STATISTICAL METHODS FOR TOXICOLOGY

Paige L. Williams
Associate Professor of Biostatistics, Harvard School of Public Health, USA

Keywords: Carcinogencity, developmental toxicology, reproductive toxicity, animal bioassay, dose-response, litter effect, generalized estimating equations, quasi-likelihood methods, risk assessment

Contents

Summary

Toxicology concerns the adverse effects of chemical compounds, drugs, or environmental agents on the health of humans or animals. Because of the importance of protecting human health, researchers in the field of regulatory toxicology have developed standard approaches for assessing the risks posed by such exposures. One of the primary statistical approaches developed for this purpose is the assessment of dose-response; that is, evaluation of whether an adverse effect appears to be increasing in frequency or severity as the "dose" or level of the exposure increases. A second key area is that of quantitative risk assessment, which aims to provide quantitative estimates of the risk of a toxic event as a function of dose, set confidence limits on the exposure level associated with a certain risk level, and develop acceptable or "safe" exposure limits. While toxicology as a field is quite broad, there have been some key areas of statistical developments addressing toxicological problems. These include applications to carcinogenicity, developmental toxicity, and reproductive toxicity. In this article, an overview of statistical approaches within each of these areas of toxicology will be provided. In addition, some of the basic approaches for evaluating dose response relationships and conducting quantitative risk assessment are discussed.

1. Introduction

Toxicological data arise both in screening new drugs for safety and efficacy and in evaluating environmental agents for possible adverse effects. In the case of therapeutic agents, ethical considerations require that they be screened on animals before pursuing clinical trials in human subjects. In the case of environmental exposures, epidemiologic studies are limited by long latency periods before the effects of interest and by the difficulties in accurately estimating exposure levels for individuals. As a result, investigation of toxicological effects has relied heavily on designed experiments in animals. Tests in animals include evaluations of acute toxicity (such as eye or skin irritation), chronic toxicity, and long-term bioassays for carcinogenicity. Assessment of possible teratological effects may involve exposure of test substances to males and/or females prior to mating and to females during gestation and lactation. Such studies may continue for several generations, and evaluate effects both on the reproductive capacity of the animals and on the fetal development of the offspring.

The need for evaluation of drugs and other chemical compounds has led to the development of statistical methods for analysis of such toxicological data. In fact, this has proven to be a fertile area for statistical research in general, and many of the modern statistical methods originated or were refined via applications to toxicological data. The goals of the various assessments described above may be very different, but there are several common themes in the approaches to analysis of such data. For example, the standard design of carcinogenicity studies, developmental toxicity studies, and reproductive toxicity studies each include a control group and 2 to 3 active exposure groups. The resulting toxicological responses can often be described as the percent responding at each exposure level, or the mean response at each exposure level. Due to the limited number of dose groups, dose-response models for toxicology are restricted to include just 2 or 3 parameters in order to be identifiable. Dose response models for cancer risk assessment have been developed to some extent on the basis of biological mechanisms of carcinogenesis. While mechanisms of reproductive and developmental toxicity are not as well understood, dose response models in these areas have been borrowed heavily from those pertaining to cancer risk assessment.

2. Applications of Biostatistics to Toxicology

2.1. Carcinogenicity Studies

Despite the advances made in diagnosing, treating, and preventing relapses in cancer, this issue remains one of the foremost health concerns of current society. Epidemiologic and laboratory research has shown that cancer occurrence is associated with a variety of factors, including diet, exercise, and genetic background. However, there is no doubt that the incidence of certain types of cancer is also increased by various environmental exposures. In addition, investigations have shown that there are often interactions between genetic background and environmental factors in predisposing individuals to cancer. Efforts to identify environmental agents and food additives that increase cancer rates are therefore of utmost concern. Cancer risk assessment has developed as an active area of research in the field of toxicology. Epidemiologic investigations, while ideal in their relevance to human populations, are

limited by the long latency period for many cancers, the difficulties in estimating individual exposure levels, and the complexity of controlling for diverse lifestyle choices. As a result, bioassay studies in animals have become an integral part of cancer research.

