8

Analyzing Toxicity Data Using Statistical Models for Time-To-Death: An Introduction

Philip M. Dixon[1] and Michael C. Newman[2]

[1,2]Savannah River Ecology Laboratory, Drawer E,
Aiken, South Carolina 29802-0005
and
[1]Biomathematics Program, Dept. of Statistics,
North Carolina State University, Raleigh, North Carolina 27695
and
[2]Savannah River Ecology Laboratory, Drawer E,
Aiken, South Carolina 29802-0005

OVERVIEW

In their research toxicologists often use routine testing protocols mandated for regulatory purposes. However, other statistical techniques provide considerably more abundant and precise information than do the standard techniques. Such techniques (called survival analysis, failure time analysis, or life data analysis) are in common use in other disciplines such as epidemiology, clinical medicine, and engineering. They are readily implemented with several common software packages (SAS, BMDP, SYSTAT, S-plus). This review develops a conceptual understanding of these alternate techniques and provides examples of their use.

INTRODUCTION

Lord Rutherford, the eminent British physicist, is often quoted as saying, ''If your experiment needs statistics, you ought to have done a better experiment.'' In some situations, the analysis of toxicity data does not need statistics; an answer is obvious when, for example, a large number of treated animals die while a large number of controls are still alive. In most other situations, however, some statistical analysis is essential. A good analysis both answers toxicological questions and quantifies the uncertainty in those answers. This chapter is concerned with the techniques available to answer some common toxicological questions.

These techniques can be divided into two groups. The first are the techniques mandated by regulation for use in routine toxicity testing. Standard bioassay procedure for a short-term, dose-response experiment is to expose animals for 96 h, count the number of death, and calculate LC_{50}'s and their 95% confidence intervals[1] (Am. Public Health Assoc., pp. 641-645). The focus in these studies is the routine toxicological evaluation of a new chemical or material of unknown constituents. Appropriate techniques are fast, easily performed, and not sensitive to violations of statistical assumptions. However, we find a common tendency for toxicological researchers to uncritically select these routine toxicity testing protocols in their research efforts. This chapter will serve to introduce researchers to some statistical techniques that provide considerably more abundant and precise information than do the standard techniques with only a small amount of additional effort. These techniques, called survival analysis, failure-time analysis, or life data analysis, are widely used in medical and engineering research.[2-4]

Standard toxicity testing analysis starts with data on the percentage of individuals that survive some period (e.g., 24 or 96 h). Survival analysis starts with the times at which individuals die. Collecting time-to-death data involves more work than recording survival to a fixed endpoint, but significant statistical benefits accrue from the small amount of additional work. Our goal in this chapter is to develop a conceptual understanding of the analysis of time-to-death data and provide examples and interpretation of data analysis using SAS[5,6] programs. Our examples were run using version 6.03 of SAS. Survival analysis programs are also available in the BMDP, SYSTAT, and S-plus systems (Table 1). More detailed and more theoretical treatments of survival analysis can be found elsewhere.[2,4,7] Additional discussion of the application of failure-time analysis in engineering and reliability studies can be found in Meeker and Hahn,[8] Nelson,[3] and Nelson.[9]

The toxicological problem is to describe the effects of factors that modify toxicity.[10] Typically, such questions involve the impact of differences in environmental conditions (water quality, pH), different test species, or individual differences (such as size, sex, or acclimation) on toxicity. Statistically, these questions can be answered by constructing a model that expresses how the time-to-death is influenced by various factors. Given a suitable model, the influences of factors can be estimated and hypotheses about them can be tested. Although we will present techniques for both, our focus will be on estimation, because

Table 1
Availability of Computer Software for Analysis of Censored Times-To-Death[a]

	BMDP	SAS	S-PLUS	SYSTAT
Estimate survival distribution	BMDP 1L	PROC LIFETEST	SURV.FIT	SURVIVAL
Test equality of survival	BMDP 1L	PROC LIFETEST	SURV.DIFF	SURVIVAL
Fit accelerated life model	—	PROC LIFEREG	—	SURVIVAL
Fit Cox model	BMDP 2L	PROC PHGLM	COXREG	SURVIVAL

[a] Absence of an analysis is indicated by a dash.

we find it more biologically informative than hypothesis testing. The biological interpretation of estimates is quite different from the interpretation of a hypothesis test, but estimation and hypothesis testing are closely linked in statistical theory.[11]

The effect of a factor can be expressed in various ways. Some typical examples include the average time to death, the median time-to-death (the time at which 50% of the group has died), the concentration at which 50% had died (the LC_{50}), or the hazard function (the instantaneous probability of dying). The effect of a factor can be expressed by a change in any one of these, not just the difference in LC_{50} typically used.

Data from three studies will be used as examples to introduce and clarify the concepts in survival analysis. We will reanalyze two classic data sets, one by Litchfield[12] on the survival of tuberculoid mice administered streptomycin or a placebo, and the second collected by Shepard[13] on the survival of speckled trout fry in water with various low concentrations of dissolved oxygen. We will also describe in some detail an analysis of the effects of individual covariates on survival of mosquitofish (*Gambusia holbrooki*) exposed to As.[14] These studies span a range of complexity from a comparison of 2 groups (mouse data), to a dose-response comparison of 10 groups (trout data), to an analysis of the effects of individual level covariates (mosquito fish data). Data for the mice and trout examples are published in the original papers and are repeated in Tables 2 and 3. Data for the mosquito fish example are available from the authors and discussed in this volume, Chapter 11.

DATA SETS

Consider the mouse data described by Litchfield,[12] a classic example of the estimation of mean time-to-death. Although it is a biomedical study, the principles can be easily applied to metal toxicology. Litchfield presents data on the survival of mice with tuberculosis. One group of 20 mice was treated with streptomycin, while a larger group of 80 was left untreated. The data consist of the number of days that each mouse survived (Table 2). It seems clear that treatment with streptomycin prolongs the life of these mice, but let us use these data to examine some different ways of describing and graphically presenting survival patterns.

The trout data set is one part of large study of the resistance and tolerance of speckled trout (*Salvelinus fontinalis*) to low oxygen levels.[13] Trout fry were acclimated to water containing 10.50 mg O_2/L; the water was then replaced with deoxygenated water containing from 0.77 to 1.77 mg O_2/L. The data are the number of minutes until the fish died (Table 3). The experiment was terminated at 5000 min; any fish still alive was recorded as censored at 5000 min.

The mosquitofish (*Gambusia holbrooki*) data set is part of a study of resistance to metal and metalloid intoxication. Field-collected mosquitofish (754) were acclimated to laboratory tanks and then exposed to 94 mg As/L in a continuous flow-through exposure system. Every 3 h, dead fish were collected, counted, sexed, and weighed. After 102 h, all remaining fish were collected, sexed, and weighed. Again, survivors at the end of the experiment are recorded as censored at 102 h. Genotypes of all fish at eight enzyme loci were determined with starch-gel electrophoresis.[14] Data are not presented here because of the large number of observations.

The routine toxicity testing approach to all of these data sets would be to calculate the percentage survivorship at some end point, often the end of the experiment (e.g., survival to 96 h). In the mouse data, none of the 80 untreated individuals were alive at 60 days, but 30% (6 out of 20) of the streptomycin-treated mice were. However, the choice of 60 days as the end of the experiment was arbitrary. If the experiment was ended at 20 days, the effect of streptomycin would have appeared larger: none of the 80 untreated mice survived, but all of

Table 2
Survival Time of Mice Infected with Tuberculosis[a]

	Number Dying in	
Day of Death	**Untreated**	**Streptomycin Group**
8	1	0
9	4	0
10	0	0
1	10	0
1	10	0
13	7	0
14	19	0
15	12	0
16	4	0
1	3	0
17	0	1
29	0	1
32	0	2
37	0	1
39	0	2
42	0	1
43	0	2
44	0	1
47	0	1
52	0	1
59	0	1
Still alive at 60 days	0	6

[a] One group of 20 was treated with streptomycin; the other group of 80 was left untreated. Experiment was terminated at 60 days, at which time six streptomycin-treated mice were still alive. Data from Litchfield.[12]

Table 3
Survival Times of Trout Exposed to Different Dissolved Oxygen Concentrations[a]

O_2 Concentration (mg/l)	Survival Times (min)
0.77	17, 18, 20, 20, 20, 21, 21, 22, 22, 23
0.94	20, 22, 24, 26, 26, 29, 29, 31, 34, 34, 41
1.10	25, 29, 33, 33, 37, 37, 37, 41, 48, 55
1.16	30, 30, 35, 35, 40, 45, 50, 58, 62, 70, 100
1.36	48, 52, 60, 85, 140, 160, 170, 190, 250
1.43	50, 50, 135, 175, 195, 215, 355, 405, 465, 600
1.55	165, 165, 195, 270, 270, 440, 440, 735, 865, 1400
1.69	195, 225, 270, 270, 440, 675, 995, 1150, 1150, 5000, 5000
1.77	240, 675, 995, 2080, 5000, 5000, 5000, 5000, 5000, 5000
1.86	400, 5000, 5000, 5000, 5000, 5000, 5000, 5000, 5000, 5000

[a] The experiment was terminated at 5000 min. From 9 to 11 fish were exposed to each concentration. Data from Shepard.[13]

the treated mice did. If the experiment had ended at 4 days, there would have appeared to be no effect of streptomycin: all of the mice in both groups were alive. What is needed is some way to describe patterns of survivorship and mortality at all times during the experiment. This can be done using either the survivor or hazard function.

SURVIVOR AND HAZARD FUNCTIONS

The survivor function [Equation (1)] describes the probability that an individuals survives longer than some time, t.

$$S(t) = P[\text{An individual dies after time } t] \tag{1}$$

When the times-to-death of all individuals in a group are known, $S(t)$ can be estimated by the proportion of individuals still alive at time t [Equation (2)].

$$\hat{S}(t) = \frac{\#\text{ alive at time } t}{\text{total }\#\text{ animals}} \tag{2}$$

At the start of an experiment, $S(0)$ is 1.00; over time, it decreases to 0.00 when the last individual dies. If the experiment is terminated before the last individual dies, the survivors are counted into the denominator, but not the numerator, so $S(t_e)$ remains above 0.00. Clearly, this function summarizes all information in a table of times-to-death, but it is not the only way to do so. The hazard function is a closely related alternative that describes different aspects of the pattern of mortality.

