Published in final edited form as:

Toxicol Lett. 2015 September 2; 237(2): 161–164. doi:10.1016/j.toxlet.2014.09.016.

Proper interpretation of chronic toxicity studies and their statistics: A critique of “Which level of evidence does the US National Toxicology Program provide? Statistical considerations using the Technical Report 578 on *Ginkgo biloba* as an example”

Grace E. Kissling^{a,*}, Joseph K. Haseman^b, and Errol Zeiger^c

^aNational Institute of Environmental Health Science, P.O. Box 12233, Research Triangle Park, NC 27709

^bJ.K. Haseman Consulting, 1054 Tacketts Pond Drive, Raleigh, NC 27614-7886

^cErrol Zeiger Consulting, 800 Indian Springs Road, Chapel Hill, NC 27514

Abstract

A recent article by Gaus (2014) demonstrates a serious misunderstanding of the NTP's statistical analysis and interpretation of rodent carcinogenicity data as reported in Technical Report 578 (*Ginkgo biloba*) (NTP 2013), as well as a failure to acknowledge the abundant literature on false positive rates in rodent carcinogenicity studies. The NTP reported *Ginkgo biloba* extract to be carcinogenic in mice and rats. Gaus claims that, in this study, 4800 statistical comparisons were possible, and that 209 of them were statistically significant ($p < 0.05$) compared with 240 (4800×0.05) expected by chance alone; thus, the carcinogenicity of *Ginkgo biloba* extract cannot be definitively established. However, his assumptions and calculations are flawed since he incorrectly assumes that the NTP uses no correction for multiple comparisons, and that significance tests for discrete data operate at exactly the nominal level. He also misrepresents the NTP's decision making process, overstates the number of statistical comparisons made, and ignores that fact that the mouse liver tumor effects were so striking (e.g., $p < 0.0000000000001$) that it is virtually impossible that they could be false positive outcomes. Gaus' conclusion that such obvious responses merely “generate a hypothesis” rather than demonstrate a real carcinogenic effect has no scientific credibility. Moreover, his claims regarding the high frequency of false positive outcomes in carcinogenicity studies are misleading because of his methodological misconceptions and errors.

Keywords

Statistical analysis; NTP carcinogenicity; *Ginkgo biloba*; False positive

*Corresponding author. Grace Kissling, Ph.D., NIEHS, P.O. Box 12233, Research Triangle Park, NC 27709 [kissling@niehs.nih.gov.] Phone: 919-541-1756.

Dr. Zeiger has no conflict of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Gaus (2014) investigates “statistical considerations” in the National Toxicology Program’s (NTP’s) carcinogenicity testing program, using the *Ginkgo biloba* Technical Report (TR) 578 as an example (NTP 2013). In doing so, he demonstrates a basic misunderstanding of the data being analyzed, the statistical tests being used to analyze the data, and the decision making process used by the NTP when evaluating the carcinogenic potential of the substances being tested. His thesis is that the NTP carries out so many statistical tests, that huge numbers of false positive outcomes are unavoidable, and thus the studies are basically uninterpretable with regard to demonstrating carcinogenic effects.

Gaus lists in his Table 1 six “studies” reported by the NTP in TR 578, and he alleges that in each one “tissues from more than 40 sites were examined for every animal.” He speculates (without evidence) that approximately 2-3 “findings” were possible for each of these 40 sites. Based on his Table 1 data, he assumes that $40 \times 2.5 \times 48$ or 4800 statistical tests were carried out by the NTP. Then, using this value, he calculates that we would expect to see 4800×0.05 or 240 significant ($p < 0.05$) pairwise effects by chance alone, compared to the 209 significant effects actually observed that he calculates somehow from the text in TR 578 (a calculation that we were unable to confirm). This calculation is based on the assumptions that these presumed 4800 statistical tests (which are not independent and are exaggerated in number) operate at exactly the $p < 0.05$ level (which they do not), that no adjustment was made for multiple comparisons by the NTP (which is incorrect), and that more than 40 sites were evaluated in each of the six studies (which is also incorrect). He then questions the NTP’s conclusion that *Ginkgo biloba* extract was carcinogenic, and relegates that conclusion to the status of an unconfirmed hypothesis.

