*Environmental Health Perspectives*
*Vol. 68, pp. 5-9, 1986*

# Methodology for Statistical Analysis of SENCAR Mouse Skin Assay Data

## by Judy A. Stober*

**Various response measures and statistical methods appropriate for the analysis of data collected in the SENCAR mouse skin assay are examined. The characteristics of the tumor response data do not readily lend themselves to the classical methods for hypothesis testing. The advantages and limitations of conventional methods of analysis and methods recommended in the literature are discussed. Several alternative response measures that were developed specifically to answer the problems inherent in the data collected in the SENCAR bioassay system are described. These measures take into account animal survival, tumor multiplicity, and tumor regression. Statistical methods for the analysis of these measures to test for a positive dose response and a dose–response relationship are discussed. Sample data from representative initiation/promotion studies are used to compare the response measures and methods of analysis.**

## Introduction

The SENCAR mouse skin initiation/promotion bioassay has been widely used to test for chemical carcinogens and tumor initiators and promoters. A general description of the SENCAR skin assay system and mouse skin tumorigenesis can be found in Slaga et al. (*1*) and Bull et al. (*2*). In general, the test substance is administered to the animals, and the appearance of skin tumors (papillomas or carcinomas) on individual animals is charted for a specified period of time. Compounds, dose levels of a given compound, and/or routes of exposure are then compared by evaluating the tumor incidence among the treatment groups of interest. A positive association is defined as a reproducible increase in the occurrence of tumors, and a negative association is defined as the absence of such an increase.

Statistics provide the necessary procedures for summarizing the relevant data and a mechanism for evaluating the strength of the association observed. As in any experimental situation, appropriate statistical methods are essential. The statistical analysis methods should be sensitive to the characteristics of the response data. The multiplicity of tumors, as well as their time of occurrence, can and should be used as indicators of carcinogenicity. Unfortunately, the characteristics of the SENCAR skin bioassay data do not lend themselves readily to the classical techniques for data analysis. In particular, early animal mortality and tumor regression complicate the analysis.

The objective of this paper is to define appropriate, sensitive, and useful statistical procedures for analyzing data obtained from a SENCAR mouse skin assay. Statistical methods commonly used or recommended in the literature will be reviewed. The advantages and disadvantages of each method will be discussed. Several alternative methods developed for the SENCAR tracking system at the Health Effects Research Laboratory of the U.S. Environmental Protection Agency (HERL/EPA) will be presented and evaluated using representative data sets.

Several general principles were used as guidelines in examining and evaluating statistical methods for the analysis of SENCAR skin tumor data. First, the questions investigated by the experiment should be addressed by the test procedure in a straight-forward manner with sufficient power to detect biologically meaningful differences. That is, the methodology selected should minimize the rates of false positives and false negatives. Second, the methods should be intuitively reasonable and sensitive to the characteristics of the data, using all relevant information available. Finally, the assumptions for the statistical validity of the method must be reasonably met, and the procedure should be robust, whether it is an omnibus or a directional test.

## Background

The SENCAR mouse skin assay can be designed to investigate two separate but related hypotheses. The first hypothesis asks: Is there evidence for an increase in tumor occurrence associated with the treatment (i.e., is there a positive dose response)? If there is, then the second hypothesis can be used to investigate evidence for and may even be used to estimate a dose–response relationship. The current work limits this investigation

*Toxicology and Microbiology Division, Health Effects Research Laboratory, U.S. Environmental Protection Agency, Cincinnati, OH 45268.

to a monotone response, thus the hypotheses can be stated as: Is there an increase in tumor occurrence associated with an increase in dose of a compound? Additionally, questions concerning the effect of exposure route and tumorigenic potency can be addressed using the same methods discussed below. The experimental endpoints on which a statistical analysis is based include animal body weight, food and water consumption, tumor counts, and pathology reports. Although each of these measurements supplies information on toxicity, this paper will focus only on tumor count data and how to use these data to calculate a reliable response measure for toxicity.

## Methods

Several studies have reported on the use of various statistical methods to analyze mouse skin assay data (*3-6*). The conventional approach to summarizing the assay results usually includes some or all of the following information: dose, number of animals exposed, and response measures such as percentage of animals with tumors, average (median) number of weeks to first tumor (±SD), time to first tumor observation, time to last tumor observation, and number of tumors per animal. The classical statistical methods (e.g., Student's *t*-test and contingency table analysis) are not necessarily appropriate for these response measures. The selection of a response measure and statistical methods for the analysis of tumor data is complicated by certain characteristics of the study data. Potential problems in the analysis arise from animal mortality occurring before the termination of the study, differences in survival among the treatment groups, and tumor regression. For example, analysis of tumor occurrence is complicated by incomplete observations in animals due to early mortality, which may be directly related to the toxicity of the compound. Tests that ignore these factors may result in a loss of power, that is, the ability to detect true differences. If no early deaths occur, the data analysis is decidedly easier. Ignoring early animal mortality and examining only surviving animals may lead to erroneous results.