The hazard function, or force of mortality,[3] describes the probability of dying as a fraction of the number alive at the beginning of the period. It is mathematically related to the survival function [Equation (3)].

$$h(t) = -d\ \log S(t)/dt = \frac{-1}{S(t)}\frac{dS(t)}{dt} \tag{3}$$

It is often useful to know whether the hazard is constant over time (e.g., 10% of the current survivors will die during any time interval), increases over time (individuals are more likely to die at longer exposure durations), or decreases over time (individuals are less likely to die as exposure duration increases).[3] The shape of the hazard function is a useful tool to help choose a model for time-to-death (see verifying assumptions in Estimation and Hypothesis Testing section). The survivor and hazard functions for the untreated group of mice are given in Figure 1. For these mice, the hazard increases as the duration of exposure increases.

CENSORING

Computing the survival curve for the streptomycin-treated group of mice introduces one complication: not all the animals have died by the end of the experiment. We have partial information on these animals, because we know that they survived 60 days in this study, but we do not know exactly when they died. Censoring is characteristic of survival data, and sophisticated methods to handle very general censoring mechanisms have been developed in the biomedical literature.[2,4] The censoring in this study is very simple; all individuals are censored at the same fixed time, the length of the experiment. Under a general

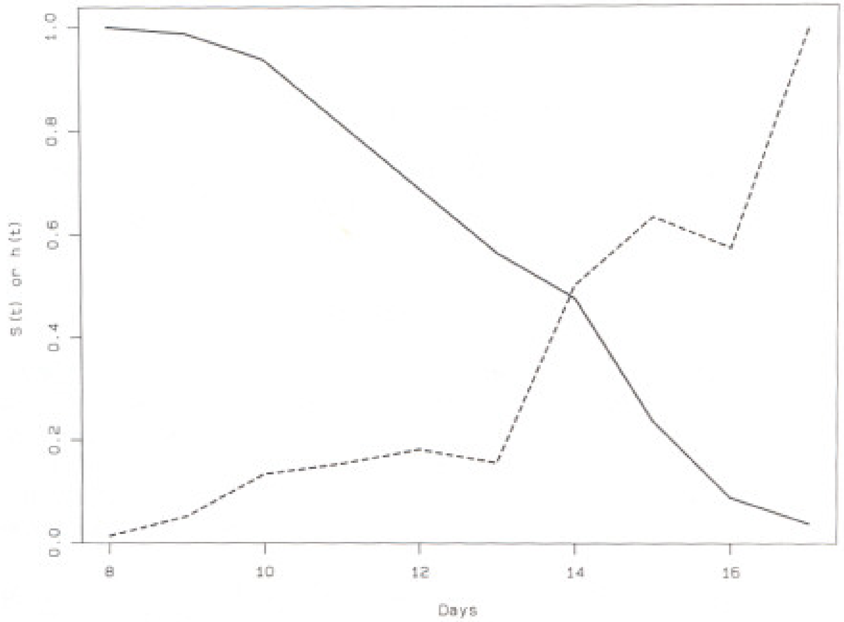

FIGURE 1. Survivor (——) and hazard (- - -) functions for untreated tuberculoid mice. Data from Litchfield.[12]

censoring mechanism, the survival distribution can be estimated by the Kaplan-Meier (also called the product-limit) estimator[2] [Equation (4)].

$$\hat{S}_{KM}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{r_j}\right) \tag{4}$$

where d_j is the number of deaths occurring at time t_j among the r_j individuals alive just before time t_j.

If there is no censoring, this reduces to Equation (2) above. When censoring occurs only at the end of the experiment, the Kaplan-Meier equation reduces to Equation (2) at all times until the end of the experiment, and it has an undefined value at later times.

Observations can be censored for many reasons. Often, in biomedical studies, patients move away and are never heard from again, or they stop participating for some other reason, or, as in the Litchfield data, the study ends before the last animals die. The Kaplan-Meier estimate provides a consistent estimator of the survival distribution for quite general censoring patterns. In general, the data that we will analyze are pairs of numbers (T_i,C_i). T_i is the observed time of death or censoring, and C_i is a censoring flag. If C_i equals 0, T_i is a time-to-death, but if C_i equals 1, T_i is a censoring time.

Like every statistic calculated from data with variability, the estimated proportion surviving to a particular time is not known exactly. The variance around the survival curve can be approximated in several different ways, but Greenwood's formula (see Cox and Oakes[2] for details) is frequently used. For the special case of end point censoring, Greenwood's estimator reduces to Equation (5), the binomial variance, at all times before the end of the experiment.

$$\text{Var } \hat{S}(t) = \frac{\hat{S}(t)[1 - \hat{S}(t)]}{N} \tag{5}$$

where N is the number of individuals in the study
$\hat{S}(t)$ is the estimated survival at time t

The variance is smallest at either end of the survival curve, when nearly all individuals are still alive or when nearly all are dead. SAS computes approximate 95% confidence intervals around the survival curve as 1.96 $\sqrt{\ }$ Var. This confidence interval assumes that the survival estimate is normally distributed, an assumption which is reasonable for large samples. For small samples, the normal assumption may not be appropriate, especially for survival estimates close to 0 or close to 1. In particular, the computed upper of lower bound to the 95% confidence interval may be larger than 1 or smaller than 0, respectively. Other ways to calculate the confidence interval avoid these problems (see Kalbfleish and Prentice,[4] pp. 14-15 for details.)

PROC LIFETEST[6] will calculate and plot the Kaplan-Meier estimate of the

Table 5
Results of Log-Rank and Wilcoxon Tests for the Equality of Survival Distributions

	Log-Rank Test		Wilcoxon Test	
	X^2	$P > X^2$	X^2	$P > X^2$
Mouse data				
Streptomycin vs untreated mice (1 d.f.)	62.67	0.0001	41.26	0.0001
Trout data				
Different O_2 exposures (9 d.f.)	259.30	0.0001	217.4	0.0001
Mosquito fish data				
Male vs female (1 d.f.)	24.95	0.0001	45.83	0.0001
Among size groups, males only (5 d.f.)	37.02	0.0001	63.43	0.0001
Among size groups, females only (5 d.f.)	13.95	0.016	25.43	0.0001
Among GPI-2 genotypes (5 d.f.)	17.30	0.004	19.50	0.0015
Among genotypes, males only (5 d.f.)	8.74	0.12	5.41	0.37
Among genotypes, females only (5 d.f.)	9.32	0.097	13.33	0.02

From each table we can calculate the expected number of deaths in each group, just as in a Chi-square test of independence. The log-rank statistic combines information from all tables and all samples into an overall squared difference that measures the similarity between the two survival curves. If the true survival curves are identical, the observed log-rank statistic has an approximate Chi-square distribution with $k-1$ degrees of freedom, where k is the number of groups. The null hypothesis that the groups have the same survival curve can be tested by comparing the observed log-rank statistic to a critical value from the appropriate Chi-square distribution and rejecting the hypothesis if the observed value is too large. For the Litchfield data, the observed log-rank statistic is 62.7 with one degree-of-freedom (Table 5). This is extremely significant and confirms our initial impression that the two survival curves are different.

The Gehan-Wilcoxon test tests the same hypothesis but differs in some mathematical details. One practical difference is that the Gehan-Wilcoxon test is more sensitive to differences at earlier survival times. The log-rank test places more emphasis on differences at later survival times, so the numerical results of the two tests usually differ. Both tests can also be viewed as survival data analogs of familiar nonparametric tests. For example, the Gehan-Wilcoxon test is the censored data analog of the Kruskal-Wallis test.[4] Practically, the choice of log-rank or Gehan-Wilcoxon test makes little difference in the interpretation of any of the three data sets considered here (Table 5). In each study, the survival curves for different groupings of the data are significantly different from each other except for some sex-specific effects of PGI-2 genotype in the mosquitofish data (Table 5).

Both the log-rank and Gehan-Wilcoxon tests can be used to test whether some factor modifies the effect of a toxicant. However, they require that the data be grouped (e.g., treated with streptomycin/not treated), and they do not estimate the size of that effect. They test only whether the two (or more) curves are different. Many factors that might modify toxicant effects are naturally grouped (e.g., sex or species), but many more are continuous (e.g., size of animals, dose of toxicant, and aspects of water chemistry). Such continuous covariates may be artificially grouped, as was done in Table 5 for mosquito fish size. However, more detailed statistical models can be used to test whether these factors have any effect on the influence of a toxicant or to estimate the size of that effect.

MODELS FOR SURVIVAL TIMES

A second approach to testing the equality of survival distributions is to find a statistical model to describe the data, use that model to estimate the differences between the survivor functions, and then assess whether the differences are large enough to be statistically significant. Two models are in common use: the proportional hazards model [Equation (6)] and the accelerated failure time model [Equation (7)].

$$h(t,x_i) = e^{f(x_i)}h_0(t) \tag{6}$$

$$\log t_i = f(x_i) + \epsilon_i \tag{7}$$

The proportional hazards model describes the effect of a particular treatment by its influence on the hazard, the probability that a surviving individual will die during a small interval of time [Equation (6)]. If a reference group (e.g., control animals) has a baseline hazard function $h_0(t)$, then the hazard of another group is some multiple of the baseline hazard. The function $e^{f(x)}$ describes how the treatment (x) determines the multiplier. If $e^{f(x)}$ equals 2 for a particular group, then the hazard for an individual in that group is twice the baseline hazard. Some choices for $f(x)$ are described below.

The accelerated life model [Equation (7)] describes differences between individuals as effects on the distribution of times-to-death, rather than effects on the hazard. The differences between groups, $f(x)$, can take any of the same forms as in the proportional hazards model, but, in the accelerated time model, $f(x)$ acts on the log of the time-to-death. The accelerated life model [Equation (7)] can be converted into a model for the hazard [Equation (8)] that is slightly different from the proportional hazards model in that the effect of the covariate appears both in the multiplier of the hazard and inside the baseline hazard function.

$$h(t,x_i) = e^{f(x_i)}h_0(te^{f(x_i)}) \tag{8}$$

This parametric approach requires that we specify the two parts of the statistical model: the function $f(x)$, which describes how the groups differ from one another, and the error distribution, which describes the variability among individuals in a group. The function $f(x)$ is chosen to reflect our assumptions about how a particular covariate x influences the response of the animal. It may be any of the types of functions used in ordinary regression or analysis of variance.[19] Some possibilities are:

1. $f(X) = a + bX$ — linear response to a continously measured covariate, e.g., water temperature
2. $f(X) = a + b \log X$ — linear response to the log of a covariate, e.g., size
3. $f(X) = a + bX + cX^2$ — polynomial response to a covariate
4. $f(X) = \begin{cases} m_1 \text{ if male} \\ m_2 \text{ if female} \end{cases}$ — different mean response in each group of individuals

The types of models and the methods for choosing independent variables used in linear regression[19] can also be applied to modeling of survival times.

