NTP Statistical Methods Working Group

Report on Analysis of Tumor Data in Photocarcinogenesis Studies

May 2004

NTP Statistical Methods Working Group

Dr. Gregory J. Carr, Procter and Gamble Corp., Cincinnati, OH
Dr. Thomas R. Fears, National Cancer Institute, Bethesda, MD
Professor Nancy Flournoy, University of Missouri, Columbia, MO
Professor Barbara C. Pence, Texas Tech University Health Sciences Center, Lubbock, TX
Professor Bruce W. Turnbull, Cornell University, Ithaca, NY
Professor Walter W. Piegorsch, University of South Carolina, Columbia, SC (Chairman)

Program Staff

Dr. Paul Howard, Food and Drug Administration
Dr. Barbara Shane, National Institute of Environmental Health Sciences

Background

The National Toxicology Program (NTP) has a long-standing history supporting the development and application of statistical methods to analyze data from carcinogenicity assays with laboratory animals. An important consideration within this regard includes recognition and adjustment for differential animal survival on the expected tumor incidence. For analysis of animal data on skin photocarcinogenicity, several statistical methods have been published over the past few years. Although many of these have been developed to analyze the important problem of tumor multiplicity in carcinogenicity studies, it is not clear which, if any, of these methods would be appropriate for analysis of data from photocarcinogenesis studies.

The NTP has in progress phototoxicology and photocarcinogenesis studies at the Center for Phototoxicology at the US National Center for Toxicological Research. These studies expose mice (e.g., SKH-1 hairless mice) to light containing Ultraviolet B, with or without an additional exposure to a potential carcinogenic agent. Since a number of these studies are at or near their final data collection phase(s), identification of proper methods for the statistical analysis is necessary, and the Program held a public meeting on 20 August 2003 to discuss the various statistical issues associated with assay data analysis. At that meeting, statisticians who had developed methods for analyzing photocarcinogenesis experiments discussed these issues, but could not come to a consensus as to how photocarcinogenesis data should be analyzed. As a result, this Working Group on Statistical Methods was convened to provide guidance to the NTP on these important issues.

The Working Group met to discuss and evaluate the available statistical methods for analysis of skin photocarcinogenicity data on March 9, 2004, at the Adams Mark Hotel in Columbia, SC. This report constitutes a summary of the Working Group's deliberations and conclusions, for submission to the NTP Board of Scientific Counselors.

Introduction – The Photocarcinogenesis Assay

The arm of the NTP overseeing and operating the photocarcinogenicity assay is the National Center for Toxicology Research (NCTR). At its Center for Phototoxicology, several types of data are collected, including the number of lesions on each animal, the day of appearance of each lesion, the size of lesion on a weekly basis, the number of lesions on each animal over time, and pathology of lesions (diagnosis of type of lesion) at sacrifice. (The presence of a lesion is recorded once it is larger than 1 mm.)

The Working Group heard a report on current NCTR studies on the potential of α and β hydroxy acids to enhance or synergize photocarcinogenesis caused by sunlight. These acids are found in a number of cosmetics, and were nominated for study by the NTP Toxicology and Carcinogenicity program. The cosmetics remove wrinkles by acidifying the skin, resulting in the loss of adhesion and subsequent sloughing off of the surface cells. This allows water to penetrate between the cells resulting in micro scale edema of the skin. The question is whether the removal of the epidermal cells will stimulate cell proliferation, a process that can lead to carcinogenesis. It is also possible that edema of the skin could alter the optical properties of the skin and this in turn would exacerbate the development of skin cancer by sunlight. Thus, it is of interest to determine whether the

application of skin creams containing over-the-counter concentrations of such acids will act synergistically with UV light in the development of mammalian skin tumors. The SKH-1 (hairless) mouse serves as the experimental animal model, due to its long-standing, successful use in these types of studies.

The chosen α hydroxy acid was glycolic acid (or hydroxy acetic acid) while the chosen β hydroxy acid was salicylic acid (or 2-hydroxy benzoic acid), which is incorporated into cosmetics at concentrations as high as 10%. These two components were singled out because mechanistically it was predicted that they would cause an effect over and above the others in the cosmetic creams.

