses. As when considering only two treatment groups, the p-value of the test can be determined exactly and does not depend on asymptotic (large sample) theory. Control groups within each set were compared with respect to the incidence of adenocarcinomas. In no case was there a significant difference. To account for heterogeneity associated with different groups of control animals, the control groups were treated as separate groups in all analyses.

As with the initial analysis, a preliminary test of the control groups versus the 160 ppm exposure group was conducted first using each set of controls (see Table 18). Here the Fisher exact test weights each of the control groups equal to the exposure group so the test is not only comparing exposure and control groups but also control groups within the same set. The difference in incidences in controls and the highest exposure level was not significant for the first two sets of controls (set 1, the control group associated with the D5 24-month exposure study and set 2, the control groups

Table 17
Sets of control groups.

Set	Study control groups
1	D5 24-month exposure
2	D5 24-month exposure; D5 12-month exposure/12-month recovery
3	D5 24-month exposure; D5 12-month exposure/12-month recovery; D4 12-month exposure/12-month recovery; HMDS 12-month exposure/12-month recovery; PDMS 24-month exposure; PDMS 12-month exposure/12-month recovery
4	D5 24-month exposure; D5 12-month exposure/12-month recovery; D4 12-month exposure/12-month recovery; HMDS 12-month exposure/12-month recovery; PDMS 24-month exposure; PDMS 12-month exposure/12-month recovery; D4 24-month exposure; HMDS 24-month exposure

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Table 18
p-Values From Fisher's exact test of constant proportions. Using all dose levels and all but the highest level.

Control set	Dose levels (ppm)	Adenocarcinoma
Set 1	0, 160	0.0573
	0, 10, 40, 160	0.0166
	0, 10, 40	>0.9999
Set 2	0, 160	0.0659
	0, 10, 40, 160	0.0173
	0, 10, 40	0.1905
Set 3	0, 160	0.0313
	0, 10, 40, 160	0.0188
	0, 10, 40	0.1554
Set 4	0, 160	0.0078
	0, 10, 40, 160	0.0188
	0, 10, 40	0.1554

associated with the D5 studies), though they were only slightly above the 0.05 level. However, for control sets 3 and 4, the effect of the highest level of exposure was significant.

Consider the cases where the difference between control groups and the highest exposure level is significant or close to being significant (as in the cases of *p*-values of 0.0573 and 0.0659). When the full study is analyzed, the effect of exposure is found to be significant when the highest dose is included and is not significant when it is not included (see Table 18).

Although not presented here, an analysis with similar results could be conducted using exact Poisson regression.

1.12. Extended max poly-K trend test

The max poly-k trend test compares the tumor rates in the control and treatment groups using information from (1) the poly-3 extension to the Cochran–Armitage linear trend test with the Bieler-Williams variance adjustment and (2) the isotonic trend test. Recall that the max poly-k trend test statistic is the maximum of the test statistics from these two tests. For the extended max poly-k trend test, Peddada et al. (2007) proposed using the maximum of two test statistics. The first is M_1 of the max poly-k trend test, which incorporates only the current control group as in (1). The second is M_2, which is analogous to M_1, except that it is based on the historical control groups and reflects their heterogeneity. Then, the test statistic is $M = \max(M_1, M_2)$. As in Casella et al. (2012), bootstrap methods were used to determine the one-sided *p*-values associated with each test.

The incidence of adenocarcinoma is significant if the 160-ppm dose level is included, and it is not significant if the 160-ppm dose level is not included (Table 19). Thus, for this test, the highest dose group is driving the conclusion of a highly significant effect, and the conclusion does not depend on which set of controls is used.

Table 19
p-Values from poly-k max trend test of constant proportions using all dose levels and all but the highest level.

Control set	Dose levels (ppm)	Adenocarcinoma
Set 1	0, 10, 40, 160	0.0017
	0, 10, 40	0.3021
Set 2	0, 10, 40, 160	0.0017
	0, 10, 40	0.6218
Set 3	0, 10, 40, 160	0.0017
	0, 10, 40	0.5043
Set 4	0, 10, 40, 160	0.0014
	0, 10, 40	0.4734