In typical regression and analysis of variance applications, the error distribution is assumed to be a normal distribution. However, a normal distribution is inappropriate for most time-to-death data because they are not symmetrical around the mean. For example, the histogram of times-to-death of mosquito fish in an acute arsenic exposure experiment has a long right tail (Figure 3). Although many fish die in the first 40 h, some are still dying between 80 and 100 h, and many are still alive at 102 h when the experiment was terminated. Other statistical distributions that may be better descriptions of the distribution of times-to-death include the exponential, Weibull, log-normal, and log-logistic distributions. Each

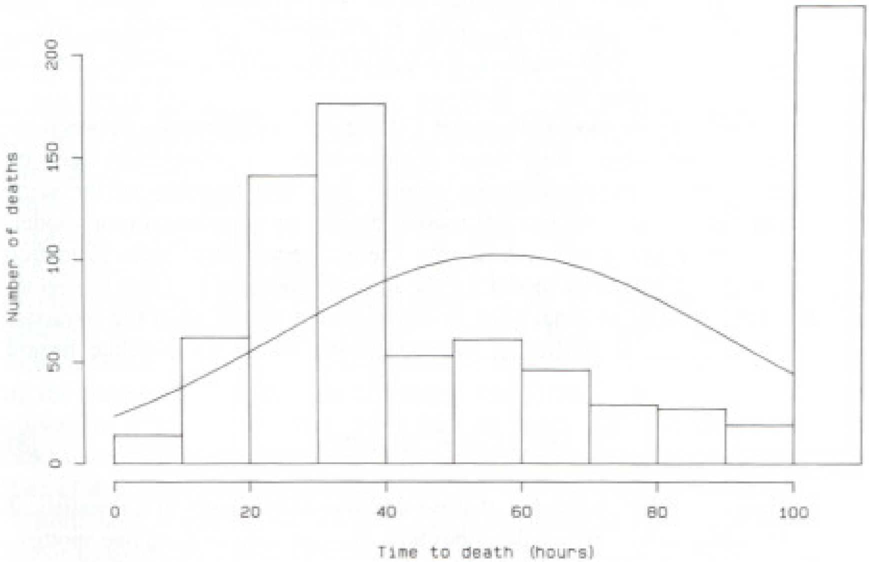

FIGURE 3. Times to death of mosquitofish in an acute arsenic exposure trial. The experiment was terminated at hour 102. The poor fit of the normal distribution (solid curve) is apparent. Data from Newman et al.[14]

Table 6
Survivor Function, Density Function, and Transformations to Linearity for Some Common Distributions of Time to Death[a]

			Transformation	
Distribution	**Survivor Function**	**Hazard Function**	**Y-axis**	**X-axis**
Exponential	$e^{-\alpha t}$	α	$\log S(t)$	t
Weibull	$e^{-(\alpha t)^{\beta}}$	$\beta\alpha(\alpha t)^{\beta-1}$	$\log[-\log S(t)]$	$\log t$
Log-normal	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\dfrac{e^{-(\log t-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\left[1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right]}$	$\text{Probit}[1-S(t)]$	$\log t$
Log-logistic	$\dfrac{1}{1+(t\alpha)^{\beta}}$	$\dfrac{\beta t^{\beta-1}\alpha^{\beta}}{1+(t\alpha)^{\beta}}$	$\text{Log}\,\dfrac{S(t)}{1-S(t)}$	$\log t$

[a] In the transformation, $S(t)$ is the survivor function, the observed proportion of individuals surviving to time t. Probit is the tabulated probit function. $\Phi(x)$ is the standard normal cumulative distribution function.

of these distributions has characteristic survival and hazard functions (Table 6).

The simplest of the distributions we will consider is the exponential distribution, which is characterized by a constant hazard function (Figure 4). A constant hazard means that the chance of a survivor dying is the same at all times. Radioactive decay is an example of a physical process with a constant hazard function. The density and survivor functions for the exponential distribution are negative exponential functions (Table 6). This distribution is described by one parameter, μ, the mean lifetime of an animal, which is equal to the reciprocal of the hazard. Larger values of μ correspond to lower hazards, more survivors, and fewer deaths in any time interval. The analysis of exponentially distributed data is relatively simple, and much early analysis of failure-time data was based on the exponential distribution,[3] but the assumption of a constant hazard function is appropriate for very few biological systems. For example, the hazard function from Litchfield's untreated mice (Figure 1) increases over time.

The Weibull distribution, a generalization of the exponential distribution, has a hazard function that can take a variety of shapes, not just a flat line like that of the exponential distribution. Weibull distributions are described by two positive parameters: α, the scale parameter that determines the spread and location of the values, and β, the shape parameter that determines the shape of the hazard or survivor functions. If $\beta = 1$, the Weibull reduces to the exponential distribution. If $\beta < 1$, the hazard is initially high and declines with time. If $\beta > 1$, the hazard increases with time, and the survivor function is S-shaped (Figure 5). An intuitive interpretation of the Weibull distribution and the role of β is that the Weibull describes the "weakest link" mode of failure.[3] Consider an individual to be composed of β parts, each of which has a constant hazard of failing. If the individual dies when any one of the β parts fails, then the time-to-death will fit a Weibull distribution. The Weibull distribution (and its special case, the exponential distribution) is the only distribution for which the accelerated time and the proportional hazards models are identical[4] because of the mathematical form of the hazard and survivor functions.

Although the Weibull distribution is quite flexible, its hazard function is monotonic: always increasing if $\beta > 1$ and always decreasing if $\beta < 1$. The log-normal and log-logistic distributions are similar distributions with hazard functions that may monotonically increase, monotonically decrease, or change directions over time (Figure 6). The hazard curves for both distributions are different for an accelerated time model and a proportional hazards model (Figure 7). For the proportional hazards model, the hazard of a treatment group is some constant proportion of the baseline hazard at all times. The hazard for the accelerated time model increases more quickly than the hazard under a proportional hazards mode, and then declines closer to the baseline level (Figure 7). Although the effect of the treatment is the same in the proportional hazard and accelerated failure models, the median time of death and the distributions of times-to-death will be different.

The choice of error distribution may or may not affect the estimation of treatment effects. For example, the shapes of the log-normal and log-logistic distributions are similar, so the substitution of one for the other usually makes little difference. However, changing the error distribution from log-normal to Weibull may substantially change the parameter estimates.

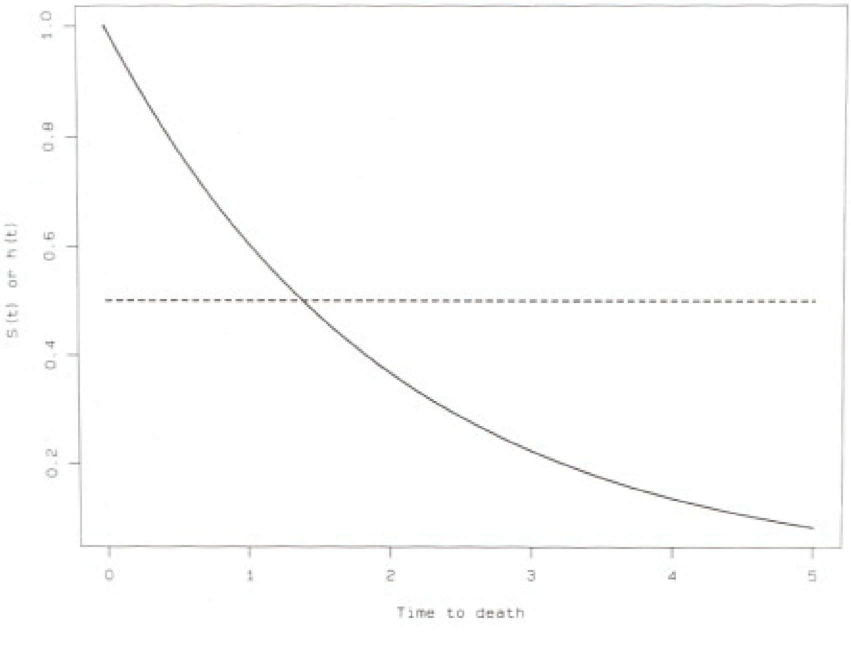

FIGURE 4. Survival (——) and hazard (- - -) functions for a exponential distribution (α = 0.5).

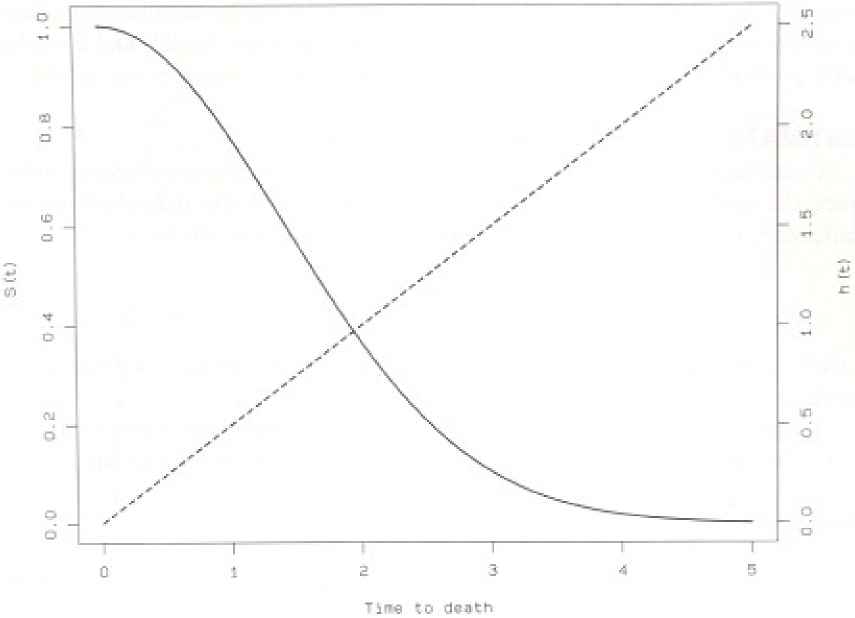

FIGURE 5. Survival (——) and hazard (- - -) functions for a Weibull distribution (α = 0.5, β = 2).

Two techniques can help in choosing an appropriate error distribution. The first is to compare maximum log-likelihoods for different models. These numbers are calculated by the common technique of fitting a model (see next section for details). For each model, the maximum value of the log-likelihood function is an index of relative fit. Larger values indicate better fit. For example, if we use a Weibull, log-normal, and log-logistic distribution to fit the mice data with MODEL DAY*CENSOR(1) = TRT, the maximized log-likelihoods are 9.74, 7.21, and 5.84, respectively. The Weibull fits better than the other two distributions because its log-likelihood is the largest. However, comparison of log-likelihoods does not show that the Weibull is a good fit. Plotting cumulative hazards, the second technique, can show that a particular distribution fits the data (see Miller[7], pp. 164-166 for details). In a cumulative hazard plot, a transformation of the observed survival is plotted against the time-to-death or the log of time-to-death (see Table 6 for details). The plot will be a straight line if the error distribution fits the data. An example and SAS code to graph a hazard plot is given in the section on Estimation and Hypothesis Testing.