The percentage of animals with skin tumors at a specific time has been traditionally analyzed by using either Fisher's Exact Test for pairwise comparisons between treatment groups or a chi-squared test. In addition, linear, probit, and logit models have been used to fit dose–response curves. Tumor counts observed at a specific time have been analyzed using various parametric and nonparametric procedures. These procedures include Student's *t*-test, analysis of variance procedures, linear regression, and log-linear models. Drinkwater and Klotz (*4*) examined the analysis of tumor multiplicity data for surviving animals using the Student's *t*-test, Wilcoxon's Rank Sum Test and a two-sample likelihood ratio test based on the negative binominal distribution. Both of these response measures, percentage of animals with skin tumors and tumor count, ignore differential survival and provide no information on the pattern of tumor development. The parametric methods require various assumptions to ensure their validity, and these assumptions may not hold true.

The pattern of tumor development can be examined by using the time to first tumor as the response measure. The time to first tumor is calculated as the time from initial exposure to the appearance of the tumor. Cumulative incidence curves can be estimated by using various parametric procedures, e.g., Weibull (*5*) and nonparametric methods, e.g., Kaplan-Meier (*3*). The cumulative incidence curves can be displayed graphically and the estimated median time to first tumor reported as a descriptive statistic. These methods include all animals at risk, adjusting for differences in mortality rates among the groups. They provide overall measures of the compound's effect on initial tumor development. The entire incidence curve, rather than a single point in time, can be statistically compared for differences among treatment groups. This method is limited by the fact that it ignores tumor multiplicity, and the median time to first tumor cannot always be calculated directly from the observed data. Additionally, the parametric methods require the assumption of a specified model distribution that may or may not be tested for "goodness of fit."

The Toxicology and Microbiology Division of HERL/EPA has developed and implemented a tracking system for collecting data in the mouse skin assay system. In addition to the daily observation of the animals for health monitoring and morbidity, the animals are observed weekly, and all grossly observed tumors are individually charted as to location, size, and type (papilloma or carcinoma). Only those tumors which are observed for a minimum of three consecutive weeks are included in the permanent and cumulative tumor count. All individual tumors are charted weekly until they regress or coalesce, or until the animal is sacrificed.

In developing the methodology to be used in the analysis of tumor data from the tracking system, the multiplicity of tumors and their time of occurrence were used as indicators of carcinogenicity. The objective was to define summary response measures and statistical methods that would simultaneously take into account survival rates and the total tumor occurrence pattern, including tumor multiplicity and regression. Crump and Ng (*7*) developed several approaches for testing for a dose-related effect and applied the methods to several tracking system data sets. Table 1 lists the response measures considered. The response is calculated over the entire study time. A response measure is calculated for each animal with its value censored at the time of the animal's death. Tumors that appear at the same location at nonoverlapping times are counted as single tumors.

The univariate response measures admittedly have inherent weaknesses. The time to first tumor is defined as previously discussed. The total number of permanent and cumulative tumors reflects tumor multiplicity, but does not take the time of occurrence and duration into

**Table 1. Tumor response measures.**

| Measure | Comments |
|---|---|
| Univariate | |
| Time to first tumor | Ignores multiplicity of tumors |
| Total number of permanent tumors | Ignores time of tumor occurrence, duration and tumor regression |
| Total number of cumulative tumors | Ignores time of tumor occurrence and duration |
| Integral with respect to (w.r.t.) time of the number of tumors | Takes duration into account but ignores exact time of occurrence |
| Weighted sum of tumors | Each tumor weighted by its time of appearance. Ignores tumor regression |
| Multivariate | |
| Multiple times to tumor | Takes multiple tumors per animal into consideration |

account. The integral with respect to time of the number of tumors is defined as follows.

