Toxicol Pathol OnlineFirst, published on July 2, 2009 as doi:10.1177/0192623309339606

Regulatory Forum

Toxicologic Pathology, 000: 1-5, 2009

ISSN: 0192-6233 print / 1533-1601 online
DOI: 10.1177/0192623309339606

Points to Consider on the Statistical Analysis of Rodent Cancer Bioassay Data When Incorporating Historical Control Data

Susan A. Elmore[1,2] and Shyamal D. Peddada[3]

[1]*National Toxicology Program, Research Triangle Park, North Carolina 27709, USA*
[2]*Cellular and Molecular Pathology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709, USA,*
[3]*Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709, USA*

Abstract

Researchers routinely use historical control data (HCD) when analyzing rodent carcinogenicity data obtained in a particular study. Although the concurrent control group is considered to be the most relevant group to compare with the dose groups, the HCD provides a broader perspective to assist in understanding the significance of the current study. The HCD is used to provide information about the incidences of spontaneous tumors and malignant systemic disorders such as lymphoma and leukemia. This article presents some possible ways of incorporating the HCD when analyzing data from a rodent cancer bioassay. Specifically, exploratory (informal) and formal statistical procedures for analyzing such data are reviewed. The boxplot is presented as an exploratory tool that describes the current data in the context of the distribution of the HCD. It will also identify potential outliers that would not be otherwise be flagged using standard methods such as the mean, standard deviation, and range. The various options for the statistical analysis of HCD presented here do not necessarily represent standard practice.

Keywords: boxplot; IQR; lower quartile; median; range; upper quartile; historical control data.

Introduction

Two-year rodent toxicology and carcinogenesis bioassays are conducted by government agencies, private companies, and research institutes to identify toxic and carcinogenic compounds that are potentially hazardous to human health through exposure to pharmaceuticals, nutraceuticals, food, water, or other environmental sources. When analyzing the tumor incidences in treatment (or dose) groups, the most appropriate control for comparison is the concurrent control group. The evaluation of the tumor incidences in the treatment groups relative to the concurrent control group is traditionally based on established statistical methods such as the Poly-3 trend test (Bailer and Portier 1988; Bieler and Williams 1993). This test, which adjusts for survival, allows one to determine the statistical significance of the tumor incidence within a treatment group and also helps to determine if there is a statistically significant trend across dose groups within a study.

To assess if the tumor responses in the current study are unusual in comparison to what is known historically about the lesion among control animals, it is customary for researchers to compare the responses in the current study with the tumor incidences in control groups from previous studies. "Historical control data" (HCD) is the term used for this compilation of data from previous studies. Thus, the HCD can be used to determine if the tumor incidence in the concurrent control group or dual control groups is consistent with the tumor incidence in the historical control groups. Comparison of the tumor incidence rates in treated groups with both concurrent control groups and HCD can, along with other study data such as the incidence of other lesions of similar cell lineage, help to determine biological relevance.

HCD is helpful in interpreting the tumor incidences in a variety of situations, such as rare tumors, common tumors (e.g., pituitary pars distalis adenomas in male and female rats), tumors with highly variable incidence rates (e.g., pancreatic islet cell tumors in male rats or thyroid C-cell adenomas in male and female rats), a tumor that has a marginal increase in incidence relative to concurrent controls, or unexpected increases or decreases of tumor incidence in study control

This is an opinion paper submitted to the Regulatory Forum and does not constitute an official position of the Society of Toxicologic Pathology or *Toxicologic Pathology*. All opinions, positions, or ideas expressed within this article are entirely those of the authors, who take full responsibility for them.

Address correspondence to: Susan A. Elmore, National Toxicology Program, National Institute of Environmental Health Sciences, Cellular and Molecular Pathology Branch, 111 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA; phone: 919-541-3474; fax: 919-541-7666; e-mail: elmore@niehs.nih.gov.