Accelerated failure time models are models for time-to-death, but they can be related back to traditional toxicological models for dose-response curves. Consider an experiment in which time-to-death is recorded for individuals exposed to different doses. A model with a linear response to dose and log-logistic errors can be transformed into a logistic dose-response model. Similarly, a model with log-normal errors can become a probit dose-response model, and a model with Weibull errors can be transformed into a Weibull dose-response model.[20]

ESTIMATION AND HYPOTHESIS TESTING

A statistical model for times-to-death specifies how various covariates influence the median time-to-death, but it does not specify the magnitude of the influence. For example, a model for the Litchfield data might be

$$\log t_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad (9)$$

where μ measures the average longevity in the control treatment the treatment effect,

τ_i, measures the change in longevity caused by streptomycin treatment

ϵ_{ij}, the errors, measure the differences among individuals in a group

Two models for the Shepard data might be

$$\log t_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad (10)$$

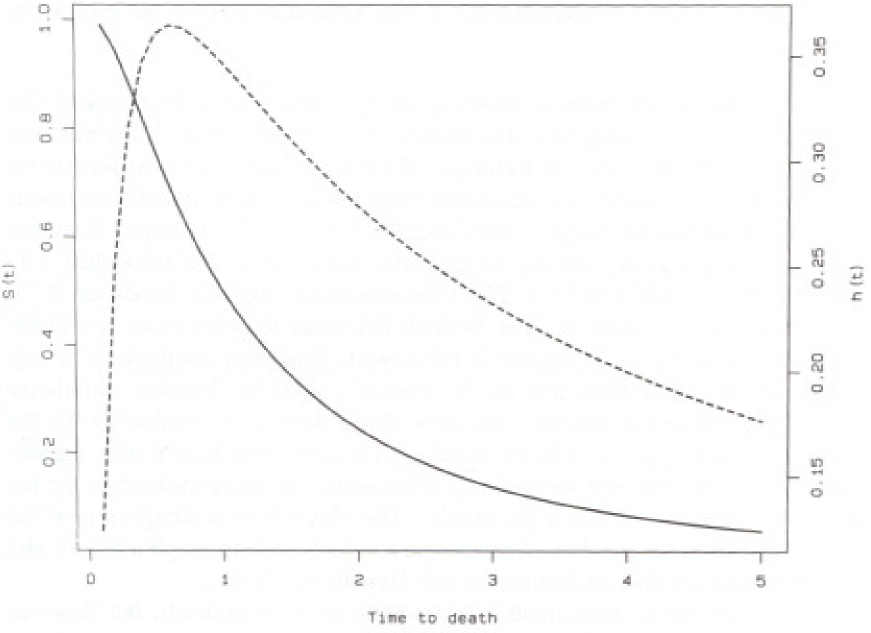

FIGURE 6. Survival (——) and hazard (- - -) functions for a log-normal distribution (μ = 40, σ = 20).

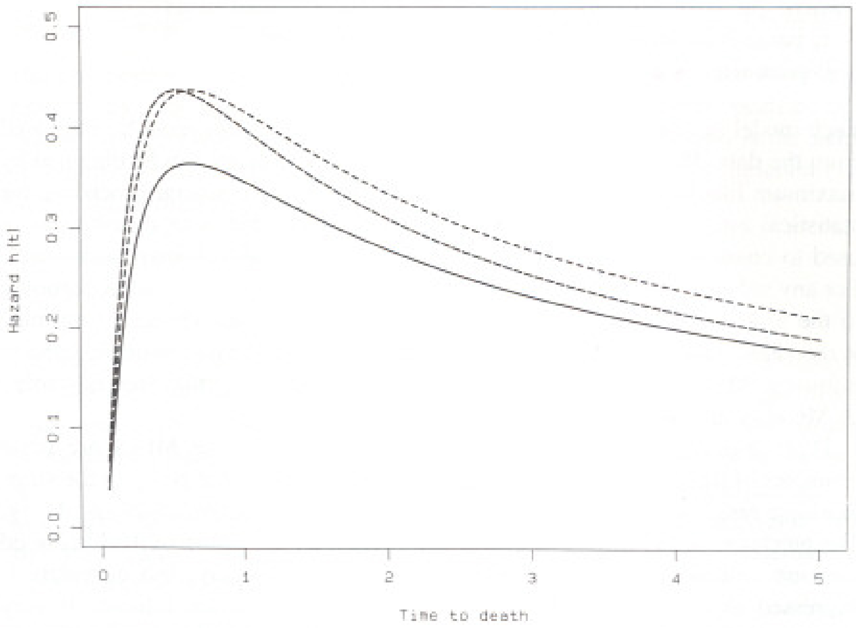

FIGURE 7. Hazard functions for the log-normal distribution: (A, ——) a baseline hazard, (B, —— ——) an accelerated time model, and (C, - - - - -) a proportional hazard models. The hazard is increased by 0.2 units in (B); the median survival time is increased by 0.2 units in (C).

where the treatment effect, τ_i, measures the difference in longevity between a reference group and the group exposed to dose i
μ measures longevity in the reference group,
ϵ_{ij} measures the differences among individuals in a group

This model specifies that the groups exposed to various doses differ in their longevity but does not specify any dose-response relationship between dose and longevity. An alternate model that specifies a linear dose-response curve is

$$\log t_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij} \tag{11}$$

Here, the dose effect, β, measures the slope of the relationship between the dose (x_{ij}) and the log-transformed time to death. A slope of 0.69 means that the median time-to-death doubles when the dose increases by 1 unit (natural log of 2 = 0.69). The intercept, α, measures the longevity for individuals with a dose of 0. Discrete and continuous parameters may be combined, as in the following model for the mosquito fish data

$$\log t_{ijk} = \mu + \alpha_i + \tau_k + \beta x_{ij} + \epsilon_{ijk} \tag{12}$$

where the α_i parameter measures the difference between sexes
τ_k parameters measure differences among genotypes
β parameter is a slope measuring the effect of size

Each model includes parameters (μ, α_i, β, τ_k) whose values must be estimated from the data. If the error distribution is specified, estimates can be obtained by maximum likelihood.[2] Briefly, maximum likelihood is a general procedure for statistical estimation.[21] The specified model, including the error distribution, is used to construct a likelihood function, a function of the unknown parameters. For any value of the unknown parameters, the likelihood function is proportional to the probability of observing the data. Intuitively, a good choice of estimate is the value which maximizes the likelihood. These are the maximum likelihood estimates (MLEs), which have many desirable statistical properties (see Edwards[21] or Mood et al.[11] for further details).

Three aspects of the likelihood function are useful. The MLEs are point estimates of the parameter, providing values for the mean of a group or the slope of a dose-response relationship, but they contain no information about variability. The precision of an estimate can be obtained from the curvature of the likelihood function around the maximum likelihood estimates. Usually, this curvature is expressed as the asympototic variance of an estimate. If the estimate is very precise, its asymptotic variance will be small, and the likelihood will decrease quickly as one considers estimates slightly different from the maximum likelihood estimate. Conversely, if the estimate is very poorly known, its asymptotic variance will be large, and estimates slightly different from the MLE will be almost as good. Finally, the log of the value of the likelihood function, calculated at the MLEs, provides a measure of fit that can be used to compare different models. This measure of fit is analogous to the model sum of squares used in regression analyses.

The likelihood function can be written down easily for any of the distributions considered here but finding its maximum is not as simple. Closed-form analytical expressions for the MLEs are available for an exponential error distribution but are not available for other distributions. Estimates for other error distributions have to be found by numerical iteration. SAS PROC LIFEREG[5] can be used to fit a wide variety of accelerated failure-time models using exponential, Weibull, log-normal, and other error distributions. If the exponential or Weibull distribution is used, then PROC LIFEREG is also fitting a proportional hazard model.

Details of the syntax of PROC LIFEREG can be found in the relevant SAS manual, but briefly: The MODEL statement specifies the desired model, including the response variable, a censoring indicator, the desired covariates, and the appropriate error distribution. Its syntax is

```
MODEL ttd*flag(number) = covariates/D = distribution;
```

where ttd is the name of the variable containing the times-to-death (or times-to-censoring) flag is a variable that identified whether that observation is a death or censored

The observation is treated as censored when the value of flag is the number in parentheses. The list of covariates to be included in the model is specified as a list of variable names to the right of the equals sign. Distribution is the name of the error distribution. For example, the model statement to fit Equation (12) with a log-normal distribution to the mosquitofish data is

```
MODEL TTD*CENSOR(1)
= SEX LOGSIZE GPI2/D = LNORMAL;
```

The variable CENSOR was created in a DATA step. CENSOR has the value of 0 if the fish died and 1 if the fish was still alive at 102 h. Multiple model statements can be included in one PROC. By default, *SAS* will treat all covariates as continuous linear variables. To declare a variable as a classification variable includes it in a CLASS statement. The following SAS code fits Equation (12) to a set of data. The CLASS statement is used to declare that SEX and GPI2 are classification variables. LOGSIZE, omitted from the CLASS statement, is a continuous variable (log size).

```
proc lifereg data = arsenic.all;
   class sex gpi2;
   model ttd * censor(1) = sex logsize gpi2/d = lnormal;
```

The following example of output from PROC LIFEREG includes a summary of the classification variables (if a CLASS statement was used), the dependent and censoring variables, counts of the number of noncensored and censored observations, and the error distribution. The maximized log-likelihood for the model is printed after the summary of variables. The parameter estimates, their standard errors, Chi-square statistics, and their tail probabilities are printed on a separate page. For each parameter, a 1 d.f. Chi-square statistic tests whether that parameter equals 0. If a CLASS statement is used, a (k − 1) d.f. Chi-square statistic tests whether any of the k level of the classification variable differs from the other levels.

```
LIFEREG Procedure
Class     Levels     Values
------------------------------------------
SEX         2        female     male
GPI2        6        100/100    100/66 100/38
                      66/66      66/38 38/38

                        Number of observations used = 751

Data Set              = WORK.ALL
Dependent Variable    = Log (TTD)
Censoring Variable    = FLAG
Censoring Value(s)    = 1
Noncensored Values    = 626 Right Censored Values    = 125
Left Censored Values  = 0 Interval Censored Values   = 0
```