Before statistically analyzing data, it is important to have an understanding of the data being analyzed and the basic study design. This apparent lack of understanding has led Gaus to a number of incorrect assumptions which are exemplified by his Table 1. The first of these assumptions is that all six studies were equivalent with regard to the variables evaluated. However, a simple reading of TR 578 will show that this assumption is false.

For example, Gaus incorrectly states that in each of the six studies “tissues from more than 40 sites were examined for every animal.” One of these studies (Study 5) was a “special study” that examined only the liver and thyroid glands (see Table 1, page 31 of NTP TR 578). The “23 day” study in rats (Study 2) was a “clinical pathology only” portion of the rat 3 month study and was not part of the “core [3 month] study” that received the complete “more than 40 sites” evaluation (see Table 1 of NTP TR 578). Thus, in these two studies, he assumes that multiple statistical tests were carried out for many organs and parameters that were never a part of the study evaluation.

Gaus also erroneously claims that the NTP used no multiple comparison tests. He states that “the multiple use of the control group requires a specific methodology, e.g., the modification of the t-test by Dunnett,” and that in TR 578 “no adjustment for multiple comparisons is mentioned”. The use of these tests is clearly described in the Statistical Methods section of the NTP report. TR 578 did in fact use Dunnett’s test (and William’s modification of

Dunnett's test) for multiple comparisons for the normally distributed continuous variables in the 3-month studies (as the NTP does for all toxicity and carcinogenicity studies). Shirley's or Dunn's multiple comparisons tests for the non-normally-distributed continuous variables were also used. The specific multiple comparisons tests used in the *Ginkgo biloba* study are identified in the Statistical Methods section of the TR as well as in footnotes to the data tables.

Appendices F, G, and H of NTP TR 578 report the results of statistical analyses of 260 variables from the 3-month studies. All of the statistical analyses controlled for multiple comparisons (contrary to Gaus' assertions), so we would expect to find 260×0.05 or 13 variables showing some statistical significance by chance alone. In fact, there were 59 statistically significant differences. Thus, even for the 3-month studies there was strong evidence that *Ginkgo biloba* was producing real biological effects.

The assumption made by Gaus in his calculations in the 3-month studies that each of the five doses contributes 0.05 to the overall false positive rate for a given variable fails to recognize that the tests used by the NTP have already been adjusted for multiple comparisons to assure an overall false positive rate of 0.05 for that variable. Thus, the contribution to the false positive rate should be 0.05 for each of these variables, not 0.25 (0.05×5) as assumed by Gaus. This, alone, makes a huge 5-fold difference in his inflated false positive rate calculations.

2. NTP rodent carcinogenicity studies

The most important studies in the *Ginkgo biloba* technical report, and the primary reason for which the report was prepared, were the rat and mouse two-year studies identified by Gaus as Studies 4 and 6. These are the only studies in which carcinogenic effects were assessed, and the results are summarized in Appendices A2, B2, C2, and D2 of NTP TR 578. Gaus asserts that 40 sites were examined for tumors in these studies, with an average of 2.5 "findings" per site. Thus, with three doses and a $p < 0.05$ level test, he calculates the number of "significant pairwise differences" that are expected to be produced by chance is $40 \times 2.5 \times 3 \times 2 \times 2 \times 0.05$ or 60; this number is greatly inflated. False positive rates of rodent carcinogenicity bioassays have been examined numerous times by a number of different authors (Fears, Tarone et al. 1977; Haseman 1983; Office of Science and Technology Policy 1985; Haseman 1990; Haseman and Elwell 1996; Lin and Rahman 1998; Rahman and Lin 2008). Gaus appears to be unaware of these publications, which have consistently shown that the overall false positive rate associated with the decision making process used to interpret rodent tumor data is reasonably close to 0.05.

Importantly, for many of the more than 40 sites in the two-year studies, no tumors were seen in any animal in any dosed sex-species group, so tests for statistical significance would be meaningless. These sites include esophagus, cecum, colon, rectum, ileum, glandular stomach, blood vessel, heart, epididymis, seminal vesicle, spinal cord, pleura, trachea, eye, gall bladder, and tooth. Other sites (e.g., salivary gland, parathyroid gland, bone marrow) showed tumors in so few animals that it is impossible to achieve statistical significance, even if statistical tests were carried out in such cases.