Long-term animal carcinogenicity bioassays represent one of the key components of risk assessment for exposure to chemical compounds and other environmental agents. The primary purpose of these experiments is to evaluate whether exposure to the chemical alters the normal pattern of tumor development. Topics of statistical research in the area of long term animal bioassays include optimal study design, tests for dose-response, survival methods for assessing time to tumor development, use of three-state models for tumor development, and quantitative risk assessment.

The standard design of a long-term animal bioassay includes a control group and 2 or 3 exposed groups, where animals are randomly assigned to exposure groups. Experimental guidelines recommend that bioassays be conducted in both males and females of at least two species of animals, which in practice are most often mice and rats. The typical sample size is 50 animals per dose group for each sex and species. Animals are exposed daily over a time period which constitutes the majority of their natural lifespan. At study termination, surviving animals are sacrificed and examined at numerous sites for tumor development. Some animal bioassays also include interim sacrifices, which aid in the assessment of occult tumors (i.e., internal tumors which cannot be observed visually or by palpation in live animals).

The objectives of carcinogenicity bioassays are to screen chemicals by testing for a dose effect, to quantify dose-response relationships, and to help elucidate carcinogenic mechanisms. Towards these goals, the analysis of data from animal bioassays tends to focus on either the lifetime tumor incidence or the time to tumor. Interim sacrifices may also provide information on pre-neoplastic lesions, but these tend to be considered more from a qualitative perspective in elucidating mechanisms of carcinogenicity than a quantitative one.

Table 1 shows the typical data summarized for the purpose of comparing lifetime tumor incidence. In this data layout, there are $K+1$ dose groups ($i=0$ for the control group), with n_i animals exposed in dose group i and x_i animals observed to have tumors prior to or at the terminal sacrifice. The two standard tests used to compare lifetime tumor incidence are Fisher's exact test, in which the number of animals with tumors in a single dose group are compared to those in the control group, and the Cochran-Armitage Trend test. The latter test is a large sample method for testing whether there is a linear increase in tumor rates with increasing exposure. If we denote the difference between the number of observed and expected tumors within dose group i as $D_i = (x_i - e_i)$, where $E_i = (x_\bullet / n_\bullet) n_i$, and define the difference vector $\mathbf{D} = (D_0, \ldots, D_K)$, dose vector $\mathbf{d} = (d_0, \ldots, d_K)$, and variance matrix $\mathbf{V}$, then the Cochran-Armitage trend test can be written as $X^2 = (\mathbf{d}^{\mathrm{T}}\mathbf{D})^2 / (\mathbf{d}^{\mathrm{T}}\mathbf{V}\mathbf{D})$. This test statistic follows a chi-square distribution with one degree of freedom, and a significant test statistic would be considered as evidence supporting carcinogenicity. However, note that multiple tests are conducted,

given that the bioassay is typically conducted in both males and females of two species, and that many target organs are examined. Significant trends in both sexes of more than one species help strengthen the claim of carcinogenicity.

Tumor Status	Dose Level				Total
	d_0	d_1	$\cdots$	d_K	
With Tumor	x_0	x_1	$\cdots$	x_K	$x_\bullet$
Without Tumor	$(n_0 - x_0)$	$(n_1 - x_1)$	$\cdots$	$(n_K - x_K)$	$(n_\bullet - x_\bullet)$
Total Exposed	n_0	n_1	$\cdots$	n_K	$n_\bullet$

Table 1: Data Layout for Lifetime Tumor Analysis

An example of bioassay data and associated trend tests is shown in Table 2. In this study, female mice were administered one of three doses of the chemical 1,2-dichloroethane, and the proportion with lung tumors was observed. The Cochran-Armitage trend test yields $X^2 = 10.64$ $(\text{df} = 1, p = 0.001)$. A goodness-of-fit test for the linear model is calculated as $X^2_{\text{gof}} = 11.09 - 10.64 = 0.45$ $(\text{df} = 1, p = 0.650)$, where $X^2_{\text{ind}} = 11.09$ is the Pearson Chi-square test for independence; the non-significance of this test suggests that the linear model is appropriate. Based on these results, we can conclude that the data provide evidence of an increasing trend in tumor rates with higher 1,2-dichloroethane doses.

