[CANCER RESEARCH 53, 6042–6050, December 15, 1993]

Statistical Analysis of *in Vivo* Tumor Growth Experiments[1]

Daniel F. Heitjan,[2] Andrea Manni, and Richard J. Santen

Center for Biostatistics and Epidemiology [D. F. H.] and Department of Medicine [A. M.], Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, and Department of Medicine, Wayne State University, Detroit, Michigan 48201 [R. J. S.]

ABSTRACT

We review and compare statistical methods for the analysis of *in vivo* tumor growth experiments. The methods most commonly used are deficient in that they have either low power or misleading type I error rates. We propose a set of multivariate statistical modeling methods that correct these problems, illustrating their application with data from a study of the effect of α-difluoromethylornithine on growth of the BT-20 human breast tumor in nude mice. All the methods find significant differences between the α-difluoromethylornithine dose groups, but recommended sample sizes for a subsequent study are much smaller with the multivariate methods. We conclude that the multivariate methods are preferable and present guidelines for their use.

INTRODUCTION

The analysis of alterations in *in vivo* tumor growth is a powerful tool for studying the effects of potential cancer treatments. In a typical experiment, one randomizes tumor-bearing animals into various treatment groups, periodically observing tumor volumes. The resulting dataset consists of a series of volumes for each animal, which one analyzes to determine whether and how the treatment affects tumor growth. In this article we discuss the statistical methods that are available for analyzing such experiments.

The standard way of demonstrating treatment effects is to establish that intergroup differences are statistically significant; thus we focus on significance tests and their properties. Classical statistics evaluates tests in terms of type I error rate and power. The type I error rate is the chance of obtaining a significant result when there is no effect, and the power is the chance of obtaining a significant result when there truly is an effect. No worthwhile test can be foolproof in the sense of being always significant when there is an effect and never significant when there is none. The best one can do is to fix the performance at preselected levels, conventionally 5% for type I error and 90% for power. For a given type I error rate and actual difference, the power increases with the sample size. Thus a common method for determining the sample size is to fix the type I error rate, estimate the size of the effect (often from past data), and choose n to be just large enough for the power to exceed 90%.

Just as there can be many ways to measure a biological parameter, not all of which are equally efficient, there can be many ways to test a statistical hypothesis, not all of which are equally powerful. A more powerful method may find significance when a less powerful method does not; consequently, the minimum sample size required to achieve the desired power is smaller with a more powerful method. Thus the choice of statistical method, far from being irrelevant, can tangibly affect the efficiency of experimentation and the credibility of results.

To determine current statistical practices among *in vivo* experimenters, we surveyed two summer 1992 issues of each of seven leading journals that commonly report such studies: *Breast Cancer Research and Treatment, Cancer Research, European Journal of Cancer, International Journal of Cancer, International Journal of Radiation Oncology, The Prostate*, and *Radiation Research*. We selected articles that presented analyses of *in vivo* tumor growth data and reviewed the statistical methods used.

Our review revealed that a variety of methods are in use. Several authors (1–7) did a separate analysis of tumor volumes at each time point, indicating all the times at which differences were significant. The analysis at each time was either a t test, a Mann-Whitney test, an ANOVA[3] F test, or a Kruskal-Wallis test. Others (8–11) executed tests only at the final measurement time or the final time when a substantial fraction of the animals were alive. Still others (9, 11–14) analyzed tumor regrowth or doubling, tripling, or quadrupling times. Two others (15, 16) analyzed animal survival times but not tumor volumes. Of the articles where ANOVA was used, three (3–5) used the Duncan multiple-range test (17) to account for multiple comparisons. One article (14) fit a Gompertz curve to tumor growth in an untreated control group, without any formal statistical testing.

Table 1 lists these methods along with a brief summary of their underlying assumptions and properties. The most popular method was to execute a test at each time point and report all the times where the test is significant. This procedure is attractive because it is simple and uses all the data. Its main weakness is that its type I error rate is not the conventional 5%. To see this, note that if there truly are no treatment differences, there is a 5% chance of significance (a type I error) at each time point. Consequently the overall chance of significance, being the chance of significance at any time, exceeds 5%. A second popular strategy is to analyze only the data from the final measurement time. This procedure has the conventional type I error rate of 5% but is deficient in power, inasmuch as it compares only the ends of the curves and may miss real differences at intermediate times.

The experimenters who looked at doubling and regrowth times analyzed their data by ANOVA or its rank-based analogues. Such methods are not applicable if the times are subject to censoring, *i.e.*, if tumors may fail to double or regrow by the end of the observation period. For this reason it is preferable to use methods that explicitly account for censoring, such as the logrank test (18). These methods generally have correct type I error rates but suboptimal power.