Log Likelihood for LNORMAL −808.9266927

Variable	d.f.	Estimate	SE	Chi-Square	Pr > Chi	Label/Value
INTERCEPT	1	3.67880488	0.195743	353.2178	0.0001	Intercept
SEX	1			12.93713	0.0003	
	1	0.22897392	0.06366	12.93713	0.0003	female
	0	0	0	.	.	male
LOGSIZE	1	0.19420818	0.039026	24.76389	0.0001	
GPI2	5			16.76149	0.0050	
	1	0.33847937	0.180507	3.51623	0.0608	100/100
	1	0.30274341	0.178672	2.871011	0.0902	100/66
	1	0.05067728	0.185455	0.07467	0.7847	100/38
	1	0.18769681	0.186692	1.010792	0.3147	66/66
	1	0.15954612	0.190794	0.699264	0.4030	66/38
	0	0	0	.	.	38/38
SCALE	1	0.70432209	0.020843		Normal scale parameter	

INTERPRETING PARAMETER ESTIMATES

The estimates and their standard errors calculated by PROC LIFEREG can be interpreted in three different ways: as hypothesis tests, as shifts in median time-to-death, and as relative risks. As a hypothesis test, they are used to answer the question: does factor *X* have any influence on when individuals die? If it does, then changing the level of factor *X* will change the median times-to death. If *X* is a classification variable, then we can consider testing the hypothesis (H_0): Different levels of *X* have the same median time-to-death, against the alternate (H_1): At least one level of *X* has a different median time-to death. An approximate test of this hypothesis is calculated by SAS and presented as a $(k-1)$ degree-of-freedom Chi-square test associated with each factor *X*, where *k* is the number of levels of factor *X*. If *X* is a continuous variable, then the 1 d.f. Chi-square statistic tests the hypothesis that the slope of the relationship between *X* and the log time-to-death is zero. For example, using the previous SAS output, we find that the two sexes are significantly different in their time-to-death ($X^2 = 12.9$, $P = 0.0003$); size significantly affects time-to-death ($X^2 = 24.8$, $P = 0.0001$); and at least one GPI-2 genotype is significantly different from the rest ($X^2 = 16.8$, $P = 0.005$).

The tests calculated by SAS for each factor are approximate, because they assume that each estimate is normally distributed. Technically, they are Wald tests.[22] An alternate test for small samples is the likelihood ratio test,[22] which can be calculated using the output from two runs of PROC LIFEREG. To test whether factor *X* influences time-to-death, fit a model including factor *X*. Then, fit a model without factor *X*. If *X* has no influence on time-to-death, then the model without *X* will fit as or almost as well as the model with *X*. Conversely, if *X* has a large effect on time-to-death, then removing it from the model will increase the lack of fit. The lack of fit of a model is quantified by the log-likelihood statistic calculated by PROC LIFEREG for each model. A better fitting model has a larger (less negative) log-likelihood value. If *X* has no effect, then twice the difference between the two log-likelihood values is approximately distributed as a Chi-square random variable.[22] As in the Wald test, the degrees-of-freedom are $(k-1)$ if X is a classification variable with k levels and 1 if X is continuous.

The estimates themselves can be used to calculate the shift in median time-to-death. Remember the statistical model fit to the data is

$$\log t_{ij} = \mu + \beta X_{ij} + \epsilon_{ij} \tag{13}$$

If X is continuous, then a slope (β) of 0.5 means that the predicted log-transformed time-to-death increases 0.5 units for every increase of 1 unit in X. Transforming back from the log scale, a slope of 0.5 means that the predicted median time-to-death increases to $e^{0.5} = 165\%$ of the original baseline value. A slope of 0, indicating no effect, leads to a predicted median time-to-death of $e^0 = 1 = 100\%$ of the baseline value. A slope less than 0 means that the predicted time-to-death decreases as X increases. If X is a classification variable, then one of the groups is used as a reference group and the change in time-to-death is relative to that group. By default, SAS uses the largest value of the classification variable as the reference group. In the example given earlier, SEX and GPI2 are class variables. Males and the 38/38 genotype were used by SAS as the reference groups for SEX and GPI2, respectively. The estimate for females (0.229) is the difference between females and males, and the estimate for each GPI2 genotype is the different between that genotype and the 38/38 genotype.

If the model is a proportional hazards model, then estimates can be transformed into relative risks. Relative risk is the ratio of the probability of dying if an individual is in group X to the probability of dying if an individual is in the reference group. Alternatively, relative risk is the ratio of the hazard for individuals in group X to the baseline hazard. It can be calculated using the estimate from an accelerated failure time model by equation (14).

$$\text{Relative Risk} = \begin{cases} e\ -\tau_i\ /\sigma & \text{for classification effects} \\ e\ -\beta_i \Delta x/\sigma & \text{for continuous effects} \end{cases} \tag{14}$$

where τ_i and β_i are estimates of treatment differences and slopes, respectively
σ is the estimated scale parameter from the SAS output
ΔX is the desired difference between two values of a continuous variable

For the model and estimates given earlier, the relative risk of females (Relative to males of the same size and genotype) is $e^{-0.229/0.704} = 0.722$. The relative risk of an 0.15-g individual (logsize $= -1.90$), relative to an 0.10-g individual (logsize $= -2.30$) is $e^{-0.194[-1.90-(-2.30)]} = 0.92$. Relative risks greater than 1 mean that, at any point in time, individuals in group X are more likely to die than are individuals in the baseline group. Relative risks less than 1 mean that individuals in group X are less likely to die. Relative risk has a clear interpretation in proportional hazards models (e.g., accelerated failure-time models with exponential or Weibull distributions) where the ratio between the two hazards is the same at all times. It is less useful for other models.

CALCULATING MEDIAN TIMES-TO-DEATH

It is often useful to calculate the predicted median time-to-death for an individual having some combination of characteristics. This can be done using SAS PROC LIFEREG. The median time-to-death and its standard error can be calculated from characteristics of the individual (X, the parameter estimates ($\hat{\beta}$, $\hat{\tau}_k$, and $\hat{\alpha}_i$), the estimated scale parameter ($\hat{\sigma}$), and the choice of error distribution [Equation (15)].

$$\text{Median TTD} = \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\tau}_k + \hat{\beta}x_{ijk} + \hat{\sigma}W_{0.5}) \tag{15}$$

$W_{0.5}$, the median of the standardized error distribution, depends on the choice of distribution. For the log-normal distribution, $W_{0.5}$ is 0. For the exponential and Weibull distributions, $W_{0.5}$ is -0.3665. These calculations can be performed by SAS (see section on Case Study for an example).

VERIFYING ASSUMPTIONS

Any statistical model makes assumptions that should be checked as part of a careful analysis. If Weibull or exponential distributions are used in an accelerated failure time model, we assume that:

1. Hazards are proportional.

Then an accelerated failure time model makes three major assumptions:

2. Baseline hazard distribution is correctly specified.
3. Response to the covariates is correctly specified.
4. Observations are independent.

Most of these assumptions can be assessed using simple graphical tools. We will present methods for verifying that the first three assumptions are appropriate. Good experimental design helps justify the last assumption that observations are independent.

The proportional hazards assumption can be checked by dividing the data into groups of similar observations and plotting a cumulative hazard curve for each group (see test in section on Censoring and below). If the data contain continuous covariates, it will be necessary to divide the covariate into a small number of groups (2-4, depending on sample size). For each group, calculate the Kaplan-Meier estimate of the survival distribution, then plot $\log[-\log S(t)]$ against t or $\log t$. If the hazards are proportional, the curves will be parallel (Figure 8). If the sample sizes are small, the variability in each curve is high, and the eye can easily see nonparallel lines, even if the hazards are proportional. It is often helpful to compute the approximate 95% confidence interval around each survival estimate, then log-log transform the upper and lower bounds for the confidence interval to help judge whether the lines are parallel.

Hazard plots can help assess the choice of error distribution (Miller,[7] pp. 164-166). If the error distribution is Weibull (or exponential), the plot of log[−log *S*(*t*)] vs log *t* will be a straight line (see Figure 8). Other transformations (Table 6) can be used to evaluate log-normal or log-logistic distributions. SAS will calculate and plot hazard curves in PROC LIFETEST, but one must do additional calculations to plot the confidence bounds. One can use PROC LIFETEST to estimate *S*(*t*) and its standard error with PROC LIFETEST, use a DATA step to calculate and transform the confidence interval, and plot the curves (see the following code). For small samples (e.g., 50 individuals per group), the variation in the curves may mask a truly straight line. To continue the analysis of the mice data (started in the censoring section) the following SAS code plots hazard curves for treated and untreated mice.

```
/*Continuation of Analysis of Litchfield data to plot hazard curves */
data curves2;
 set curves;
 log_ttd = log(day);
 lls = log(−log(survival));
 lower = log(−log(sdf_lcl));
 upper = log(−log(sdf_ucl));

plot plot;
 plot lower*log_ttd = “−” upper*log_ttd = “+” lls*log_ttd = trt/overlay;
 title ‘Log(−Log S(t)) vs Log TTD for each group’;
 title2 ‘with confidence interval’;
```

The form of response to covariates can be tested graphically or by fitting augmented models. Consider what might be a reasonable way to improve the functional form and see if that augmented model actually does fit the data better. If the original model includes a linear response to a continuous covariate, a possible augmented model is a quadratic, or perhaps a cubic, equation. If the model includes two continuous covariates, a possible augmented model includes the cross-product of the two variables. If the model includes two classification variables, a possible augmented model may include the interaction between the two variables. The augmented model will always fit the data better because we are using more variables. The relevant statistical question is whether the improvement in the fit is significant. The log-likelihoods reported by PROC LIFEREG for each model can be used to construct a likelihood ratio test[21] for the significance of the improvement.

We will test whether a linear dose-response model is adequate to describe the mortality patterns of trout exposed to low oxygen. In the subsequent SAS code (following the next paragraph), the first MODEL statement, labeled ‘‘Linear,’’ fits a model with a linear response to oxygen level. The second, labeled ‘‘Quad,’’ augments the linear model to include a quadratic response. The third, labeled ‘‘Loglin,’’ fits a model with a linear response to log-transformed oxygen level. The last, labeled ‘‘Groups,’’ fits a model with a different mean for each group.

If any of the first three models is appropriate, then that model will fit almost as well as the groups model. These calculations are repeated for just the intermediate group of oxygen levels in the second PROC LIFEREG step. Log-normal error distributions were chosen because plots of log [− log $S(t)$] vs log t were not straight lines.