Tumor Status	Dose Level			Total
	0	1	2	
With Tumor	2	7	15	24
Without Tumor	38	43	33	114
Total Exposed	40	50	48	138

Table 2: Lung Tumors in Female Mice Exposed to 1,2-dichloroethane

One difficulty in interpretation of tests for lifetime tumor incidence is that the higher dose levels may be associated with an increased mortality rate, which would reduce the time at risk of developing tumors. For example, consider the carcinogencity study of 1,2-dichloroethane: 92% of the control animals survived to 62 weeks, whereas only 72% of the highest dose animals survived this long. Several survival adjustments have been proposed to account for such treatment toxicity. One approach is to eliminate deaths that occur before the first tumor is observed in any dose group, and then use this adjusted number at risk in conducting Fisher's exact test or the Cochran-Armitage trend test. For the 1,2-dichloroethane data, the number of animals eliminated from the "total" and "without tumor" rows is 3 for dose 0, 2 for dose 1, and 12 for dose 2, yielding a revised Cochran-Armitage trend test statistic of $X^2 = 14.89$.

A comparison of the lifetime tumor incidence between dose groups provides an evaluation of whether exposure increases the frequency of tumors, but does not provide any indication of whether exposure is associated with both more frequent *and* earlier

tumors as compared to control. Statistical methods for comparing time to tumor have been developed for this purpose. If we denote T as the time to tumor onset, then the tumor onset rate for dose group i is

$$\lambda(t) = \frac{\lim_{\Delta \to 0} \Pr\left(t \le T < t+\Delta \,\middle|\, T \ge t, d_i\right)}{\Delta}. \qquad (1)$$

If the tumor of interest is observable (i.e., can be identified visually or by palpation, such as many mammary tumors in rats), then the standard logrank test can be applied to compare dose group *i* to the control group. Extensions of the logrank test can be used to test for trend in increasing exposure, by treating exposure level as a continuous covariate. However, the statistical analysis of such experiments is often complicated by the fact that the progression of many types of tumors cannot be directly observed (occult tumors), and thus the tumor status can only be determined at death. Statistical methods for such analyses take into account the relationship between tumor development and death by either (1) ascertaining or assuming the cause of death for tumor-bearing animals, or (2) using interim sacrifice data to provide additional information on the progression of tumors.

Two primary types of analyses have historically been applied when the cause of death for tumor-bearing animals can be ascertained or assumed. In the situation in which the tumor type of interest has no effect on the animals' risk of death, tumors are referred to as "incidental", and the prevalence method can be applied based on a Mantel-Haenszel statistic to test for equal tumor incidence rates among dose groups. More specifically, a table is constructed for each time of death identifying the number of animals with tumors and the number without in each dose group, among all animals that died at this time point. The Mantel-Haenszel test is then used to pool the results over death times. A special case of this type of analysis is referred to as a Hoel-Walburg analysis; this is based on dividing the length of the study into separate mutually exclusive time intervals and pooling the separate chi-square tests for trend over the time intervals. The National Toxicology Program follows this approach, using the intervals 0-52 weeks, 53-78 weeks, 79-92 weeks, 93 weeks – terminal sacrifice, and terminal sacrifice.

An alternative approach for comparison of tumor prevalence is to use logistic regression to model the probability of tumor adjusting for time of death, and conduct a score test for the addition of dose to the model. The key assumption in treating tumors as incidental is the "representativeness" assumption; that is, that the animals which die with a tumor at time *t* are representative of the animals which remain alive and at risk at time *t*. For the opposite situation in which the tumor type under consideration causes death, a fatal tumor analysis is conducted. In order for the fatal tumor analysis to be an appropriate test of differences in tumor incidence rates, the assumption must be made the tumor presence causes death so rapidly that the time of death for tumor-bearing animals can be used as a surrogate for the time of tumor onset. In this case, a standard lifetable analysis is conducted via a logrank test, treating the time of death with tumor as the time of tumor onset.