Let $t_i < \ldots < t_k$ be the observation times. A tumor seen at time $t_i$, for $i = 2, \ldots, k-1$, contributes a factor of $(t_{i+1} - t_{i-1})/2$ to the integral. A tumor seen at time $t_1$ and $t_k$ contributes a factor of $(t_2 - t_1)$ and $(t_k - t_{k-1})$, respectively to the integral. That is, the integral is defined as

$$n_1(t_2 - t_1) + \sum_{i=2}^{k-1} n_i(t_{i+1} - t_{i-1})/2 + n_k(t_k - t_{k-1})$$

where $n_i$ is the number of tumors seen at time $t_i$, for $i = 1, \ldots, k$. The integral takes duration into account but not the specific time of occurrence. The weighted sum of tumors weights each tumor by the time from its appearance until the end of the experiment.

To test for a dose–response effect using the univariate response measures, both trend and multiple comparison "survival"-type analysis tests were examined by Crump and Ng. The multiple comparison test of Gehan and Wilcoxon (*8–10*) was examined. It is an omnibus test, in the sense that it has power to detect nearly all types of differences. The trend test considered was a Cox-type trend test discussed by Tarone (*11*). It is based on the proportional hazard model and is a one-sided directional test. For the univariate tests, it was assumed that the response measure, either the total number of tumors or the integral with respect to time of the number of tumors, was censored at the value attained at the time of death or study termination. The test statistics were calculated by using the PHGLM and SURVTEST procedures from the Statistical Analysis System (SAS) (*12*). The multivariate regression analysis method of Prentice et al. (*13*) was used to examine the multiple times to tumor data. It is a generalization of the Cox-type trend test applied to the univariate data. The multivariate test statistic was calculated using a modified version of the program R-COX developed at the Fred Hutchison Cancer Research Center.

**Table 2. Tumor multiplicity data from SENCAR mouse studies.[a]**

| No. of studies | No. of dose groups | No. of animals per group | Phorbol myristate acetate used as promoter? | Length of study, weeks |
|---|---|---|---|---|
| 8 | 2 | 40 | Yes | 31–36 |
| 10 | 3 | 40 | Yes | 52 |
| 6 | 4 | 40 | Yes | 52 |

[a] Only those tumors, both papillomas and carcinomas, observed for 3 consecutive weeks are included in tumor count.

**Table 3. Gehan-Wilcoxon multiple comparison test vs. Cox-type trend.**

| Response measure | Total no. of data sets | Percentage of data sets: GW ≤ Cox[d] | Percentage of data sets: GW > Cox[d] |
|---|---|---|---|
| Time to first tumor | | | |
| +/+[a] | 15 | 60 | 40 |
| −/−[b] | 8 | 63 | 37 |
| +/−[c] | 1 | 0 | 100 |
| Integral w.r.t. time (week = 24) | | | |
| +/+ | 7 | 100 | 0 |
| −/− | 8 | 100 | 0 |
| +/− | 5 | *100* | *0* |
| Integral w.r.t. time (week = last) | | | |
| +/+ | 8 | 75 | 25 |
| −/− | 9 | 89 | 11 |
| +/− | 7 | 100 | 0 |
| Total no. permanent tumors (week = 24) | | | |
| +/+ | 7 | 100 | 0 |
| −/− | 9 | 89 | 11 |
| +/− | 4 | 100 | 0 |
| Total no. permanent tumors (week = last) | | | |
| +/+ | 12 | 92 | 0 |
| −/− | 9 | 78 | 0 |
| +/− | 3 | 100 | 0 |
| Total no. cumulative tumors (week = 24) | | | |
| +/+ | 7 | 100 | 0 |
| −/− | 8 | 100 | 0 |
| +/− | 5 | 100 | 0 |
| Total no. cumulative tumors (week = last) | | | |
| +/+ | 7 | 86 | 14 |
| −/− | 9 | 89 | 11 |
| +/− | 4 | *100* | *0* |

[a] Significant ($p \leq 0.05$) by both statistical methods of analysis.
[b] Nonsignificant ($p > 0.05$) by both statistical methods of analysis.
[c] Significant by only one statistical method.
[d] Comparison based on test statistic $p$-value.

## Analysis of SENCAR Tracking System Data

SENCAR skin assay data collected on the tracking system were analyzed using the methods developed. The various response measures and statistical methods were compared. Table 2 describes the studies included in the present report. The evaluation included 24 data sets from initiation/promotion studies that followed, in general, the protocol procedure outlined in Bull et al. (*2*). The univariate response measures were calculated for two study duration times: 24 weeks and the entire study period.