Statistical tests of the improvements in fit can be constructed from the SAS output (Table 7). If the groups model fits just as well as a linear model, then twice the difference in likelihoods has a Chi-square distribution. The degrees-of-freedom for the Chi-square is the difference of the number of parameters in each model. For the complete data the Chi-square statistic to test the fit of the linear or loglin models has 8 d.f. because the ''Groups'' model has 11 parameters (10 means and 1 scale parameter), while the linear models each have 3 parameters (1 intercept, 1 slope, and 1 scale parameter). If twice the observed difference in log-likelihoods exceeds a critical Chi-square statistic, once concludes that the full model fits significantly better. For these data, the linear model is not sufficient to describe the response to all doses, but it is adequate for the intermediate doses (Table 7). This approach cannot be used to test between a linear response and a log-linear response because both models have 3 parameters. Technically, the

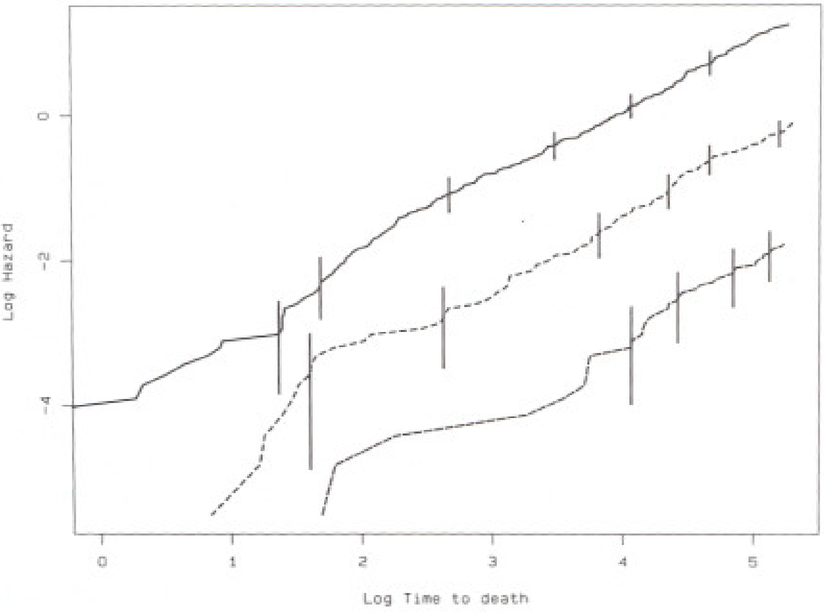

FIGURE 8. Hazard plots for simulated data. Data were generated from a proportional hazards model with Weibull errors. Vertical bars indicate 95% confidence intervals for *S(t)* at selected times. The relative risk of group 2 (- - -), relative to group 1 (——— ———), was 4.48. The relative risk of group 3 (————), relative to group 1, was 20.

Table 7
Summary of Log-Likelihood and Tests of Fit of Linear Models for Trout Data

	Log-Likelihood (d.f. in model)	
Model	**All Doses**	**Intermediate Doses**
Groups	−112.48(11)	−67.73(7)
Linear response	−132.82(3)	−68.14(3)
Quadratic response	−115.45(4)	−68.06(4)
Log-linear response	−146.13(3)	−68.71(3)
	−2Δ Log-Likelihood (d.f.,*p*)	
Test of Fit	**All Doses**	**Intermediate Doses**
Linear response vs groups	40.68(8, 0.0001)	0.82(4, 0.94)
Quadratic response vs groups	5.94(7, 0.55)	0.66(3, 0.69)
Linear vs quadratic response	34.74(1, 0.0001)	0.16(1, 0.88)
Log-linear response vs groups	67.30(8, 0.0001)	1.96(4, 0.74)

likelihood ratio test is valid when one model is nested in the full model.[21] Both the linear and log-linear models are nested in the groups model, but the linear model is not nested in the log-linear model.

```
data shepard;
 infile cards missover;
 input o ttd @;
 log_o = log(o);
 o_group = o;
 o2 = o * o;
 do until (ttd =.);
   if ttd = 5000 then censor = 1;
     else censor = 0;
  output;
  input ttd @;
  end;
cards;
  0.77  17   18   20   20   21   21   22   22   23
  0.94  20   22   24   26   26   29   29   31   34     34   41
  1.10  25   29   33   33   37   37   37   41   48   55
  1.16  30   30   35   35   40   45   50   58   62   70   100
  1.36  48   52   60   85   140   160   170   190   250
  1.43  50   50   135   175   195   215   355   405   465   600
  1.55  165   165   195   270   440   440   735   865   1400
  1.69  195   225   270   270   440   675   995   1150   1150   5000   5000
  1.77  240   675   995   2080   5000   5000   5000   5000   5000   5000
  1.86  400   5000   5000   5000   5000   5000   5000   5000   5000   5000

proc lifereg;
  title 'all groups';
  class o_group;
  Linear: model ttd*censor(1) = o/d = lnormal;
  Quad: model ttd*censor(1) = o o2/d = lnormal;
  Loglin: model ttd*censor(1) = log_o/d = lnormal;
  Groups: model ttd*censor(1) = o_group/d = lnormal;
```

```
proc lifereg;
 title 'intermediate groups';
 where o between 0.95 and 1.76;
 class o_group;
 Linear: model ttd*censor(1) = o/d = lnormal;
 Quad: model ttd*censor(1) = o o2/d = lnormal;
 Loglin: model ttd*censor(1) = log_o/d = lnormal;
 Groups: model ttd*censor(1) = o_group/d = lnormal;
```

Unlike SAS PROC GLM, PROC LIFEREG will not automatically generate the quadratic, cross-product, or interaction terms. To fit such models, one should create a new variable for each new term in the model. For example, to test for possible interaction between two classification variables, we can create a new variable containing a unique value for each combination of the two original variables. An easy way to do this in SAS is to use the string concatenation operator, ||, if the two original variables are character variables. The importance of a quadratic or cross-product term can be tested by including new variables containing the square of the original variable or the product of two original variables, respectively.

GROUPED TIMES-TO-DEATH

All of the preceding analyses have assumed that the exact time-to-death was recorded; hence, time-to-death is a continuous variable. As in any measurement, the fineness of the measuring scale imposes some discreteness on what theoretically could be a continuous variable. In the mosquitofish data, dead fish were collected every 3 h, so that time-to-death is recorded as 9 h, or 12 h, but never as 10.3 h. Even so, time-to-death may be considered continuous in these data because 3 h is short relative to the 102-h duration of the experiment.

What if deaths were recorded weekly in a 12-week experiment (e.g., Quattro and Vrijenhoek[23])? Here the interval between measurements is a sizable fraction of the duration. An animal recorded as a death at week 3 actually died between the second and third weeks. Although we do not know the exact time-to-death, we know that it falls within a certain interval. PROC LIFEREG will fit parametric survival models to such interval censored data (see following code). Animals still alive at the end of the experiment are right censored, just as before. Parameter interpretations are the same as those for exact data.

```
/* SAS code to fit interval censored model to discrete times to death          */

/* experiment terminated at week 12,                                           */
/* survivors to week 12 are coded in the raw data as ttd = 13 and are          */
/* right censored,                                                             */
/* observed deaths occurred between observed ttd and previous week             */

data interval;
 input trt ttd;

/* If the animal was alive at the end of the experiment, code it as right censored at week
12                                                                             */
```

```
if ttd > 12 then do;
  lower = 12;
  upper = .;
  end;

/* But if it died during the experiment, code it as interval censored between the previous
week and this week                                                                      */

  else do;
  lower = ttd - 1;
  upper = ttd;
  end;

proc lifereg;
  class trt;
  model (lower,upper) = trt;
```

CASE STUDY: THE INFLUENCE OF SEX, SIZE, AND GENOTYPE ON ARSENIC INTOXICATION

As part of a larger study of metal tolerance in mosquitofish, we examined the roles of sex, size, and genotype as modifying factors affecting As intoxication. A summary of the data collection protocol is provided in Data Sets; further details are given in Newman et al.[14] The primary concern in this study was whether fish with particular genotypes were more sensitive than other genotypes to As intoxication. To increase the precision of comparisons between genotypes, it was necessary to control potential variation due to the effects of size and sex. We analyzed the effects of different genotypes at eight enzyme loci, but this analysis will be restricted to the effects at one locus, GPI-2. The questions we wish to answer are:

1. Are males and females equally sensitive to the effects of As?
2. Are larger individuals of either sex more resistant to As?
3. Does the genotype at the GPI-2 locus influence survival?

The first step in the analysis is to group individuals into size classes and calculate survival curves for each sex, size, and genotype class. Hazard plots for each sex (Figure 9a) are linear until approximately 42 h (log t = 3.75), when the slopes decrease, but are not parallel. Similar patterns are found among different sizes classes of male or female fish and among different genotypes (Figure 9b). A Weibull error distribution is not appropriate because the hazard plots are not linear and parallel across all the data. The Weibull would not be an appropriate distribution even if two modes of action were postulated for As, one responsible for deaths before 42 h and one after 42 h, because the lines are not parallel.

The differences between survival curves for different sexes, different size classes, and different genotypes are statistically significant by either the log rank test or the Wilcoxon test (Table 5). When each sex is considered separately, the

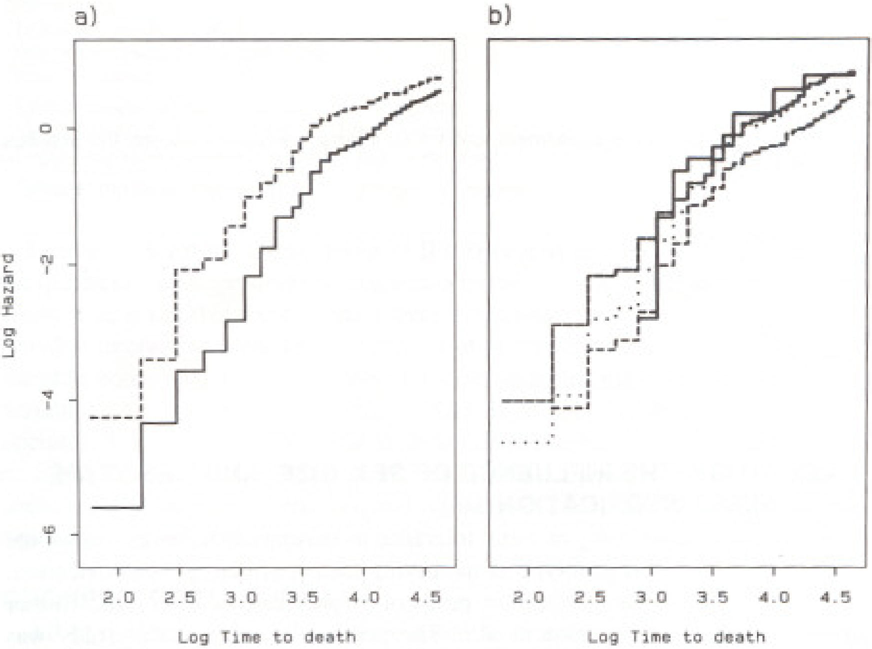

FIGURE 9. Hazard plots for mosquitofish exposed to arsenic (a) males (- - -) and females (——) and (b) four of the six genotypes at the GPI-2 locus; 38/38 (——), 66/66 (....), 100/100 (- - -), 100/66 (—— ——).

survival curves for different GPI-2 genotypes are significantly different in females (using the Wilcoxon test) but are not significantly different in males (using either test). Fewer of the experimental animals were males, so the lack of significant differences between GPI-2 genotypes may just be a result of low statistical power. Accelerated time models can be used to verify that the effects of GPI-2 genotype are different in the two sexes.