Abbreviations: CC, concurrent control; F344, Fischer 344; FDA, Food and Drug Administration; HCD, historical control data; HD, high dose; IQR, interquartile range; LD, low dose; MD, medium dose; NTP, National Toxicology Program; Q_1, lower quartile; Q_2, median; Q_3, upper quartile.

TABLE 1.—Effect of a single potential outlier on the nonrobust measures.

	Tumor incidence (percentage)	Nonrobust measures	Robust measures
Hepatoblastoma	6, 2, 2, 0, 2, 6, 0, 8, ***34***, 8, 0, 0, 4, 0, 2, 0, 0, 0, 4, 0, 4, 2, 2, 2, 4, 0, 2, 0, 2	3.21 (mean) 6.47 (*SD*) 0–34 (range)	2 (median) 0 (Q_1) 4 (Q_3) 0–4 (IQR)
Largest incidence in row 2 increased from 34 to 70	6, 2, 2, 0, 2, 6, 0, 8, ***70***, 8, 0, 0, 4, 0, 2, 0, 0, 0, 4, 0, 4, 2, 2, 2, 4, 0, 2, 0, 2	4.5 (mean) 13 (*SD*) 0–70 (range)	2 (median) 0 (Q_1) 4 (Q_3) 0–4 (IQR)

Tumor incidences in row 2 are the data reported in the NTP HC database for male B6C3F1 mice all routes and vehicles (November 2008). Tumor incidences in row 3 are the same as those in row 2, except that the 34% (in row 2) has been artificially replaced by 70% (in row 3) to illustrate the effect that an outlier may have on the mean, standard deviation (*SD*), and range.

animals (Baldrick 2005; Eiben and Bomhard 1999; Haseman, Arnold, and Eustis 1990; http://ntp.niehs.nih.gov/ntpweb; http://www.criver.com/sitecollectiondocuments/rm_rm_r_survival_wistar_han_rats_compilation_data.pdf).

Comparison of the tumor incidence data from the current study with HCD can be performed in two different ways. One may use an exploratory (informal) analysis of the data or take a more formal statistical approach to the analysis. Although these procedures provide a statistical evaluation of the data, one should weigh this information along with other biological/toxicological information when making a final assessment regarding a chemical.

EXPLORATORY (INFORMAL) STATISTICAL ANALYSIS OF TUMOR INCIDENCE

For a given lesion, a common informal method of using HCD is to provide the mean, standard deviation, and range of tumor incidence from a historical control database or published literature. Usually, for a given species, strain, sex, and vehicle, two different sets of means, standard deviations, and ranges are provided: one set that is specific to the route of exposure and another set that combines all routes. In some situations, statistical inferences based on the range of the distribution alone may provide misleading results. As the number of studies increases, the range of the distribution increases so that the range of historical control rates may be too high to be useful. Also, summary statistics such as the mean, the standard deviation, and the range can be affected by one anomalous study in the historical control database.

For example, consider the incidence of hepatoblastoma in male B6C3F1 mice. In the NTP's database for the twenty-nine studies conducted during the period September 13, 1999, to February 10, 2004 (based on NTP-2000 diet), there were 48 male mice out of 1,447 diagnosed with hepatoblastoma (Table 1) (http://ntp.niehs.nih.gov/ntpweb). These studies had a mean of 3.31%, a standard deviation of 6.47%, and a range of 0% to 34% for all routes and vehicles. However, the 34% incidence (17 out of 50 mice) was found in only one oral study with water as the vehicle. The next largest incidence was 8% (4 out of 50 mice). Without this "unusually" large incidence, the range would have been 0% to 8%. Thus, the range quadrupled as a consequence of one study. Other examples include hepatocellular and adrenal cortex adenomas in female F344 rats. In the NTP's database for the twenty-seven studies (all routes and vehicles) conducted during the period August 16, 1999, to January 22, 2004 (based on NTP-2000 diet), there were 16 out of 1,350 female rats diagnosed with hepatocellular adenoma (mean 1.19%, standard deviation 2.62%, range 0% to 12%) and 24 female rats out of 1,346 diagnosed with adrenal cortex adenoma (mean 1.78%, standard deviation 3.39%, range 0% to 16%) (http://ntp.niehs.nih.gov/ntpweb). The 12% incidence (6 out of 50) of hepatocellular adenoma was found in one skin study with ethanol as the vehicle. The range without this outlier is 0% to 4%. Thus, the range tripled as a consequence of this one study. The 16% incidence (8 out of 50 rats) of adrenal cortex adenoma was found in only one inhalation study (vehicle air). The next largest incidence was 6%. Without this one outlier the range would have been 0% to 6%. Thus, the range more than doubled as a consequence of one study.