Gaus apparently believes that tumor-less sites, and other sites for which statistical significance is impossible to achieve, nevertheless contribute to the overall false positive rate. He argues that “the fact that tests without a chance of being significant were not computed to save superfluous workload still does not reduce the number of significant tests.” However, a statistical test comparing tumor incidence at a site with no tumors is meaningless and does not contribute to the overall false positive rate, as Gaus assumes.

Clearly, to estimate the true underlying false positive rate associated with carcinogenicity in the *Ginkgo biloba* study, one must base the calculations on those tumors that occur in sufficient numbers to make a false positive outcome theoretically possible. To evaluate this matter, we examined the NTP historical control tumor database used in the *Ginkgo biloba* study (<http://ntp.niehs.nih.gov/go/34284>). In this database there were a total of 40 tumor sites/types in male and female rats and mice that occurred with a sufficient frequency (i.e., >2% incidence) to contribute meaningfully to an overall false positive rate. These historical control tumor rates can be used to assess the number of significant ($p<0.05$ or $p<0.01$) pairwise dosed group vs. control differences in cancer rates expected to occur by chance alone.

When performing these calculations, it is important to understand that statistical tests based on discrete data cannot achieve exactly a $p<0.05$ level of statistical significance, especially for a low incidence tumor. For example, suppose a tumor has a background rate of 2%, 50 dosed animals and 50 controls are examined, and a $p<0.05$ Fisher’s exact test is used to compare observed tumor rates. In this example, it can be shown that the actual level of significance is only 0.0012 because of the discreteness of the data and the relatively low background tumor rate. Even for tumors with a higher background rate, the actual significance level can never reach the nominal level. Thus, using the nominal significance level of 0.05 for false positive rate estimation, as Gaus did, will greatly exaggerate the number of false positives expected.

Although the NTP uses the poly-3 test for pairwise comparisons, not Fisher’s exact test, the same argument holds. The point is that when discrete data are used in false positive rate calculations, it is inappropriate to use the “nominal” level of $p=0.05$; one must use the “actual” level of significance, and sometimes it is much lower than $p=0.05$.

Gaus also assumes that the NTP rigidly adheres to a $p<0.05$ statistical decision rule when interpreting experimental results. This is not the case. The ultimate decision on carcinogenicity is a complex “weight of evidence” scientific judgment process that considers, in addition to p values, factors such as the presence or absence of a dose-response, the concurrent control value, historical control data, the presence (or absence) of pre-neoplastic lesions, the consistency of the effect across sexes, and whether or not the effect was seen in a suspected target organ. Haseman (1983) estimated that the overall false positive rate associated with the actual NTP decision-making process is about 7%, and could be closely approximated by a statistical decision rule that, in the case of a rare tumor, required $p<0.05$ for a dosed group vs. control comparison, and $p<0.01$ for a common tumor dosed group vs. control comparison. Other, more recent, papers have reported similar results (Haseman and Elwell 1996; Lin and Rahman 1998; Rahman and Lin 2008).

In the *Ginkgo biloba* study, there were several reported $p < 0.05$ pairwise effects (e.g., leukemia in male rats, and uterine polyps and mammary gland fibroadenoma in female rats) that were not considered to be “real” carcinogenic effects by the NTP for a variety of reasons (NTP 2013). In contrast, the effects that were considered to provide “clear evidence of carcinogenic activity” in mice were quite dramatic, as is discussed in more detail below.

As we have noted, in order for a chemical to be called carcinogenic in a rodent cancer bioassay, its response(s) must not only be statistically significant, but also be coherent, and part of a pattern of responses. For example, if a male mouse liver tumor response is statistically significant at the low dose, but not at the two higher doses, it would not be considered a positive tumor response. However, if the statistically significant response is at the highest dose, and the overall pattern of tumor response shows a monotonically increasing trend, it will be considered a positive tumor response. The *Ginkgo biloba* report provides an excellent example of this latter pattern (Table 1). Not only are the incidences of the different liver tumor types increased in the male and female mice, but the responses also exhibited significant, monotonically increasing trends.