Since it is known that the highest peak of cell proliferation is about 16 hours after exposure to the cream, it was decided to treat the mice on weekdays between 8:00 am and 10:40 am with 7.5 μL of cream (to cover an area of approximately 35 cm^2) and with simulated solar light between 12:00 and 3:30 pm. (This is a departure from previous studies at, e.g., Argus Laboratories, where animals were treated on alternate weekdays with cream followed by light and on intermediate days with light followed by cream.) The animals are placed 2 m away from the light source, which is a 6.5 kW xenon light solar simulator. The amount of light generated from this lamp is similar to sunlight over the wavelengths of 300 to 800 nm. The ratio of UVA/UVB is 21:1. (This is quite different from tanning lights that emit a 1:1 ratio of UVA/UVB.) The amount of light that the animals receive is measured on a daily basis and the lamps are calibrated weekly.

Randomization of animals in cages on racks and position of racks in relation to the light source is built into the experimental design. Each mouse is caged separately. During exposure to the simulated sunlight, 72 mice are placed on a rack so that each cage is at the same angle to the light. The room housing the lights can accommodate 8 racks, so that a total of 576 mice can be exposed simultaneously. It is recognized that the amount of light received at the outer end of the rack is 15% less than the amount in the middle of the rack, thus the cages are rotated daily to mitigate any positional effects of the lights. The racks are also rotated daily between exposure positions to further eliminate any positional effects. Racks are rotated weekly in the animal holding room. These efforts are thought to reduce any biases in the study due to rack position, room position in the light, or location in holding room.

Animals are assigned randomly to each of 48 dose groups (see Table 1, below), thus minimizing loading or a scheduled-removal bias in the results. The same randomization is used in all studies.

The animals are treated for 40 weeks and then sacrificed 12 weeks later. Body weights are collected throughout the study at weekly intervals and standard statistical analyses are performed on this observed outcome. At sacrifice, pathology of the lesions on the skin is performed, as well as gross pathology of the rest of the animal. The lesions on the skin are classified as to whether they are hyperplastic, squamous cell carcinomas, carcinomas, or carcinomas *in situ*. (The lesions are measured with a micrometer and their location on the animal is determined along with the week in which each lesion is first seen.)

Biopsies of lesions are not taken during the course of study. If an animal has a lesion that is 10 mm or greater at the weekly evaluation, the animal is removed from the study and sacrificed. (This is done to keep lesions from growing and merging together, which would making it difficult or

confusing to determine the discrete number of lesions per animal.) The location of all the lesions are made on a weekly basis and the size measured. Lesions are classified as being <1 mm, 1-2 mm, 2-3 mm, 3-5 mm, 5-7 mm, 7-10 mm, or > 10 mm. Lesion multiplicity of specific sizes are noted and each lesion is classified according to the type of histological change. To identify location and size of lesion, all mice are photographed at sacrifice and the location of the lesions is marked on the photograph. This permits the measurement of the lesion at a later date if required.

Table 1: Sample treatment design for glycolic acid and salicylic acid photocarcinogenesis studies.

Treatment	Amount of simulated solar light			
	No light	7J CIE/cm^2	14J CIE/cm^2	21J CIE/cm^2
None	18*	18	18	18
Control cream	18	18	18	18
4% Glycolic, pH 3.5	18	18	18	18
10% Glycolic, pH 3.5	18	18	18	18
2% Salicylic acid, pH 4.0	18	18	18	18
4% Salicylic acid, pH 4.0	18	18	18	18

* Number of males and females on each treatment

The gross lesions are documented and standard histological sections are taken of the skin in the region of lesions and away from lesions (control areas) on each mouse. Histological analysis of each lesion is made and each gross skin lesion is linked to the histological slide of the lesion by a numerical identifier. A table is constructed showing the number of lesions of the various sizes from male and female mice. The multiplicity of lesions of specific sizes (based on clinical observations) and tumor multiplicity (based on histological classification) are recorded.

Statistical Issues

As presented to the Working Group, a number of statistical issues associated with analysis and interpretation of data from these photocarcinogenesis studies appeared relevant for current and future statistical study; see Table 2 for a summary list, many of which are admittedly inter-related. In some cases, these issues are being or have already been addressed by NTP scientists as part of the assay's design and implementation, while in other cases the issues remain open for further consideration.