These examples emphasize that range uses only the two endpoints of the entire distribution of the HCD and the intervening data are not considered. Consequently, comparison with the historical control range may result in overlooking a potential effect because the current study tumor incidences are "within" the historical range. Therefore, when comparing the current study data with the historical control range, the potential effect that outliers may have on the range should be considered along with all other relevant biological and toxicological data. When outliers are identified, they are not to be discounted but considered along with other relevant data to determine the significance of any potential treatment-related effect in dosed groups.

Since the range, mean, and standard deviation can be influenced by a single study, if there are a sufficient number of studies available, one could also report the median and interquartile range (IQR) of the HCD, which are not influenced by extreme data points (Figure 1). The median is the midpoint of all values sorted from smallest to largest. The range is the difference between the minimum and maximum values. The IQR is a measure of statistical dispersion and is equal to the difference between the upper and lower quartiles (75th and 25th percentiles), which are usually denoted as Q_3 and Q_1, respectively (Dawson and Trapp 2004) (Figure 1). The lower quartile is the median of the first half of the data sorted from smallest to

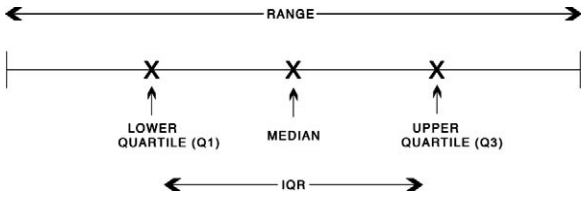

FIGURE 1.—Schematic drawing of the interquartile range (IQR). The median is the midpoint of all HCD values sorted from smallest to largest, and the range is the difference between the minimum and maximum HCD values. The IQR is the difference between the upper and lower quartiles (75th and 25th percentiles). The lower quartile (Q1) is the median of the first half of the data sorted from smallest to largest. The upper quartile (Q3) is the median of the second half of the data sorted from smallest to largest.