Moreover, Gaus’ calculations ignore that fact that the carcinogenic effects that provided clear evidence of carcinogenicity in mice were so dramatic (see Table 1) that it is virtually impossible for them to be false positive outcomes. It is scientifically (and statistically) irresponsible to argue that there is no “statistical proof” that these are real carcinogenic effects, and that these dramatic results “only generates a hypothesis which has to be proved in a new study.”

It is curious that Gaus does not present, or consider, the dramatic, and consistent, liver tumor incidences in response to *Ginkgo biloba* extract in his article. For example, by a Fisher’s exact test, the high dose hepatoblastoma effect in male mice is highly significant ($p = 0.000000000000001$), that is, a probability of 1 in one quadrillion. Even using Gaus’ assumed “total statistical tests” value of 4800, the expected number of findings this significant by chance alone is essentially zero (i.e., 0.0000000000004800). Similar results hold for the other liver tumor effects in male and female mice (see Table 1). These are obviously real carcinogenic effects and not potential false positive outcomes that require confirmation by additional testing, as Gaus recommends.

We next investigate this important issue: How many significant pairwise carcinogenic effects would we expect to see by chance alone in a negative rodent carcinogenicity study if we use (1) Gaus’s approach of making all pairwise comparisons at the $p < 0.05$ level, compared with (2) the complex NTP decision making process that is approximated by the procedure suggested by Haseman (1983)?

Using the tumor rates in the NTP historical control database cited above (and assuming a Fisher’s exact test based on 50 animals per group) we calculated the false positive rate associated with each of the 40 individual tumor sites/types that could contribute meaningfully to an overall false positive rate, and then combined this information across sex-species groups to obtain the number of false positives expected by chance alone (details available upon request). We found that in a negative rodent carcinogenicity study we would

expect to see on average 2.1 “significant” ($p < 0.05$) increases in tumor incidence, but only 0.3 significant effects using the NTP decision making process.

These estimates agree closely with empirical NTP data. We consider three randomly selected, recent NTP negative mouse and rat carcinogenicity studies (NTP TR’s 566, Diethylamine (NTP 2011a); 567, Ginseng (NTP 2011b); and 572, Methyl trans-Styryl Ketone (NTP 2012)). All three studies reported no evidence of carcinogenic activity in male or female mice and rats. From these technical reports, we can calculate both the frequency of $p < 0.05$ pairwise carcinogenic “effects”, as well as the frequency of significant carcinogenic effects using the statistical decision rule that closely approximates the NTP decision making process (Haseman 1983). The results are summarized in Table 2. These data are totally consistent with the above calculations, and are far less than the 60 “significant” results predicted by Gaus’ approach.

3. Conclusions

Gaus’ conclusion that the reported carcinogenic outcomes in the *Ginkgo biloba* study are not a “proof [of carcinogenicity],” but a hypothesis needing confirmation is based on faulty statistical methodology that greatly exaggerates the actual likelihood of a false positive outcome, and a lack of understanding of the experimental designs and the NTP’s analyses of the tests. The carcinogenic effects seen in this study (especially the dramatic liver tumor effects seen in male and female mice) are strongly supported by the data and clearly represent real carcinogenic effects. It defies logic to conclude that these compelling data merely “generate a hypothesis” that *Ginkgo biloba* extract might cause liver tumors, and that such a hypothesis requires additional confirmation before these tumor increases can be accepted as real biological effects. Moreover, the general claims made by Gaus regarding the high frequency of false positive outcomes in rodent carcinogenicity studies is a result of numerous methodological errors, is in opposition to numerous published estimates of false positive rates, and is not borne out by the data and conclusions in the many NTP cancer bioassay reports.

Acknowledgments

This research was supported in part (G.E.K.) by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences, under projects Z01 ES45003 and Z01 ES45004. Dr. Kissling, chief statistician for the NTP, is currently responsible for the statistical analyses in the NTP Technical Reports, and was a contributor to the analysis addressed by Gaus (2014). Dr. Haseman is the former chief statistician for the NTP and was responsible for the development of the statistical approaches used in the Technical Reports and for the analyses used for numerous published reports.