This report will comment specifically on the two issues in Table 2 felt to be most critical in terms of having the greatest impact on interpretation of the data: (1) how to account for multiplicity, i.e., how to adjust standard methods of analysis or to devise new methods of analysis that can account for the presence of one lesion on an animal vs. the presence of two or more lesions on another animal and how to view this endpoint among the multiple endpoints being recorded; and (2) how to adjust and/or perform dependent censoring of animal data, i.e., how to correctly base the censoring on tumor size and/or tumor multiplicity and how to incorporate this into the statistical analysis. A third issue, how to incorporate the important aspect of interaction ('synergy') between UV

exposure and chemical agent exposure, was also felt to be important; the Working Group strongly advises the NTP to build this aspect into any methods derived from the recommendations given in this report.

The emphasis on these selected issues should not be interpreted as a dismissal of the other topics listed in Table 2: The Working Group encourages the NTP to visit or revisit, as appropriate, these other issues when considering where and how to direct resources for further statistical research and applications.

Table 2: Potential statistical issues in the design and analysis of NTP photocarcinogenesis studies (in alphabetical order).

• Analysis of body weight data
• Analysis of multiple tumors per animal
• Analysis of survival time data
• Analysis of time-to-tumor data and lesion progression data
• Cage rotation/randomization
• Clustering (spatial) of lesions
• Dependent censoring
• Dose selection, number, and spacing
• Dosimetry
• Interaction (a.k.a. synergy) assessment of UV and chemical agent
• Interim sacrifices
• Multiple types of study outcomes
• Nature of 'zero-dose' controls (untreated, vehicle, etc.)
• Order of UV/chemical agent exposure regime
• Rater/observer bias
• Sample size allocation/selection
• Severity indices for tumor progression and ordered categorical regression analysis of tumor burden
• Standardized stopping rules
• Tumor growth models

Note that use of the term *multiplicity* describes two inter-related aspects of the photocarcinogenicity data. First is the occurrence and analysis of *multiple skin tumors* per animal. The Working Group is informed that if every cell on the back of an exposed or control animal has an equal chance of developing into a tumor, and if the treatment alters this capacity, then multiplicity will be very important from a biological point of view. Second, the Group also recognizes that *multiple endpoints* are being recorded in these photocarcinogenesis studies, including: (a) animal survival, (b) time to (first) tumor of a specified size for each animal or group, (c) tumor burden on each animal during the course of the study, and (d) pathological nature of the tumors. Table 2 distinguishes these two issues by referring to the former as "Analysis of multiple tumors per animal" but to the latter as "Multiple types of study outcomes."

Multiplicity

During the Working Group's meeting, the issue of tumor multiplicity was raised a number of times. Clearly, multiplicity of tumors on the skin — typically on the backs — of the test animals is an important consideration, since squamous cell tumors can occur in multiple numbers in humans. Also, differential multiplicity can be used as an important distinguishing factor when assessing differences among exposure groups. Thus accounting for multiplicity can build the *rate* of tumor formation/development into the statistical analysis.

The Working Group recognized that one can also build many of the other endpoints highlighted in Table 2 — such as tumor burden, type, or clustering — into a statistical analysis. If resources permit, this would be advantageous. In any case, the Group felt that any consequent analytic methodology should be as parsimonious as is possible. (Simply put, the simplest goal of interest is one of screening potential tumorigenic agents. We don't need to "model cancer" here, we just need to develop models and methods that can analyze multiple tumor data in an effective fashion.) As such, the Working Group recommends some form of semi-parametric model, flexible and robust enough to overcome possible violations of highly parameterized model assumptions that future data sets might present; see Gail *et al.* (1980) for an early example of the pertinent paradigm. This contrasts with an early, informal charge to the Group, which encouraged a review and selection of two competing methods for analyzing photocarcinogenesis experiments that had been proposed by other Program staff. Both those approaches were acknowledged to be excellent efforts in statistical modeling, and certainly deserving of both the Group's consideration and of publication in peer-reviewed journals. It was felt, however, that both approaches seem to have already "bought in" to an earlier work by Kokoska *et al.* (1993) that appealed to a single-hit exposure paradigm for tumor development. The Working Group felt that any statistical analysis should instead view the assay data afresh, and in particular try to incorporate the fact that exposure is essentially continuous, or at least has a multiple-induction capability. (It was also felt that an extensive literature search should be conducted as part of this effort. The Program did provide the beginnings of such for the Working Group, but time constraints prevented any in-depth effort to be completed prior to the Group's meeting. The NTP is encouraged to continue this effort, including review of both the biomedical and the reliability literature for sources that might be pertinent to the issues raised in this report.)