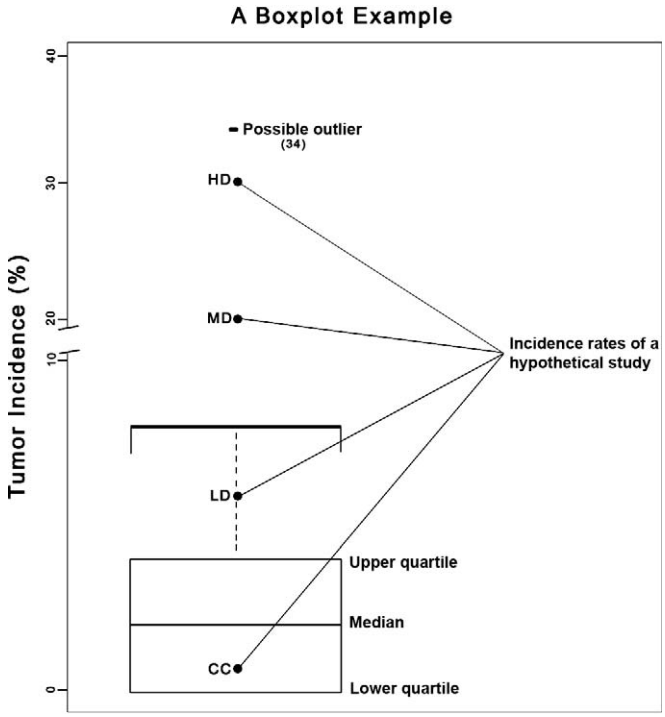

FIGURE 2.—The historical control data from Table 1 was used to construct this boxplot. A boxplot can be used to graphically summarize the distribution of study data and concurrent control with regard to historical control data. The IQR is the height of the box with the bottom of the box representing the lower quartile (Q1) of the HCD and the top of the box representing the upper quartile (Q3) of the HCD. The median of the HCD is a horizontal line inside the box. The "whisker" is a hatched line that extends from the top of the box to the largest value in the HCD data, not exceeding Q3 + 1.5 IQR. In this example, the potential HCD outlier of 34% is indicated. Superimposed on the boxplot are hypothetical examples of concurrent control (CC) and study data consisting of low-dose (LD), mid-dose (MD), and high-dose (HD) groups. The presence of a possible outlier (34%) results in data from the MD and HD groups falling within the HCD range. Without this outlier, these data would be outside of the upper range of the HCD.

largest. The upper quartile is the median of the second half of the data sorted from smallest to largest. In the above hepatoblastoma example, the 34% rate does not contribute to the IQR, but it does contribute to the range. The IQR for this example would be 6% (0% to 6%) compared to the range, which is 34% (0% to 34%). In this example if the maximum data point were anywhere beyond 8%, the IQR and median would not change, but the mean, standard deviation, and range would increase, illustrating the robustness of the IQR and median.

The IQR can be used to build a boxplot, which is a simple graphical way to summarize the distribution of the HCD and is most informative when there are fifteen or more studies to assess (Benjamini 1988; Dawson and Trapp 2004). The boxplot depicts groups of numerical historical control data through their five-number summaries: the smallest observation, lower quartile (Q_1), median (Q_2), upper quartile (Q_3), and largest observation (Figure 2). It consists of a vertical box, capturing the middle 50% of the data, with the bottom of the box representing Q_1 (lower quartile) and the top of the box representing Q_3 (upper quartile). The median of the distribution is the horizontal line inside the box. The IQR is the height of the box. From the top of the box a vertical line segment is drawn (known as a "whisker"). This line extends to the largest value in the data not exceeding $Q_3 + 1.5$ IQR. All data points (34% in this example) beyond this value are regarded as unusually large observations (potential outliers). Similarly, a whisker below the lower end

of the box may also be drawn. This line segment extends to the smallest value in the data not less than $Q_1 - 1.5$ IQR. Thus, all observations that go below this point are regarded as unusually small observations (potential outliers). Note, as in Figure 2, that if no data exist within the interval ($Q_1 - 1.5$ IQR, Q_1) then there will be no whisker at the lower end of the distribution. Similarly, if no data exist within the interval (Q_3, $Q_3 + 1.5$ IQR), then there will be no whisker at the higher end of the distribution.

On a boxplot of the HCD, one may mark the tumor rates in the concurrent control (CC), low-dose (LD), medium-dose (MD), and high-dose (HD) groups (Figure 2). Such a plot would clearly display the current experimental data in the context of the distribution of historical controls. For instance, if HD falls outside the upper whisker of the boxplot, while CC is within the box or whiskers of the distribution, then one may conclude that there is a possible treatment effect observed in the current data. The evidence of a potential treatment effect is stronger if the CC falls within the box. Due to potential differences in survival rates between dose groups and control groups, we recommend that the boxplot be constructed using the Poly-3 survival adjusted tumor incidence rates (Bailer and Portier 1988) rather than the raw incidence rates.