A final criticism that applies to all the methods reviewed is that they generally yield little biological insight; *i.e.*, with these methods one can state which groups are significantly different and possibly rank the groups, but otherwise, because no modeling is being done, it is difficult to relate results to underlying mechanisms.

The past three decades have seen the development of classes of statistical methods designed to avoid these criticisms. Our purpose in this article is to present a subset of these methods that we find best suited to the analysis of tumor growth experiments. We call the methods "multivariate" because they treat the series of tumor volumes on an animal as a single multivariate observation. They use the entire data series and permit detailed modeling of growth curves and intra-animal correlation patterns, thus substantially improving the efficiency of testing and reducing sample size requirements. The methods

Received 3/9/93; accepted 10/12/93.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported by USPHS Grant CA-40011.

[2] To whom requests for reprints should be addressed, at Center for Biostatistics and Epidemiology, Pennsylvania State University College of Medicine, Hershey, PA 17033.

[3] The abbreviations used are: ANOVA, analysis of variance; d.f., degrees of freedom; DFMO, α-difluoromethylornithine; GEE, generalized estimating equations; LR, likelihood ratio; MANOVA, multivariate analysis of variance; RE/AR, random effects/autoregressive.

Table 1 *Methods for statistical analysis of in vivo tumor growth data*

Data used	Analysis method	Key assumptions	Critique
Volumes at all times	ANOVA or Kruskal-Wallis, repeat until significance		Inflated type I error rate
Volumes at final time	ANOVA (*t* test)	Independence Normality Same variance in all treatment groups	Suboptimal power Sensitive to normality
	Kruskal-Wallis (Mann-Whitney)	Independence	Suboptimal power
Doubling times	Logrank test	Independence Proportional hazards	Suboptimal power Sensitivity to hazards assumption
Entire curve	MANOVA	Independence Normality Same variance in all treatment groups	Excludes cases having missing values Suboptimal power
	Multivariate growth-curve analysis	Independence Normality Same variance in all treatment groups Growth curve	Same as MANOVA, but more powerful when growth curve is correct
	Regression with RE/AR errors	Independence Normality Same variance in all groups and at all times RE/AR correlations Growth curve	Most powerful, but potentially sensitive to assumptions Includes all cases

are not new to statistics or even to cancer research (see Ref. 19, Chap. 8), although evidently they are not well known to *in vivo* experimenters. We illustrate the methods by applying them to previously published data on the effect of DFMO, a polyamine biosynthetic inhibitor, on the growth of BT-20 human breast cancer cells in nude mice.

MATERIALS AND METHODS

Experimental Methods: The BT-20 Experiment

The objective of the experiment (20) was to determine whether hormone-independent human breast cancer cells growing in nude mice manifest sensitivity to the polyamine-biosynthetic inhibitor DFMO. Tumors from the BT-20 cell line were established in 4–6-week-old ovariectomized athymic Ncr-*nu* mice (National Cancer Institute, Bethesda, MD) by injecting 5×10^6 cells resuspended in 0.25 ml medium into two mammary fat pads per mouse. We randomized size-matched mice bearing established tumors to one of six DFMO dose groups: 0% (control), 0.5%, 1%, 2%, and 3% in drinking water. We continued treatment until the mice in the 0% group had to be sacrificed because of large tumor burden, measuring tumor volume on days 0 (baseline), 3, 7, 10, 14, and 16 posttreatment. We measured the length (l), width (w), and height (h) of the tumors with a Jamison caliper and calculated volume from the hemiellipsoid formula

$$Y = \pi lwh/6$$

Statistical Methods

In this section we describe three multivariate methods for analyzing tumor growth data (for details see the "Technical Appendix"). All three correct the main flaws of the currently popular methods by achieving their nominal type I error rates and using the entire volume series. Note that the methods require that the data be normally distributed. If the volumes are not normal they can often be made so by transformation; we assume in this section that the logarithmic transformation is appropriate.

The MANOVA Model. The multivariate linear model (21) (see "Appendix Section A.1") takes the animal's vector of log tumor volumes to be the unit of data. In other words, it assumes that the animals are independent but that the observations within an animal may be correlated. The underlying mean vector is the same for all animals in a treatment group, and the unknown variance-covariance matrix is the same for all animals in the population. Mathematically, the model is

$$Y = X_M B_M + \epsilon \qquad \text{(A)}$$

where Y is the matrix of observed log volumes, X_M is a design matrix, B_M is a matrix of regression coefficients, and ϵ is a matrix of random errors, the rows of which are independent and multivariate normal with mean 0.