References

- Fears TR, Tarone RE, et al. False-positive and false-negative rates for carcinogenicity screens. *Cancer Res.* 1977; 37(7 Pt 1):1941–1945. [PubMed: 861930]
- Gaus W. Which level of evidence does the US National Toxicology Program provide? Statistical considerations using the Technical Report 578 on *Ginkgo biloba* as an example. *Toxicology Letters.* 2014; 229(2):402–404. [PubMed: 24998428]
- Haseman JK. A reexamination of false-positive rates for carcinogenesis studies. *Fundam Appl Toxicol.* 1983; 3(4):334–339. [PubMed: 6628896]

- Haseman JK. Use of statistical decision rules for evaluating laboratory animal carcinogenicity studies. *Fundam Appl Toxicol.* 1990; 14(4):637–648. [PubMed: 2193843]
- Haseman JK, Elwell MR. Evaluation of false positive and false negative outcomes in NTP long-term rodent carcinogenicity studies. *Risk Anal.* 1996; 16(6):813–820. [PubMed: 8972110]
- Lin KK, Rahman MA. Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. *J Biopharm Stat.* 1998; 8(1):1–15. discussion 17–22. [PubMed: 9547425]
- NTP. Toxicology and carcinogenesis studies of diethylamine (CAS No. 109-89-7) in F344/N rats and B6C3F1 mice (inhalation studies). *Natl Toxicol Program Tech Rep Ser.* 2011a; (566):1–174.
- NTP. Toxicology and carcinogenesis studies of ginseng (CAS No. 50647-08-0) in F344/N rats and B6C3F1 mice (gavage studies). *Natl Toxicol Program Tech Rep Ser.* 2011b; (567):1–149.
- NTP. Toxicology and carcinogenesis studies of methyl trans-styryl ketone (CAS NO 1896-62-4) in F344/N rats and B6C3F1 mice (feed and dermal studies). *Natl Toxicol Program Tech Rep Ser.* 2012; (572):1–188.
- NTP. Toxicology and carcinogenesis studies of Ginkgo biloba extract (CAS No. 90045-36-6) in F344/N rats and B6C3F1/N mice (Gavage studies). *Natl Toxicol Program Tech Rep Ser.* 2013; (578):1–183.
- Office of Science and Technology Policy. Chemical carcinogens: A review of the science and its associated principles. *Federal Register.* 1985; 10:10371–10442.
- Rahman MA, Lin KK. 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.* 2008; 18(5):949–958. [PubMed: 18781527]

Abbreviations

NTP	National Toxicology Program
TR	Technical Report
p	p-value

Highlights

- Gaus (2014) demonstrates a serious misunderstanding of the NTP's statistical analysis and interpretation of rodent carcinogenicity data, thereby greatly exaggerating the anticipated false positive rate.
- False positive rates of rodent carcinogenicity bioassays have been examined numerous times and have been established to be reasonably close to 0.05.
- The dramatic liver effects in the *Ginkgo biloba* mouse study are obviously real carcinogenic effects and not false positive outcomes.
- Gaus' conclusion that NTP's positive rodent carcinogenicity studies merely "generate a hypothesis" rather than demonstrate carcinogenic effects is not credible and is not supported by the data.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Mouse liver tumor effects in the NTP *Ginkgo biloba* study

	Males				Females			
	Control	Low dose	Mid dose	High dose	Control	Low dose	Mid dose	High dose
Hepatocellular adenoma	31/50	46/50	33/50	33/50	17/50	37/50	41/50	48/50
Hepatocellular carcinoma	22/50	31/50	41/50	47/50	9/50	10/50	15/50	44/50
Hepatoblastoma	3/50	28/50	36/50	38/50	1/50	1/50	8/50	11/50

Table 2

Frequency of significant pairwise carcinogenic effects in three randomly selected, recent negative NTP rodent carcinogenicity studies

NTP TR	Frequency of $p < 0.05$ pairwise differences	Frequency of carcinogenic effects using approximate NTP decision rule
566	1 ^a	0
567	2 ^b	1
572	2 ^c	0

^a Male rats, low dose, skin tumors, $p = 0.044$

^b Male mice, all benign tumors, high dose, $p = 0.050$; female mice, all malignant tumors, low dose, $p = 0.009$

^c Female rats, uterine polyps, mid dose, $p = 0.040$; female mice, all benign tumors, mid dose, $p = 0.020$