The Group's specific recommendation is to start "from the ground, up" and consider a semi-parametric modeling strategy. It is suggested that a non-homogenous Poisson process framework be used, including compound aspects to account for as many as possible/desired of the complexities the tumor data present (many of which are mentioned above). This could include joint modeling of important longitudinal/recurrent events along with terminal events seen in the data. Thus outcomes such as differential tumor type(s) and longitudinal tumor growth could also be incorporated into the larger modeling/analysis strategy. These issues are summarized in more detail in a technical Appendix to this report, below. The effort to develop these ideas will likely involve substantive statistical and subject-matter expertise, including careful interactions between the scientists performing the assay and the statistical research staff. The Working Group strongly encourages the NTP to devote research resources towards this goal.

Dependent Censoring

A second, important consideration identified a number of times during the Working Group's deliberations was the concern that, apparently, the assay protocol requires animals whose tumors grow too large be removed from the study. While biological considerations for this strategy are reasonable (coalescing tumors make it or confusing to determine the discrete number of lesions per animal; animals with large tumors may also be so moribund as to require sacrifice for humane reasons), it gives rise to a concern about the presence of *dependent censoring*. Clearly, removal of the animals truncates the malignancy process. If left unadjusted, such censoring can detrimentally affect the final inferences made on the tumorigenicity of the test agent and/or the UV exposure. While it has no specific recommendations for how to address this issue, the Working Group strongly advises the NTP to (i) study the full impacts of this form of dependent censoring on the statistical methods finally chosen for use with this assay; (ii) determine how such censoring should be performed to avoid problems identified in (i); and (iii) use this information to develop statistical adjustments for the censoring effect. An in-depth literature search would include the growing literature on joint modeling of longitudinal process (e.g., appearance times of multiple lesions) and a dependent terminating event (e.g., early, outcome-dependent sacrifice); useful touchstone articles in this regard include Cook and Lawless (1997) or Huang and Wang (2003). This material would be a natural place to begin the effort.

Other Issues & Closing Comments

The Working Group is pleased to present this short set of recommendations to the NTP, and applauds the Program for its foresight in considering the statistical issues associated with this important assay system. The recommendations made herein are meant to provide guidance and counsel, and should not be interpreted as any form of suppression or criticism of the Program's efforts to date.

The Working Group's primary recommendation is to revisit the modeling and analysis of the data from these photocarcinogenesis studies, in order to consider simple, flexible models for the tumorigenic outcome(s). The Group suggests a point-process regression modeling approach, but other parsimonious models may also be considered that can provide accurate assessments of the response of SKH-1 hairless mouse skin to chemical and/or UV exposure.

The Working Group also recommends that the Program engage in study of the important issue of dependent censoring. This issue was seen to be a major concern. A number of other issues were also noted that could affect any methods finally chosen for analysis of the tumorigenicity data; see selected entries in Table 2. Note that for any methods chosen for the data analysis, the Working Group felt unequivocally that the important feature of interaction ('synergy') between UV exposure and chemical agent exposure should be built into the statistical model.

The Working Group also noted with interest that the (spatial) location of observed lesions on the skin of each mouse could be used to enhance the distinguishability/identifiability of differences among exposure groups. Such *location clustering* may be important if some parts of the skin are more exposed than others due to the position each mouse assumes during the light exposure. Alternatively one may be interested whether the same number of lesions on a mouse are in close

proximity (i.e., in a cluster) or whether they are more dispersed. A large literature on spatial point processes is available here that could be studied to some profit; see, e.g., Diggle (2003).

Lastly, the Working Group did recognize that more advanced modeling efforts could be applied to these data. The Group's deliberations were centered on the important issue of carcinogen screening, and this drove much of its desire to see simple, flexible methods developed and employed for the data analysis. It is understood, however, that development of this assay system presents an important opportunity to also study selected mechanisms of carcinogenesis, using perhaps different designs and model characterizations. The Group encourages the NTP to take every opportunity to study issues such as tumor initiation/promotion/progression, immune suppression, and possible (differential) stress due to handling of these animals. This could be an area where multi-stage, birth-death, biomathematical models may be useful. (Note however, that the Working Group specifically felt that the simple multistage model often seen in cancer risk assessment did not seem to be appropriate for the skin tumor endpoint observed in this photocarcinogenicity assay.)

Technical Appendix: Framework for the Point Process Approach

A flexible model using a point process approach for the analysis of tumor data from a photocarcinogenicity experiment consists of three parts, each corresponding to an important endpoint of the photocarcinogenicity assay.