Perhaps the simplest reasonable model that includes the effects of sex, size, and genotype is Equation (12), in which sex and genotype are classification variables and some function of size is a continuous variable. The form of the response to size can be estimated from plots of log time-to-death versus various transformations of size. There is considerable scatter in these plots, but the plots of log(size) appear to be more linear than those against untransformed size. A log transformation of size is commonly a better predictor of other measures of toxic potency, such as LD_{50} (ref. 24) (see also this volume, Chapter 4). The following SAS code fits this model using both Weibull and log-normal error distributions.

```
proc lifereg data = arsenic.all;
  class sex gpi2;
  model ttd*censor(1) = sex logsize gpi2;
  model ttd*censor(1) = sex logsize gpi2/d = lnormal;
  title 'Basic model – comparison of Weibull and lognormal errors';
```

The log-likelihood of the model using the log-normal error distribution (− 808.9) is less negative than that from the Weibull distribution (− 884.8), indicating that the log-normal is the more appropriate error distribution, as expected from the hazard plots (Figures 9a, b). This model assumes that the response to size and GPI-2 genotype is the same in male and females. These assumptions can be tested by fitting more general models. The following SAS code constructs new variables and fits two models to test whether the response to size differs between males and females.

```
data sexsize;
  set arsenic.all;
  mlsize = 0;
  flsize = 0;
  if sex = 'M' then mlsize = logsize;
    else flsize = logsize;
  sexsize = flsize

proc lifereg;
  class sex gpi2;
  2sex: model ttd*censor(1) = sex mlsize flsize gpi2/d = lnormal;
  sexdiff: model ttd*censor(1) = sex logsize sexsize gpi2/d = lnormal;
```

The log-likelihood for either model is − 804.8, a statistically significant improvement from a model with the same slope for both sexes. The two models differ in how the differences between the sexes are parameterized. In the model labeled ''2sex,'' separate slopes are fit for males and females. These estimates of these slopes can be directly interpreted, but the test of whether the two slopes are different must be hand calculated from the log-likelihood statistics. The model labeled ''sexdiff'' parameterizes the slope as the slope for males (LOGSIZE) and the difference between the male and female slopes (SEXSIZE). The SEXSIZE coefficient is an estimate of the difference between the slopes; the test of whether SEXSIZE equals 0 tests the hypothesis that the slopes are equal. The log-likelihoods of the two models are the same, because they are different forms of exactly the same model.

A similar procedure could be used to construct five new variables to test the SEX*GPI2 interaction (five variables are necessary because the GPI2 classification variable has five degrees-of-freedom), but an even more general model can be used to test whether there are any other interactions. The following SAS code fits separate models to the male and female data.

```
proc sort data = arsenic.all;
  by sex;

proc lifereg;
  by sex;
  class gpi2;
  model ttd*censor(1) = logsize gpi2 /d = lnormal;
```

The log-likelihood statistics (and degrees-of-freedom) for each sex can be added together to calculate the log-likelihood and degrees-of-freedom for a model in which each parameter (intercept, slope for log of size, effect of GPI-2 genotype, and scale) is allowed to differ between the sexes. The improvement in fit for this model is small (log-likelihood of full model is −803.6) and is not statistically significant (Table 8).

Estimates and their standard errors can be obtained directly from the SAS output (Table 9). Each estimate can be interpreted as the difference in log-transformed median time-to-death caused by a 1 unit change in that variable. Sizes of male and female fish are continuous variables, so the estimates can be interpreted like slope estimates from an analysis of covariance. For example, for a male fish, each 0.2-unit increase in log(size) increases the log-transformed median time-to-death by (0.2)(0.53) = 0.106 units. For females, the same change results in an increase of (0.2)(0.157) = 0.031 units. Transformed from the log scale, an 0.2-unit increase in size leads to $e^{0.106}$ = 1.11 (or 11% larger) predicted median time-to-death in males and $e^{0.031}$ = 1.03 (or 3% larger) in females. SEX and GPI-2 are classification variables that can be interpreted like groups in analysis of variance. The estimate of the SEX effect is the difference between the intercept for male fish and the intercept for female fish. The estimates for each GPI-2 effect are the differences between that genotype and the last genotype (in this case, the 38/38 genotype). For example, the estimated effect of the 100/100 genotype is 0.35. Hence, the 100/100 genotypes have a predicted median time-to-death that is $e^{0.35}$ = 1.42 times as long as the median time-to-death for fish with the 38/38 genotype. Because the error distribution was log-normal rather than Weibull, the estimates cannot be interpreted as relative risks.

Rather than interpret estimates directly, it is possible to calculate the median times-to-death for individuals with specified characteristics and interpret them. For example, Table 10 presents characteristics of 12 individuals with different combinations of sexes, genotypes, and sizes that span realistic values. The

Table 8
Summary of Log-Likelihood Tests for Mosquito Fish Date

Model	Error Dist.	Log-Likelihood	d.f. in Model
1. Sex Log(size) GPI-2	Weibull	−884.8	9
2. Sex Log(size) GPI-2	Log-normal	−808.9	9
3. Sex Log(size) Sex*Log(size) GPI-2	Log-normal	−804.8	10
4. Both sexes: Log(size) GPI-2	Log-normal	−803.6	16
Females only: Log(size) GPI-2	Log-normal	−565.4	8
Males only: Log(size) GPI-2	Log-normal	−238.2	8

Tests of fit	−2Δ Log-likelihood	Δd.f.	$P > X^2$
Does response to size differ between sexes?			
Model 2 vs 3	8.2	1	0.0044
Do any other parameters differ between sexes?			
Model 3 vs 4	2.4	6	0.88

Table 9
Parameter Estimates and Hypothesis Tests for Model "2sex" (see Case Study section) Fit to Mosquito Fish Data

Parameter	d.f.	Estimate(SE)	X^2	$P > X^2$
Intercept	1	4.30(0.29)	216.7	0.0001
Sex	1	−0.45(0.24)	3.3	0.0678
Log(size) for males	1	0.53(0.12)	18.4	0.0001
Log(size) for females	1	0.157(0.041)	14.8	0.0001
GPI-2 (overall)	5		17.1	0.0043
GPI-2, effect of 100/100	1	0.35(0.18)		
GPI-2, effect of 100/66	1	0.31(0.18)		
GPI-2, effect of 100/38	1	0.06(0.18)		
GPI-2, effect of 66/66	1	0.18(0.18)		
GPI-2, effect of 66/38	1	0.18(0.19)		
GPI-2, effect of 38/38	0	0.00(0.00)		

Table 10
Estimated Median Times to Death for Fish with Various Combinations of Traits Using Model "2sex" (see Case Study section) Fit to Mosquito Fish Data

Sex	Size(g)	GPI-2 genotype	Median TTD (h)	SE Median
Female	0.15	100/100	49.6	3.0
Female	0.25	100/100	53.8	2.9
Female	0.35	100/100	56.7	3.1
Male	0.15	100/100	38.2	2.6
Male	0.25	100/100	50.1	4.6
Male	0.35	100/100	59.6	7.4
Female	0.30	100/100	55.4	3.0
Female	0.30	100/66	53.2	2.6
Female	0.30	100/38	41.4	2.9
Female	0.30	66/66	47.1	3.4
Female	0.30	66/38	46.7	3.8
Female	0.30	38/38	39.1	6.7

following SAS code calculates predicted median times-to-death for individuals with those characteristics.

```
data controls;
  input sex $ size gpi2 $;
  logsize = log(size);
  mlsize = 0;
  flsize = 0;
  if sex = 'M' then mlsize = logsize;
    else flsize = logsize;
  predict = 1;
  ttd = .;
  censor = .;
cards;
F 0.15 100/100
F 0.25 100/100
F O.35 100/100
M 0.15 100/100
M 0.25 100/100
M 0.35 100/100
```

```
F 0.30 100/100
F 0.30 100/66
F 0.30 100/38
F 0.30 66/66
F 0.30 66/38
F 0.30 38/38
;
data all;
 set sexsize controls;

proc lifereg data = all;
 class sex gpi2;
 model ttd*censor(1) = sex mlsize flsize gpi2 /d = lnormal;
 output out = preds control = predict p = med _ ttd std = se _ med;

proc print;
 title 'Predicted median times to death';
 var sex size gpi2 med _ ttd se _ ttd;
```

The strategy in this SAS code is to create a data set containing a combination of characteristics for each new individual. The time-to-death for the new individuals is set to the SAS missing value code, so that the new individuals are not used to estimate the parameters of the model. The predicted median time-to-death and its standard error are written to a new SAS data set by the OUTPUT command. The CONTROL = option includes only the new individuals in the output data set. The relevant parts of this new data set are then printed onto the SAS listing file.

Median times-to-death (Table 10) are computed from the estimates, so they should show the same patterns as the parameter estimates; however, we feel that the median times-to-death are more easily interpreted. Small males die more quickly than either larger males or females (Table 10), and the response to size is different between the sexes. Increasing the size of female fish from 0.15 to 0.25 increases the median time-to-death by 4.2 h (Table 10). A similar increase in size of male fish leads to a much larger increase (11.9 h) in median TTD. Genotype has a large effect on either sex fish; there is a 16.3 h difference between the median time-to-death for the most resistant genotype (100/100) and the most susceptible genotype (38/38).

COX PROPORTIONAL HAZARDS MODEL

The accelerated life models fit by SAS PROC LIFEREG assume specific forms for the baseline hazard function and error distribution. An alternative class of models, called Cox proportional hazard models,[25] makes no assumptions about the form of the baseline hazard function. Instead, they make the proportional hazards assumption, i.e., that the effect of a covariate is to multiply the hazard function by a constant that does not change over time [see Equation (6)]. Such models are commonly used in the analysis of medical data, where the assumption of proportional hazards appears to be generally acceptable.[18] As discussed before (see Verifying Assumptions), plots of the hazard function can be used to verify the appropriate choice of error distribution. If none of the distributions adequately fits the data, a Cox model might be appropriate.