Results of the Gehan-Wilcoxon multiple comparison test were compared to the Cox-type trend test for the analyses of the univariate measures (Table 3). The probability level of the test statistic was the basis for comparison. Using an $\alpha$ level of 0.05, the data sets were

**Table 4. Comparison of response measures.**

| | Agreement, %[a] | |
|---|---|---|
| | 24-Week data | Last-week data |
| Univariate | | |
| Time to first tumor vs. | | |
| Integral w.r.t. time | 95 | 92 |
| Total no. permanent tumors | 90 | 92 |
| Total no. cumulative tumors | 95 | 90 |
| Integral w.r.t. time vs. | | |
| Total no. permanent tumors | 95 | 100 |
| Total no. cumulative tumors | 100 | 100 |
| Total no. permanent tumors vs. | | |
| Total no. cumulative tumors | 95 | 100 |
| Multivariate | | |
| Agreed with univariate results in 8 of the 11 data sets analyzed. | | |
| Multivariate test found significant differences in 2 of the 3 analyses for which there was disagreement. | | |

[a] Based on Gehan-Wilcoxon test statistic: $p$-value $\leq 0.05$ vs. $p$-value $> 0.05$.

**Table 5. Effect of early cut-off time.**

| | Number of studies | | | | |
|---|---|---|---|---|---|
| | Week[b] | | | | |
| Response measure | 24 | 28 | 32 | >32 | Total |
| Integral w.r.t. time | | | | | |
| Significant[a] | 11 | 3 | 1 | 3 | 18 |
| Nonsignificant | 3 | – | – | 3 | 6 |
| Total no. current tumors | | | | | |
| Significant | 10 | 4 | 1 | 2 | 17 |
| Nonsignificant | 3 | – | – | 4 | 7 |
| Total no. cumulative tumors | | | | | |
| Significant | 9 | 4 | – | – | 13 |
| Nonsignificant | 3 | – | – | 4 | 7 |

[a] $p \leq 0.05$, Gehan-Wilcoxon Multiple Comparison Test.
[b] Earliest week minimum $p$-value obtained (i.e., maximum difference observed).

categorized according to the agreement of the results from the two test methods. For all response measures and time periods, the Gehan-Wilcoxon test showed greater power that the Cox-type trend test. The difference between the methods was less in the analysis of time to first tumor than in the integral or total tumor count analyses. In every analysis of tumor count data where the results differed between the two test methods, the Gehan-Wilcoxon test had a higher significance level, probably because it tests for a wide range of alternative hypotheses, whereas the trend test tests for a monotone dose response.

The univariate response measures were compared based on the significance level of their test statistic using the Gehan-Wilcoxon method (Table 4). An $\alpha$ level of 0.05 was used as the cut-off point for significance. For both the 24-week data analysis and the entire study period analysis, the percent agreement between response measures was at least 90%.

Eleven of the 24 data sets were analyzed by using the multivariate stratified proportional hazard model of Prentice et al. (*13*). In 8 of the 11 cases, the results agreed with those of the univariate analysis. In two of the three cases where there was disagreement, the multivariate test statistic showed a significant association between dose of the compound and tumor occurrence.

The univariate methods of analysis for the tumor count integral and total tumor count assume that an animal's tumor observation is censored at its value at the time of death. This approach could possibly bias the analysis and lead to erroneous results. With both measures, animals dying at different times with equal responses are treated equivalently. Intuitively, tumor-free deaths occurring at the end of the study should be considered greater evidence for a no-treatment effect than tumor-free deaths occurring earlier in the study. Similarly, a death occurring early in the study in an animal with a specific tumor response should be considered greater evidence for a toxic effect than a death occurring later in the study in an animal with the same tumor response. To examine the effect of mortality, the study time periods for data analysis were varied. The maximum difference (minimum test statistic $p$-value) in tumor occurrence was observed by 28 weeks in most of the data sets analyzed (Table 5).

## Discussion

Statistical methods supply the necessary procedures to quantify the strength of the evidence in support of the hypotheses under study. Thus, the methods used should be intuitively reasonable and easily interpretable, with sufficient power to detect meaningful differences. The advantages and disadvantages of various response measures and statistical methods for hypothesis testing of mouse skin assay data have been examined. Traditional methods that ignore survival differences, multiplicity of tumors, and the pattern of tumor development may result in a loss of statistical power. Each method has its own deficiencies. None individually are sufficient to reliably describe and test assay data for toxicity. For example, unadjusted tumor incidence data should only be interpreted in conjunction with an examination of the survival patterns in the treatment groups. Methods that account for these factors eliminate biases and thus aid in ensuring the validity of the test results. Adjustment for these factors would be especially important for less potent compounds or studies based on low doses. The exact consequence of ignoring the potential analyses problems would depend on the characteristics of the data set in question.