Part 1. Incidence model for multiple tumors

To account for multiple tumors, one can apply an Andersen-Gill multiplicative intensity point process regression model. Predictor variables would include UV exposure level, treatment (cream) dose, UV×treatment interaction, animal's sex, animal's initial weight, and other pertinent covariates. In this model, the lesions are essentially assumed to appear according to a (time) inhomogeneous Poisson process, such as described in sources such as Hougaard (2000, Ch. 9), Kalbfleisch and Prentice (2002, Ch. 9), or Therneau and Grambsch (2000, Chs. 8.5, 9.5). These methods can be easily implemented using standard survival analysis software for performing Cox proportional hazards analyses. The software can be manipulated into performing the maximum (partial) likelihood analysis of the multiple tumor times with this model. Use of PROC PHREG in SAS is described in Hougaard (2000, p.317); use of the *coxph* function in S-plus is in Therneau and Grambsch (2000, p. 190). If the mice cannot be considered homogeneous, or if interest exists in enlarging the model, one may also introduce a random or *frailty* effect for "mouse." Details on how to do the computations in S-plus are given in Therneau and Grambsch (2000, Ch. 9). The model is also discussed in Hougaard (2000, Ch. 9).

Since mice are only examined weekly, it may be more natural to consider a discrete time version of the continuous intensity model; see, e.g., Jiang *et al*. (1999). Here again, standard software can be applied for the analysis.

Part 2. The tumor type model

Multiple skin tumor types are recorded as outcomes from the NTP photocarcinogenesis experiments. To incorporate tumor type into the analysis, suppose observed lesions can be categorized as belonging to one of $J > 1$ per-specified types. (If $J = 1$, skip this model

component.) Conditional on a tumor occurring (see #1, above), its type is determined according to a multinomial probability. (For a more general model, one can use instead a Dirichlet mixture of multinomials.) Unless J is very large, it is a straightforward operation to write down score equations which can be solved to obtain MLEs using numerical analysis software such as MATLAB; see Abu-Libdeh *et al*. (1990).

Part 3. The tumor growth model

A third outcome associated with tumorigenic response and recorded in the photocarcinogenesis assay is how tumors grow over time. Given the occurrence of a lesion and (if included in the model) its type, tumor growth over time can be modeled using standard longitudinal constructions. Typically these are linear or nonlinear mixed-effects regression models. In SAS, such a mixed-effects model can be analyzed using PROC MIXED or PROC NLMIXED, respectively; in S-plus, it can be analyzed using the functions *lme* or *nlme*, respectively. Among the fixed covariates that could be considered in the model, besides the natural ones noted in Part 1, one might consider week of first appearance, number of previous lesions, etc.

Note for all three modeling components (or just two, if we omit tumor type) that the features are *conditionally independent*. Hence the likelihood factors and parameters for each model can be estimated separately, along for inferential purposes with their standard errors. Of particular interest will be the regression coefficients corresponding to the effect of treatment in each of the three components. First, a global hypothesis of no treatment effect in all components can be tested at some significance level, α. If this is rejected, a step-down multiple comparisons procedure can be used to test the effect of treatment in each component separately, while protecting the experiment-wise α level (Hochberg and Tamhane, 1987, §A1.3; Hsu, 1996, §5.1.6).

References Cited

Abu-Libdeh, H., Turnbull, B. W., and Clark, L. C. (1990). Analysis of multi-type recurrent events in longitudinal studies: Application to a skin cancer prevention trial. *Biometrics* **46**, 1017-1034.

Cook, R. J., and Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16**, 911-924.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd Edn. London: Arnold.

Gail, M. H., Santner, T. J., and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **36**, 255-266.

Hochberg, Y., and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: John Wiley & Sons.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

Hsu, J. C. (1996). *Multiple Comparisons*. New York: Chapman & Hall.

Huang, Y., and Wang, M. C. (2003). Frequency of recurrent events at failure time: Modeling and inference. *Journal of the American Statistical Association* **98**, 663-670.

Jiang, W. X., Turnbull, B. W., and Clark, L. C. (1999). Semiparametric regression models for repeated events with random effects and measurement error. *Journal of the American Statistical Association* **94**, 111-124.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edn. New York: John Wiley & Sons.

Kokoska, S. M., Hardin, J. M., Grubbs, C. J., and Hsu, C. C. (1993). The statistical analysis of cancer inhibition promotion experiments. *Anticancer Research* **13**, 1357-1363.

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.