When individual tumors within a carcinogenicity experiment can be classified as either incidental or fatal, the separate analyses of incidental and fatal tumors described above can be combined. Some researchers have proposed models for estimating the joint distributions of time to tumor and time to death from tumor. However, a drawback of the methods which require classification of individual tumors as incidental or fatal is that pathologists are often very reluctant to make such identifications, recognizing that many tumors are of intermediate lethality. The U.S. National Toxicology Program (NTP) follows the pragmatic approach of conducting both a prevalence analysis and a fatal tumor analysis; if both are significant, this provides evidence for carcinogenicity, while if neither is significant, a lack of carcinogenicity is suggested. In most animal carcinogenicity experiments, the cause of death cannot be ascertained, and the assumptions of incidental or fatal tumors are extreme and difficult to justify. The recognition of the biases resulting from making such assumptions have led to the development of methods which rely on interim sacrifice data in addition to survival information.

Three-state models for time to tumor have been used extensively to formalize the ideas underlying estimation of tumor incidence and death rates in the presence of interim sacrifice data. Figure 1 illustrates the framework of the three-state model. Based on this framework, "**N**" represents a normal animal that is tumor-free, which can either progress to the tumor state "**T(tumor)**" with transition rate $\lambda(t)$ or die without a tumor by progressing directly to state "**D**" with transition rate $\beta(t)$. Animals that develop tumors at time *x* can also go from state "**T(tumor)**" to state "**D**" with corresponding transition rate $\gamma(t \mid x)$. The transition rates are defined on the basis of the random variables T, which represents the time to first event (either tumor or death), *D*=time to death, and the indicator of tumor presence by time t, denoted as $\delta(t) = I(T < t)$. These transition rates can then be expressed as:

$$\lambda_T(t) = \frac{\lim_{\Delta \to 0} \Pr\left(t \le T < t + \Delta,\ \delta(t) = 1 \middle| T \ge t\right)}{\Delta} \tag{2}$$

$$\beta(t) = \frac{\lim_{\Delta \to 0} \Pr\left(t \le T < t + \Delta,\ \delta(t + \Delta) = 0 \middle| T \ge t\right)}{\Delta} \tag{3}$$

$$\gamma(t|x) = \frac{\lim_{\Delta \to 0} \Pr\left(t \le D < t + \Delta \middle| D \ge t, T = x < t\right)}{\Delta} \tag{4}$$

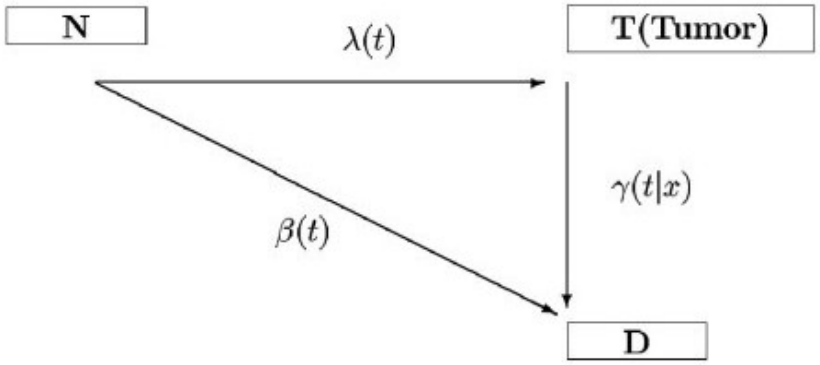

Figure 1. Three-state model for Time to Tumor and Time to Death

Note that $\lambda_T(t)$ differs from the tumor incidence rate $\lambda(t)$ defined previously in that it conditions on the animal being both alive and tumor free. Because of this conditioning, the indicator $\delta(t)=1$ defines that the event is a tumor occurrence rather than death. As a result, $\lambda_T(t)$ is often referred to as the cause-specific hazard for tumor onset or the tumor incidence function.