The data from Table 1 demonstrate that commonly used measures such as the mean, standard deviation, and range can be highly affected by extreme data, whereas the median and IQR do not change. In this example, for illustration purposes, if the highest data point of 34 were increased to 70, then the range and standard deviation would almost double, and the mean would increase by 40% (3.31 to 4.5). However, the robust measures such as Q_1, median, Q_3, and IQR would be unaffected.

Formal Statistical Analysis of Tumor Incidence

Another approach to analyzing the concurrent data using information from historical controls would be to apply formal statistical methods. Over the past two decades, several attempts have been made by statisticians to develop a statistical procedure for analyzing concurrent experimental data by formally making use of the HCD (Tarone 1982; Dempster, Selwyn, and Weeks 1983; Hoel 1983; Hoel and Yanagawa 1986; Tamura and Young 1986, 1987; Prentice et al. 1992; Ibrahim and Ryan 1996; Ibrahim, Ryan, and Chen 1998; Dunson and Dinse 2001). Each of these methods has strengths and limitations. For instance, Tarone (1982) treats tumor incidence as a binomial proportion and accounts for the extra binomial variation among historical studies using a probability distribution called the beta distribution. While this model seems reasonable and intuitive, the method does not take into consideration that not all animals survive to the end of the study. The statistical methodology of Ibrahim and Ryan (1996) assumes that tumors are lethal and cause instantaneous death, while Ibrahim, Ryan, and Chen (1998) assume that the tumors are nonlethal. Both of these assumptions are extreme and may not be true in practice. Dunson and Dinse (2001) overcame the above deficiencies by using a Bayesian methodology that does not make any of the above assumptions regarding the tumor. However, their method requires carefully chosen values for some of the statistical parameters in the model, termed *prior parameters*. From a practical point of view, it may be difficult to choose values for such prior parameters of the statistical model as it requires the toxicologist and pathologist to have a sound understanding of the underlying statistical model and the impact of the prior parameters on the data. Similarly, the statistician would require an understanding of the underlying biological/toxicological mechanisms when choosing the prior parameters.

Recently, Peddada, Dinse, and Kissling (2007) have proposed a nonparametric statistical method, which overcomes the above deficiencies. This methodology can be modified to compare the dose group with concurrent control and historical controls separately, thus resulting in a pair of *p*-values rather than one single *p*-value. It can also be modified to compare concurrent control with historical controls, resulting in a third *p*-value. From a ''weight of evidence'' point of view, the three *p*-values may be useful in understanding the significance of the current data. No distributional assumptions are made by this methodology regarding tumor incidences or tumor lethality. Similar to the Poly-3 trend test (Bailer and Portier 1988; Portier and Bailer 1989), it uses the Poly-3 correction to the sample size to account for differences in survival rates among dose groups. Such survival adjustments cannot be made without having survival times for individual animals. These data are usually not publicly available for the historical controls, and it would be useful to report this information as a part of the HCD. If survival adjustments are not made, then there is a potential for bias due to survival differences.

Summary

While the concurrent control group provides the most relevant control data for determining treatment-related effects in a study, evaluation of HCD may be useful in certain situations. These include the interpretation of rare tumors, high-incidence tumors, tumors with a highly variable incidence, tumors with a marginal increase in incidence relative to concurrent controls, or unexpected increases or decreases of tumor incidences in study control animals. All of the statistical approaches described here (exploratory and formal) may be used in combination to evaluate the HCD and to determine its appropriateness for comparison to a set of test data. However, HCD should be used as one of many sources of information that can add to the ''weight of evidence'' approach when assessing the potential carcinogenic effect of a compound. Other data to consider may be the incidences of other lesions of similar cell lineage, body weight, survival, time of tumor onset, if the tumor occurs in both species or both sexes, if there is a positive dose-related response, or if there are bilateral lesions in paired organs. The goal of using HCD is to gain additional information that might aid in the overall evaluation of a carcinogenicity study. The various statistical tools that are available to evaluate HCD should be considered and discussed in the context of

sound biological principles. For further comments on statistical approaches, one may refer to the U.S. Food and Drug Administration Center for Drug Evaluation and Research (FDA CDER; 2001) guidance for industry document.