The Cox model has other advantages over accelerated life models. Often, one is more interested in describing the influence of covariates on survival than in describing the baseline hazard. By using a Cox model, one can estimate the effects of covariates while ignoring the unknown baseline hazard function. Also, the Cox model is less affected by outliers, unusually large or small failure-times, because the computations use only the rank ordering of failure and censoring times.[4] Cox proportional hazard models can be fit by the supplemental SAS procedure PHGLM,[26] which is available in some implementations of SAS, and by procedures in BMDP, SYSTAT, and S-Plus (Table 1). The covariates in a Cox proportional hazard model may be any combination of continuous or classification variables, like the covariates in an accelerated failure-time mode. The one difference is that a Cox model does not include an intercept.

If a Cox model is fit to the As data, the conclusions are similar to those obtained by fitting accelerated failure-time models. The parameter estimates (Table 11) are different from those obtained from an accelerated life model with a log-normal error distribution (Table 9), but the results of the hypotheses tests are the same, except for influence of size in females. In particular, there is still evidence that certain genotypes survive longer than do others. The difference in sign and magnitude of the estimates has two causes: (1) a mathematical difference in parameterization between proportional hazards and accelerated failure-time models[4] and (2) use of the log-normal error distribution (for which an accelerated time model is not a proportional hazards model) rather than the Weibull or exponential distributions.

Estimates from the Cox model can be interpreted as relative risks (see section Interpreting Parameter Estimates), because different hazard functions are as-

Table 11
Results from Fitting a Cox Proportional Hazards Model to the Arsenic Data[a]

Parameter	d.f.	Estimate (SE)	X^2	$P > X^2$
Sex	1	0.364(0.365)	0.99	0.32
Log(size) for males	1	−0.508(0.183)	7.67	0.0056
Log(size) for females	1	−0.129(0.067)	3.73	0.053
GPI-2 Overall	5		15.04	0.010
GPI-2, 100/100		−0.59(0.27)		
GPI-2, 100/66		−0.50(0.27)		
GPI-2, 100/38		−0.18(0.28)		
GPI-2, 66/66		−0.40(0.28)		
GPI-2, 66/38		−0.28(0.29)		
GPI-2, 38/38		0.00(0.00)		

[a] The model included the same covariates used in the model 2sex in Case Study section. Computations done with the COXREG and COXREG.PRINT functions in S-Plus. Overall test of GPI-2 genotypes computed from output of the COXREG function.

sumed to be proportional, so the relative risk is constant over time. For estimates from Cox models, the relative risk is computed using Equation (16).

$$\text{Relative Risk} = \begin{cases} e\ \mathrm{t}_i & \text{for classification effects} \\ e^{\beta_i \Delta x} & \text{for continuous effects} \end{cases} \qquad (16)$$

For example, the estimate of the effect of the 100/100 genotype at the GPI-2 locus is -0.586, so the relative risk for an individual with that genotype is $e^{-0.586} = 0.557$. In other words, such an individual is more likely to survive than is a reference individual.

Unlike the accelerated time models, the Cox model does not directly provide estimates of the median time-to-death. However, the entire survival curve (including the median) can be described if the baseline survival function is estimated. An estimate of the baseline survival curve is usually available in the output of the program that fits the Cox model.

The major assumption in a Cox model is that the baseline survival function is the same for all individuals. This assumption can be relaxed by stratifying the data into groups and estimating a separate baseline survival curve for each group. This is not quite the same as a separate analysis for each group; in a stratified analysis, a covariate is assumed to have the same effect in each stratum. Any potential covariate (e.g., GPI-2 genotype in the As data set), can be modelled it either as a covariate or as a variable that defines strata. When included as a covariate, its effect is estimated and tested, but the hazards are assumed to be proportional. When included as strata, the effects are not estimated, but the hazards do not have to be proportional.

Including stratification variables provides a way to check the proportional hazards assumption.[4] Calculate the baseline survival curve for each stratum, then plot log [$-$ log $S(t)$] against log t for each stratum. Just as in the hazard plots seen previously (in Verifying Assumptions), parallel lines on the plot suggest that the hazards are proportional. If a Weibull or exponential error distribution is appropriate, then the lines will also be straight. To check whether the proportional hazards assumption is appropriate for the effects of the GPI-2 genotype in the As data, a Cox model was fit using sex, male size, and female size as covariates and GPI-2 genotypes as strata. The hazard plots for each genotype are essentially parallel. Hence, we can assume that hazard functions for each genotype are proportional.

SUMMARY

There is a tendency for environmental toxicologists to uncritically select routine toxicity testing protocols in their research efforts. This can be unfortunate, as statistical techniques are available that provide considerably more abundant and precise information than do the standard techniques. One such class of techniques that uses time-to-death information to construct survival models is described.

A general review of survival analysis is provided using three data sets of increasing complexity. Modeling survival times with proportional hazards and accelerated failure-time models is introduced and compared briefly to standard techniques. The choice of distribution for survival analysis is discussed, with emphasis on the Weibull, exponential, and log-normal distributions. Cox proportional hazards models are described briefly. Techniques for parameter estimation and hypothesis testing are illustrated using an As toxicity data set.

These models provide a statistically powerful and conceptually easy way to assess the modifying effects of environmental conditions (e.g., water quality) or subject characteristics (e.g., fish size) in acute toxicity tests.

ACKNOWLEDGMENTS

The authors are grateful for comments and suggestions made to two anonymous reviewers. J. Quattro and R. Vrijenhock provided a fitness data set for analysis. Financial support was obtained from contract DE-AC09-76SROO-819 between the U.S. Department of Energy and the University of Georgia.

REFERENCES

1. American Public Health Association. *Standard Methods for the Examination of Water and Wastewater*, 15th ed. (Washington, DC: American Public Health Association, 1981), p. 1134.
2. Cox, D. R. and D. Oakes. *Analysis of Survival Data* (London: Chapman and Hall, 1984), 201 pp.
3. Nelson, W. *Applied Life Data Analysis* (New York: John Wiley & Sons, 1982), p. 634.
4. Kalbfleisch, J. D. and R. L. Prentice. *The Statistical Analysis of Failure Time Data* (New York: John Wiley & Sons, 1980), p.321.
5. SAS Institute Inc. *SAS/Stat User's Guide, Release 6.03 Edition* (Cary, NC:SAS Institute Inc, 1988), p. 1028.
6. SAS Institute Inc. *SAS Technical Report P-179, Additional SAS/Stat Procedures, Release 6.03 Edition* (Cary, NC:SAS Institute Inc, 1988), p. 255.
7. Miller, R. G. Jr. *Survival Analysis* (New York: John Wiley & Sons, 1981), p. 238.
8. Meeker, W. Q. and G. J. Hahn. *How to Plan an Accelerated Life Test — Some Practical Guidelines*. (Milwaukee: ASQC, 1985), p. 36.
9. Nelson, W. *How to Analyze Reliability Data* (Milwaukee: ASQC, 1983), p. 54.
10. Sprague, J. B. "Factors that Modify Toxicity," Fundamentals of Aquatic Toxicology, G. M. Rand, and S. R. Petrocelli, Eds., (Washington: Hemisphere Publ. Co., 1985), pp. 124-163.
11. Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics, 3rd ed.* (New York: McGraw Hill, 1974), p. 564.
12. Litchfield, J. T. Jr. "A Method for Rapid Graphic Solution of Time-Percent Effort Curves," *J. Pharmacol. Exp. Ther.* 97:399-408 (1949).
13. Shepard, M. P. "Resistance and Tolerance of Young Speckled Trout *(Salvelinus fontinalis)* to Oxygen Lack, with Special Reference to Low Oxygen Acclimation," *J. Fish Res. B. Can.* 12:387-446 (1955).

14. Newman, M. C., S. A. Diamond, M. Mulvey, and P. Dixon. ''Allozyme Genotype and Time to Death of Mosquitofish, *Gambusia affinis* (Baird and Girard) during Acute Toxicant Exposure: A Comparison of Arsenate and Inorganic Mercury,'' *Aquat. Toxicol.*, 15:141-156 (1989).
15. Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. G. Smith. ''Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. II. Analysis and Examples,'' *Brit. J. Cancer* 35:1-39 (1977).
16. Gehan, E. A. ''A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples,'' *Biometrika* 52:203-223 (1965).
17. Mantel, N. ''Evaluation of Survival Data and Two New Rank Statistics Arising in its Consideration,'' *Cancer Chemother. Rep.* 50:163-170 (1966).
18. Tibshirani, R. ''A Plain Man's Guide to The Proportional Hazards Model,'' *Clin. Invest. Med.* 5:63-68 (1982).
19. Draper, N. R. and H. Smith. *Applied Regression Analysis, 2nd ed.* (New York: John Wiley & Sons, 1981), p. 709.
20. Christensen, E. R. ''Dose-Response Functions in Aquatic Toxicity Testing and the Weibull Model,'' *Water Res.* 18:213-221 (1984).
21. Edwards, A. W. F. *Likelihood* (Cambridge: Cambridge Univ. Press, 1972), p. 235.
22. Rao, C. R. *Linear Statistical Inference and Its Applications, 2nd ed.* (New York: John Wiley & Sons, (1973), p. 625.
23. Quattro, J. M. and R. C. Vrijenhoek. ''Fitness Differences among Remnant Populations of the Endangered Sonoran Topminnow,'' *Science* 245:976-978 (1989).
24. Anderson, P. D. and L. J. Weber. ''Toxic Response as A Quantitative Function of Body Size,'' *Toxicol. Appl. Pharmacol.* 33:471-483 (1975).
25. Cox, D. R. ''Regression Models and Life Tables,'' *J. R. Stat. Soc. B* 34:187-202, (1972).
26. SAS Institute Inc. *SUGI Supplemental Library User's Guide, Version 5 Edition* (Cary, N.C.:SAS Institute Inc., 1986), p. 437.

Metal Ecotoxicology

Concepts & Applications

Edited by

Michael C. Newman
Alan W. McIntosh

LEWIS PUBLISHERS

Library of Congress Cataloging-in-Publication Data

Metal ecotoxicology: concepts and applications/edited by Michael C.
Newman and Alan W. McIntosh
Materials developed for a special session of the International
Conference on Metals in Soils, Waters, Plants, and Animals, held in
Orlando, Fla., Apr. 30-May 3, 1990.
Includes bibliographical references and index.
ISBN 0-87371-411-3
1. Metals--Environmental aspects--Congresses. 2. Metals-
-Toxicology--Congresses. 3. Pollution--Environmental aspects-
-Congresses. I. Newman, Michael C. II. McIntosh, Alan W.
III. International Conference on Metals in Soils, Waters, Plants,
and Animals (1990: Orlando, Fla.)
QH545.M45M46 1991
574.5′222--dc20 91-13943
CIP

This book represents information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Every reasonable effort has been made to give reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

LEWIS PUBLISHERS, INC.
121 South Main Street, P.O. Drawer 519, Chelsea, Michigan
48118

PRINTED IN THE UNITED STATES OF AMERICA
1 2 3 4 5 6 7 8 9 0