The primary interest in evaluating data from a carcinogenicity study in the context of three-state models for time to tumor is to assess exposure effects on $\lambda_T(t)$. The death rates $\beta(t)$ and $\gamma(t \mid x)$ may also depend on exposure levels, but are considered nuisance parameters. In order to obtain estimates of the parameters, the Markov assumption is often made that $\gamma(t \mid x) = \gamma(t)$; in other words, given that the tumor has occurred, the instantaneous rate of death does not depend on when the tumor occurred. Since we are assuming that tumors are occult, the observable data at each time point are only D and $\delta(D)$. However, since deaths include both natural deaths and sacrifices, we have the following four types of events: (1) sacrifice, no tumor, (2) death, no tumor, (3) sacrifice with tumor, and (4) death with tumor. The likelihood contribution for each of these events can be written in terms of the transition rates $\lambda_T(t)$, $\beta(t)$ and $\gamma(t)$. To obtain maximum likelihood estimates, particular forms for survival distributions must be chosen, such as the exponential model. Alternatively, discrete time models can be employed by breaking up the length of the study into mutually exclusive time intervals defined by the interim sacrifices, and then computing non-parametric estimates of the transition rates within each of these time intervals. Another useful way to maximize the likelihood is via the EM algorithm. In this context, the missing data are the tumor onset times. In the "E" step, the algorithm estimates the sufficient statistics based on the observed data. In the "M" step, the complete data log-likelihood is maximized given the sufficient statistics estimated by the E step. The algorithm iterates between the E and M step until convergence.

There is a host of other statistical issues which relate to long-term animal carcinogenicity studies. A variety of approaches have been developed to attempt to account for multiplicity in response. Such multiplicity may exist at the level of the tumor, in that there are multiple tumors at a particular target organ, or at the level of the

animal, in that there are multiple tumor sites examined. Another area that has received quite a bit of attention is the incorporation of data on historical controls into analysis of animal bioassay data. This approach is especially useful when the tumor type of interest is known to be quite rare, but the low incidence rate of tumors in exposed groups does not attain significance due to the relatively small sample sized employed.

-
-
-

Bibliography

Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, 1-12. [This is a classic reference for development of biologically based models of carcinogenesis]

Clayson, D., Krewski, D., & Munro, I., eds. (1985). *Toxicological Risk Assessment*. Boca Raton, Fl, CRC Press. [This textbook contains many articles on design and analysis of long-term animal bioassays for carcinogenicity, including methods for low dose extrapolation and confidence bound estimation.]

Diggle, P.J., Liang, K.-Y., & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press. [This textbook provides a review of the methods of estimation by quasi-likelihood and generalized estimating equations].

Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E., & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research; Volume III – The design and analysis of long-term animal experiments*. Lyon, International Agency for Research on Cancer. [This textbook describes methods for design and analysis of long-term animal bioassays, including discussion of incidental and fatal tumor methods and 3-state models of carcinogenesis.]

Hood, R.D. (1996). *Handbook of Developmental Toxicology*. New York, CRC Press. [This textbook provides a review of the study design, outcomes, and methods of analysis for developmental toxicity studies.]

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. London, Chapman and Hall. [This textbook reviews dose-response models for binary outcomes, and describes both maximum likelihood estimation methods and estimation methods which account for over-dispersion.]

Whittemore, A.S. & Keller, J.B. (1978). Quantitative Theories of Carcinogenesis. *SIAM Review* **20**, 1-30. [Classic reference for mechanistic models of carcinogenesis.]

Biographical Sketch

Paige Leigh Williams was born in Hartford, Connecticut, USA; she received her Bachelor of Science in Public Health (1983), Master of Science in Biostatistics (1984), and PhD in Biostatistics (1990) from the University of North Carolina School of Public Health. Dr. Williams has been on the faculty of the Department of Biostatistics at Harvard School of Public Health in Boston, Massachusetts since 1991. She has taught numerous course in statistical theory and methods to graduate students, and has served on

thesis committees for 20-25 doctoral students. Dr. Williams has also served as a senior statistician for the U.S. AIDS Clinical Trials Group (ACTG) since 1991, and in this role directs the design and analysis of clinical trials and observational follow-up studies among HIV-infected subjects. Her primary research interests are quantitative risk assessment, environmental statistics, AIDS clinical trials, and statistical methods for survival data. She has published over 70 articles in peer-reviewed statistical, environmental, and infectious disease journals, serves on several editorial boards, and has served on advisory boards, executive committees, and as a council member of US-based and international statistical societies.