Acknowledgments

The authors wish to thank Ms. Elizabeth Ney of the National Institute of Environmental Health Sciences (NIEHS) for preparation of the IQR and boxplot figures and members of the Society of Toxicologic Pathology Historical Control Data Working Group for helpful discussion and review. This research was supported (in part) by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04). We also wish to thank Drs. Gregg Dinse (NIEHS), Grace Kissling (NIEHS), Walter Piegorsch (U. Arizona) and John Peckham (EPL) for their helpful comments which improved the presentation of this manuscript.

References

Bailer, A. J., and Portier, C. J. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics* **44**, 417–31.

Baldrick, P. (2005). Carcinogenicity evaluation: Comparison of tumor data from dual control groups in the Sprague-Dawley rat. *Toxicol Pathol* **33**, 283–91.

Benjamini, Y. (1988). Opening the box of a boxplot. *The American Statistician* **42**, 257–62.

Bieler, G. S., and Williams, R. L. (1993). Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. *Biometrics* **49**, 793–801.

Dawson, B., and Trapp, R. (2004). Basic and clinical biostatistics, 4th ed. McGraw-Hill Medical Publishers, New York.

Dempster, A. P., Selwyn, M. R., and Weeks, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J Amer Stat Assoc* **78**, 221–27.

Dunson, D. B., and Dinse, G. E. (2001). Bayesian incidence analysis of animal tumorigenicity data. *Appl Stat* **50**, 125–41.

Eiben, R., and Bomhard, E. M. (1999). Trends in mortality, body weights and tumor incidences of Wistar rats over 20 years. *Exp Toxicol Pathol* **51**, 523–36.

Haseman, J. K., Arnold, J., and Eustis, S. L. (1990). Tumor incidences in Fischer 344 rats: NTP historical data. In Pathology of the Fischer rat reference and atlas (G. A. Boorman, S. L. Eustis, M. R. Elwell, C. A. Montgomery, and W. F. Mackenzie eds.), pp. 555–64. Academic Press, San Diego, CA.

Hoel, D. G. (1983). Conditional two sample tests with historical controls. In Contributions to Statistics (P. K. Sen, ed.), pp. 229-36. North-Holland, Amsterdam, the Netherlands.

Hoel, D. G., and Yanagawa, T. (1986). Incorporating historical controls in testing for a trend in proportions. *J Amer Stat Assoc* **81**, 1095–99.

Ibrahim, J. G., and Ryan, L. M. (1996). Use of historical controls in time-adjusted trend tests for carcinogenicity. *Biometrics* **52**, 1478–85.

Ibrahim, J. G., Ryan, L. M., and Chen, M. (1998). Using historical controls to adjust for covariates in trend tests for binary data. *J Amer Stat Assoc* **93**, 1282–93.

Peddada, S., Dinse, G., and Kissling, G. (2007). Incorporating historical control data when comparing tumor incidence rates. *J Amer Stat Assoc* **102**, 1212–20.

Portier, C. J., and Bailer, A. J. (1989). Testing for increased carcinogenicity using survival–adjusted quantal response test. *Fundam Appl Toxicol* **12**, 731–37.

Prentice, R. L., Smythe, R. T., Krewski, D., and Mason, M. (1992). On the use of historical control data to estimate dose response trends in quantal bioassay. *Biometrics* **48,** 459–78.

Tamura, R. N., and Young, S. S. (1986). The incorporation of historical information in tests of proportions: Simulation study of Tarone's procedure. *Biometrics* **42**, 343–49.

Tamura, R. N., and Young, S. S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics* **43**, 813–24.

Tarone, R. E. (1982). The use of historical control information in testing for a trend in proportions. *Biometrics* **38**, 215–20.

U.S. 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 (section C). http://www.fda.gov/cder/guidance/815dft.htm