1.13. Meta-analysis

Another approach to accounting for the heterogeneity introduced by using the historical controls is to include that source of variability in the probit analysis considered earlier. To do this, again let z_{ijk} be the 0/1 (no tumor/tumor) response of rat k receiving dose i in study j. As before,

$$z_{ijk} \sim \text{Bernoulli}\left(p_{ijk}\right)$$

but the model becomes

$$\text{probit}\left(p_{ijk}\right) = \mu + \beta w_i \mathrm{I}(j = 0) + s_j + \gamma t_{ik} \tag{3}$$

where

p_{ijk} = probability of a tumor for rat k in dose group i
μ = overall mean
β = treatment effect
w_i = 0, 1, 4, and 16 for dose group i = 0, 10, 40 and 160, respectively
$s_j \sim N(0, \sigma^2_{s_j})$ = random effect of jth study
t_{ijk} = the cube of the proportion of time that rat k in dose group i was alive in the study
γ = time effect

The indicator function $\mathrm{I}(j = 0)$ reflects the fact that dose levels other than the control (0 ppm) are only considered for the current study. The prior distribution for $s_j|\mu, \sigma^2_s$ was $N(\mu, \sigma^2_s)$. Normal priors with a mean of zero and a variance of four were used for μ, β, and γ. The prior distribution of σ^2_s was exponential with a mean of two. The Gibbs sampler had a burn-in of 10,000 iterations, a thinning rate of 300, and 10,000 recorded samples.

As with the original probit analysis, interest lies in whether or not there is a treatment effect. Thus, the test of interest is of H_0: $\beta = 0$ versus H_1: $\beta \neq 0$. The mean and standard deviation of the posterior distribution as well as percentiles of the posterior distribution were computed for each set of historical controls (see Table 20). The 2.5 and 97.5 percentiles correspond to a test of the hypotheses at the 5% level. If both endpoints are positive (or negative), the null hypothesis is rejected, and the conclusion is that the incidence of tumors increases (decreases) with dose. To ensure that the inferences on β are not sensitive to the choice of prior, a prior with a mean of zero and a variance of one was also considered. The results did not change in any meaningful way.

Regardless of which set of controls is used in the analysis, the conclusions are the same and agree with the results based solely on concurrent controls. (Note: In reviewing the results of Casella et al., 2012 in the context of this work, a small programming error was found in the code. This was corrected with the help of those authors, and the results presented here reflect the revised code.) The effect of D5 is significant for adenocarcinomas. However, if the highest dose of 160 ppm is excluded from the analysis, the effect of D5 is not significant, not even marginally so.

2. Discussion

Five uterine endometrial adenocarcinomas were observed in the highest dose group of 160 ppm. Only one was found in the other dose groups, and it was in the lowest positive dose group of 10 ppm. When using Fisher's exact test for testing the effect of exposure, the incidence of adenocarcinomas in the highest dose group compared to the control group was either significantly different or close (p = 0.0572 or 0.0659) to being significantly different from the

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005

Table 20
Assessment of the Effect of Dose, β, Relative to the incidence of adenocarcinomas from Probit model using various control sets.

Tumor type	Dose Levels (ppm)	Mean	Standard deviation	Percentile of posterior distribution				
				2.5	5	50	95	97.5
Control	0, 10, 40, 160	0.077	0.030	0.023	0.031	0.075	0.128	0.140
Group 2	0, 10, 40	−0.451	0.410	−1.422	−1.234	−0.385	0.084	0.157
Control	0, 10, 40, 160	0.084	0.028	0.032	0.040	0.083	0.133	0.144
Group 3	0, 10, 40	−0.319	0.366	−1.184	−1.006	−0.253	0.159	0.225
Control	0, 10, 40, 160	0.076	0.026	0.029	0.0370	0.075	0.121	0.131
Group 4	0, 10, 40	−0.378	0.372	−1.261	−1.093	−0.308	0.095	0.154

control. When all exposure groups were included in the analysis, a significant effect was observed when the high dose group was included in the analysis, but the effect was not significant when the high dose group was not included. These results do not depend on which test was used.

Fisher's exact test and the test based on exact Poisson regression can be conducted using most major statistical software packages, such as SAS, Stata and R. The Bayesian probit model was programmed using JAGS in R and could be programmed in other software packages that accommodate Bayesian models. However, this particular Bayesian model is somewhat challenging to implement and routine implementation would be facilitated by having a specialized R package focusing on this type of model. The max poly-k trend test is based on newer methods and, to our knowledge, software packages that can be used to conduct this test are not readily available. As software does become more accessible, the max poly k trend test and the Bayesian probit model will surely be used more because they have better statistical properties and can account for heterogeneity introduced when concurrent and historical controls are used in the analysis.