One can represent a number of tumor growth models with Equation A by appropriate selection of X_M. The model we consider here is called the MANOVA model. In it, the columns of X_M are indicators of dose group membership. The matrix B_M has as many columns as there are measurement times, with each column representing the mean log volume for the five groups at that time. This model asserts that a separate ANOVA model obtains at each time, with the random errors of measurements within the same animal possibly correlated.

If there is a dose effect, between-group differences depend on the measurement time; *e.g.*, the difference between the 0% mean and the 0.5% mean would be one thing on day 0, another on day 6 (a pattern of effects called a dose-by-time interaction). To test this, we translate it into a linear hypothesis about the coefficient matrix B_M (Appendix Equations A.2 and A.3), which we test using the Hotelling-Lawley trace test (21). Henceforth we refer to this as the MANOVA dose-by-time interaction test. It can be executed in the GLM procedure of the SAS System (SAS Institute, Inc., Cary, NC) and other commercial statistical programs. Its power function is complicated, although a noncentral F approximation is available (22). As commonly practiced, *e.g.*, in SAS, the test requires complete data on all animals; thus animals for which any volume measurements are missing are excluded from the analysis. If the test is significant one can proceed to univariate tests at each time; the requirement of a significant MANOVA pretest guarantees that the type I error rate is preserved at 5%.

The Multivariate Growth Curve Model. This model (23) (see Appendix Section A.2) assumes that

$$Y = X_G B_G P_G + \epsilon \qquad \text{(B)}$$

where Y is the matrix of observed log volume data, X_G is a between-animals design matrix, B_G is a matrix of regression coefficients, P_G is a within-individuals design matrix, and ϵ is a matrix of random errors, the rows of which are independent and multivariate normal. It states that in each treatment group the data follow the same kind of curve (specified by P_G), although the coefficients (the rows of B_G) may differ from group to group. The matrix X_G consists of indicator variables denoting group membership, analogous to the X_M matrix in the MANOVA model of Equation A.

The growth curve model is not a special case of the general multivariate linear model, but it can be made so by transforming the log volume data. A popular approach is to compute a set of regression coefficients from each animal and analyze these by MANOVA. In the BT-20 data, for example, sup-

pose that in each dose group the log volume curve is a straight line. To test for dose effects, we test whether the slopes are the same. The Hotelling-Lawley test of this hypothesis reduces to a univariate ANOVA F test applied to the within-animal slopes. Any good statistical package can do such a test, although computing the within-animal slopes may require some effort. The power can again be computed using the noncentral F (22). As in MANOVA, animals with observations missing are typically deleted, although some adaptations (24) retain all the data. See "Discussion" for more on this point.

Regression with Random Effects/Autoregressive Errors. This model (25, 26) (see Appendix Section A.3) asserts that

$$Y_i = X_i\beta_R + \epsilon_i \qquad (C)$$

where for animal i, Y_i is the vector of log tumor volumes, X_i is the matrix of predictors, and ϵ_i is the vector of random errors; β_R is a regression coefficient vector common to all animals. The error is assumed to have a RE/AR variance-covariance matrix, as described below. We henceforth refer to this as the RE/AR regression model.

In the BT-20 study we want to fit a linear model with different slopes but a common intercept. The RE/AR model, unlike the MANOVA and growth curve models, can accommodate this assumption. To assess dose effects one simply tests the hypothesis that all the slopes are equal.

The RE/AR model assumes that the error term is the sum of an animal-specific random effect and an autoregressive process. By "random effect" we mean a random difference between animal i and the mean volume for all animals; *i.e.*, the tendency for the tumor on one animal to be always larger or always smaller than the mean among all animals. An "autoregressive process" is a random process in which the correlation between observations decreases with increasing separation in time. In tumor growth series and other biological data, we commonly observe that the shorter the time between measurements, the higher is their correlation (27). Fitting an autoregressive error model is one way to model this phenomenon.

We have fit the RE/AR regression model using a program we have written in Fortran and *S-Plus* Version 3.1 (Statistical Sciences Inc., Seattle, WA). It can also be fit in BMDP program 5V (BMDP, Inc., Los Angeles, CA) and SAS procedure MIXED [available in Versions 6.07 and later (SAS Institute, Inc., Cary, NC)]. Hypotheses about the regression parameters can be tested in a number of ways; we have used LR tests, the power of which we approximate with the noncentral χ^2 (28). The RE/AR model uses all the data, and consequently there is no need to exclude incompletely observed animals.