Response measures and methods of analysis were derived for use in testing the toxicity of a compound when survival data, tumor count data, and time of tumor occurrence are available. Only nonparametric and semiparametric techniques were considered to eliminate the need to specify a response distribution. The methods developed are advantageous for several reasons. The analyses include every animal, not just animals surviving to a given time. Tumor multiplicity is accounted for by each of the response measures, and the duration and regression of tumors are accounted for by several of the univariate methods. The multivariate response measure

considered is the only measure examined that is not biased by any of the characteristics of the data. It accounts for mortality and time of tumor occurrence. Intuitively, this method seems the most reasonable; however, several assumptions are necessary to ensure its validity. The stratified proportional hazard model assumes that the hazard of the dose groups is proportional to that of the control group and that tumors occur in a single animal as a nonhomogenous Poisson process. The model does not allow for extra variability to account for differences in susceptibility among animals (*7*). Additionally, this method is computationally the most difficult, since the computer software needed for computation is not readily available.

Each of the univariate response measures has limitations, and thus the analysis of the mouse skin assay data should not be confined to one of these measures. A thorough analysis should include an examination and test for equality of survival distributions, a comparison of time to first tumor distributions and a comparison of tumor multiplicity. Each of these measures contains important information that should be used to interpret the results of an experiment. Although there were relatively few cases in which the analysis of the integral and total tumor count response measures differed, the integral accounts for tumor duration and is thus preferable to the total tumor counts as a response measure. The strong performance of the Gehan-Wilcoxon multiple comparison test in the analysis of the SENCAR tracking system data leads to the recommendation for its use. As a result of these analyses, it can be stated that if an association between tumor response and dose of a compound is found and it is desirable to examine the data for a monotone dose response, a trend analysis using the Cox-type trend test can be performed. By conducting the analysis in this manner, all available information on the toxicity of a compound is fully used.

The research described in this paper has been peer reviewed by the U.S. Environmental Protection Agency and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

## REFERENCES

1. Slaga, T. J., Fischer, S. M., and Triplett, L. L. Comparison of complete carcinogenesis and tumor initiation and promotion in moue skin: The indiction of papillomas by tumor initiation-promotion—a reliable short term assay. J. Am. Coll. Toxicol. 1: 83–99 (1982).
2. Bull, R. J., Robinson, M., Laurie, R. D., Stoner, G. D., Greisinger, E., Meier, J. R., and Stober, J. A. Carcinogenic effect of acrylamide in SENCAR and A/J Mice. Cancer Res. 44: 107–111 (1984).
3. Coomes, R. M., and Hazer, K. A. Statistical analyses of crude oil and shale oil carcinogenic test data. In: Advances in Modern Environmental Toxicology, Vol. 6 (H. N. MacFarland, C. E. Holdsworth, J. A. MacGregor, R. W. Call, and M. L. Lane, Eds.), Princeton Scientific Publishers, Princeton, NJ, 1984, pp. 167–186.
4. Drinkwater, N. R., and Klotz, J. H. Statistical methods for the analysis of tumor multiplicity data. Cancer Res. 41: 113–119 (1981).
5. Holland, J. M., and Frome, E. L. Statistical evaluations in the carcinogenesis bioassay of petroleum hydrocarbons. In: Advances in Modern Environmental Toxicology, Vol. 6 (H. N. MacFarland, C. E. Holdsworth, J. A. MacGregor, R. W. Call, and M. L. Lane, Eds.), Princeton Scientific Publishers, Princeton, NJ, 1984, pp. 151–166.
6. Hasselblad, V., Stead, A., and Herderschert, T. Bioassay: a system for fitting dose-response curves, a user's guide. In-house documentation, U.S. EPA, Health Effects Research Laboratory, RTP, NC, December, 1983.
7. Crump, K. S., and Ng, T. A study of statistical methods for testing for a dose-related effect in skin painting studies. Response report for U.S. EPA, Health Effects Research Laboratory, Cincinnati, OH, Meta Systems, Inc., 1983.
8. Gehan, E. A. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 52: 203–222 (1965).
9. Mantel, N. Ranking procedures for arbitrarily restricted observations. Biometrics 23: 65–78 (1967).
10. Tarone, R. E., and Ware, J. On distribution-free tests to equality of survival distributions. Biometrika 64: 156–160 (1977).
11. Tarone, R. E. Test for trend in life table analysis. Biometrika 62: 679–682 (1975).
12. SAS Institute Inc. SUGI Supplemental Users Guide: Statistics. SAS Institute Inc., Cary, NC, 1983.
13. Prentice, R. L., Williams, B. J., and Peterson, A. V. On the regression analysis of multivariate failure time data. Biometrika 68: 373–379 (1981).