Because the highest dose of D5 (160 ppm) is clearly driving the significant effects observed in these analyses, a threshold effect may be present. Models can be used to estimate the threshold. However, these are not explored here.

Jean et al. (2015) present a summary of the test results for 129 responses observed for the female rats in the 24-month inhalation study (see Smith et al., 2005, for details). For seven responses, the test could not be conducted due to insufficient data. Among the 122 one-sided Fisher exact tests performed, two were significant at the 5% level. One test indicated an increase and the other a decline in response as the dose level increased. The comparison between adenocarcinoma incidence for dose groups 160 ppm and 0 ppm, which had a one-sided p-value from the Fisher exact test of 0.0287, was the test resulting in a significant increase in response. However, based on the Benjamini and Hochberg (1995) adjustment for false discoveries, the first and second smallest p-values associated with the two tests that were significant at the 5% level would need to be less than 0.0002 and 0.0004, respectively, to be declared significant. In either case, the observed value of 0.0287 is not significant, if multiple testing is considered.

Given the large number of tests conducted in the study, the question naturally rises as to whether the result could be simply a chance finding. As with any statistical test, this is a possibility; the p-value simply provides the probability of such a finding if there is no effect. The evidence tends to support the conclusion that D5 at the highest dose level results in an increased incidence of adenocarcinomas, but not adenomas. However, the biological investigation by Klaunig et al. (2015) addresses whether an excess is plausible on mechanistic grounds and the potential biological relevance.

As noted in the EPA (2005, p. 48) guidelines, "In cases where there may be reason to discount the biological relevance to humans of increases in common animal tumors, such considerations should be weighed on their own merits and clearly distinguished from statistical concerns." To assess human relevance, besides the statistics, the whole toxicological profile of the substance must be considered in an overall evaluation.

Conflicts of interest

The work was funded by the American Chemistry Council (ACC, http://www.americanchemistry.com/). Evonik Industries applies D5 in the production of special chemicals (http://www.evonik.com).

Acknowledgment

Data for both treated and control rats were provided by Dow Corning. The methods presented here build upon the work of George Casella and his co-workers, Andrew Womack and Luis Leon Novelo. Andrew Womack provided all programs the group had used and corrected the probit analysis program.

References

Armitage, P., 1955. Tests for linear trends in proportions and frequencies. Biometrics 11, 375–386.
Bailar, A.J., Portier, C.J., 1988. Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics 44, 417–431.
Bieler, G.S., Williams, R.L., 1993. Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. Biometrics 49, 793–801.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.
Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics. Methods and Applications. University Press, Cambridge.
Casella, G., Leon-Novelo, L., Womack, A., 2012. Analysis to Determine Effects of D5 on Fischer 344 Rats (Final Report).
Dekant, W., Klaunig, J.E., 2014. Toxicology of Decamethylcyclopentasiloxane (D5). Submitted to Regulatory Toxicol. Pharmacol.
Dinse, G.E., Peddada, S.D., Harris, S.E., Elmore, S.A., 2010. Comparison of NTP historical control tumor incidence rates in female Harlan Sprague Dawley and Fischer 344/N rats. Toxicol. Pathol. 38, 765–775.
Dotti, A., Smith, P.A., Chevalier, H.J., 2005. Hexamethyldisiloxane: a 24-Month Combined Chronic Toxicity and Oncogenicity Whole Body Vapor Inhalation Study in Fischer-344 Rats. Study No. 2004_I0000–53896.
Gelfand, A.E., Hills, S.E., Racine-Poon, A.S., Adrian, F.M., 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. J. Am. Stat. Assoc. 85, 972–985.
Haseman, J.K., 1995. Data analysis: statistical analysis and use of historical control data. Regul. Toxicol. Pharmacol. 21, 52–59.
Haseman, J.K., Hailey, J.R., Morris, R.W., 1998. Spontaneous neoplasm incidences in Fischer 344 rats and B6C3F1 mice in two-year carcinogenicity studies: a National Toxicology program update. Toxicol. Pathol. 26, 428–441.
Hauck, W.W., Anderson, S., 1984. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. J. Pharmacokinet. Biopharm. 12, 83–91.
Heimann, G., Heuhaus, G., 2001. On the asymptotic distribution for Peto's combined test for carcinogenicity assays under equal and unequal censoring. Biometrika 88, 435–445.
Hirji, K., 2006. Exact Analysis of Discrete Data. Chapman and Hall, Boca Raton FL.
Jean, Paul A., Plotzke, K.P., Scialli, A.R., 2015. Chronic toxicity and oncogenicity of Decamethylcyclopentasiloxane in the Fischer 344 rat. Submitted to Regulatory Toxicol. Pharmacol.
Klaunig, J.E., Dekant, W., Plotzke, K., Scialli, A.R., 2015. Bioloigcal relevance of

ARTICLE IN PRESS

Decamethylcyclopentasiloxane (D5): analysis of the potential mode of action of Decamethylcyclopentasiloxane induced rat uterine endometrial adenocarcinoma tumorigenesis. Submitted to Regulatory Toxicol. Pharmacology.
Kuroiwa, Y., Ando, R., Kasahara, K., Nagatani, M., Yamakawa, S., Okazaki, S., 2013. Transition of historical control data for high incidence tumors in F344 rats. J. Toxicol. Pathol. 24, 227–230.
Lee, M., 2004. 24-Month Combined Chronic Toxicity and Oncogenicity Whole Body Vapor Inhalation Study of Octamethylcyclotetrasiloxane (D4) in Fischer 344 Rats. Study No. 2004-I0000-54091.
Lee, P.M., 2003. Bayesian Statistics: an Introduction, second ed. Arnold, London.
Maekawa, A., Kurokawa, Y., Takahashi, M., Kokubo, T., Ogiu, T., Onodera, H., Tanigawa, H., Ohno, Y., Furukawa, F., Hayashi, Y., 1983. Sponteaneous tumors in F-err/DuCrj rats. Gann 74, 365–372.
Mertens, J.W.M., 2003. A 24-Month Combined Chronic Toxicity and Oncogenicity Dietary Study of Polydimethylsiloxane (PDMS) 10 Cst Fluid in Fischer 344 Rats. Report No. 2003-I0000-53254.
Nyska, A., Klein, T., Skolnik, M., Wayner, T., Klein, B., 1994. Unusually high incidence of spontaneous endometrial adenocarcinoma in aged virgin Fischer rats. Exp. Toxicol. Pathol. 46, 7–9.
Organisation for Economic Co-operation and Development (OECD), 2012. Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies. Supporting Test Guidelines 451, 452 and 453. 2nd Ed. Series on Testing and Assessment. No. 116.
Peddada, S.D., Dinse, G.E., Haseman, J., 2005. A survival-adjusted quantile response test for comparing tumor incidence rates. Appl. Stat. 54, 51–61.
Peddada, S.D., Dinse, G.E., Kissling, G.E., 2007. Incorporating historical control data when comparing tumor incidence rates. J. Am. Stat. Assoc. 102, 1212–1220.
Portier, C., Hedges, J., Hoel, D., 1986. Age-specific models of mortality and tumor onset for historical control animals in the National Toxicology Program's carcinogenicity experiments. Cancer Res. 46, 4372–4378.
Regan, K. 2014. Pathology report: images and descriptions of uterine epithelial tumors from a two-year chronic bioassay with Decamethylcyclopentasiloxane (Study DCC9346/RCC753390). Prepared by ReganPath/Tox Services Inc., for SEHSC.
Rothman, K.J., Greenland, S., Lash, T.L., 2008. Modern Epidemiology, 3. ed. Lippincott Williams & Wilkins, Philadelphia.
Smith, P.A., Burri, R., Chevalier, H.J., 2005. Decamethylcyclopentasiloxane (D5): a 24-month Combined Chronic Toxicity and Oncogenicity Whole Body Vapour Inhalation Study in Fischer-344 Rats. 2005. RCC laboratories. Report No. 2005-I0000-54953; Study No. 9346; Report date 2005-06-09.
Tarone, R.E., 1982. The use of historical control information in testing for a trend in proportions. Biometrics 38, 215–220.
US Environmental Protection Agency (EPA), 2005. Guidelines for Carcinogen Risk Assessment. U.S. Environmental Protection Agency, Washington, DC. EPA/630/P-03/001F.

Please cite this article in press as: Young, L.J., Morfeld, P., Statistical considerations for a chronic bioassay study: Exposure to Decamethylcyclopentasiloxane (D5) and incidence of uterine endometrial adenocarcinomas in a 2-year inhalation study with Fischer rats, Regulatory Toxicology and Pharmacology (2015), http://dx.doi.org/10.1016/j.yrtph.2015.12.005