Comparison of the Multivariate Models. The MANOVA model is the most general of the multivariate models in that it places no restrictions on the shape of the growth curves or the variance matrix. The growth curve model is similar to MANOVA except that it restricts the mean growth curves to be of the same parametric type (in our example a straight line), with different coefficients in each group. The RE/AR model differs from the growth-curve model in allowing the dose groups to have common regression coefficients (in our example the intercept). Unlike the other models, it assumes that the standard deviation is the same at each time in all groups, and correlation within animals follows the RE/AR pattern. Incorporating these more restrictive assumptions makes the analysis more powerful, if they are justified. If not, both type I error rate and power may suffer.

RESULTS

In this section we analyze the BT-20 data and compare the candidate methods in terms of power and type I error rate. To give an idea of the steps involved in a multivariate analysis, we present detailed results for the RE/AR model.

Regression Modeling of the BT-20 Data. The first step in RE/AR modeling is to verify that the data reflect the model assumptions, at least approximately. The key assumptions, outlined in Table 2, are that (a) the SDs are equal at all times in all groups, (b) the errors are symmetrical, (c) correlations follow the RE/AR model, and (d) the growth curve is adequately represented by the proposed parametric form (*e.g.*, a straight line, a quadratic or a spline). Note that in assumption b we attempt to assess error symmetry rather than normality because it is difficult to formally test normality, and symmetry is presumably the critical feature of normality.

Table 2 *Key assumptions of the RE/AR regression model*

Assumption	Diagnostic	Action to take
Equal SD at all times in all groups	Spread-*vs.*-level plot (29)	Power transformation
Symmetrical errors	Symmetry plot (29)	Power transformation
RE/AR covariances	Semivariogram (26)	Adjust variance model
Shape of the curve	Straightness plot (29)	Power transformation
	Goodness-of-fit significance tests	Select best-fitting model within appropriate family

A class of data-analytic techniques (29) is available for assessing assumptions (i), (ii) and (iv). These tools highlight departures from the assumptions and suggest ways to transform the data to make the assumptions more nearly true, as shown in Table 2. Application to the BT-20 data directed us to a range of possible transformations, including the log and the square root. We chose the log because it has been selected by many previous datasets, and slopes of log-scale data have a simple biological interpretation.

Fig. 1 displays boxplots (30) of tumor volume by time for the five dose groups. The comparable sizes of the boxes demonstrate that the log transformation has rendered the SDs nearly equal. The whiskers show that the distributions are roughly symmetrical, although there are occasional outliers (displayed as *dots*), most often on the low side. The plot of mean log growth curves (Fig. 1*f*) shows that growth is roughly log-linear, with slope decreasing as dose increases.

To select an empirical best model we fit a sequence of models, comparing their fits via significance tests. A first question is the basic shape of curves to use in subsequent modeling. Solid tumor growth curves are often Gompertzian in that they start out nearly log-linear but later flatten at a limiting volume. We had expected to see flattened curves in this experiment, although we had some idea that the time sampling was so short that the curves would be nearly linear. We thus compared two key models: a "full linear" model with a common intercept and dose-specific slopes (Equations A.10–A.12), and a "full quadratic" model with a common intercept and dose-specific linear and quadratic terms. Although the quadratic model is not Gompertzian, it should be a much better approximation than the linear model. The full models are nested (*i.e.*, one obtains the linear from the quadratic by setting the time-squared coefficients to 0) and thus can be compared by a LR test. For the BT-20 data the LR χ^2 statistic is 3.4 on 5 d.f. for a P value of 0.64. This suggests that departures from linearity are small relative to the discriminating power of the data.

To test the RE/AR Assumption iii, we reestimated the variance and covariance empirically using the semivariogram (26), a plot of the covariance of observations 0 units part minus the covariance at Δ units apart, as a function of Δ. Fig. 2 compares the empirical and best-fitting RE/AR semivariograms from the BT-20 data. Agreement is good, suggesting that the RE/AR assumption is adequate for these data. The semivariogram estimate of the variance of a single log volume is 0.541, in good agreement with the model-based estimate of 0.536. If the fit had been less encouraging we could have adopted one of the more general covariance models proposed by Diggle (26).

As noted above, Fig. 1 suggests a monotonic dependence of the growth rate on the DFMO dose. The slope estimates (Table 3) bear this out. To assess this empirically we conducted a series of LR tests, first comparing the full linear model to a linear model with a common slope and intercept, essentially a test for any dose effects. As Fig. 1 suggests, the full model fits significantly better (LR = 74.5 on 4 d.f., $P = 3 \times 10^{-15}$). Having established that the groups differ, we sought

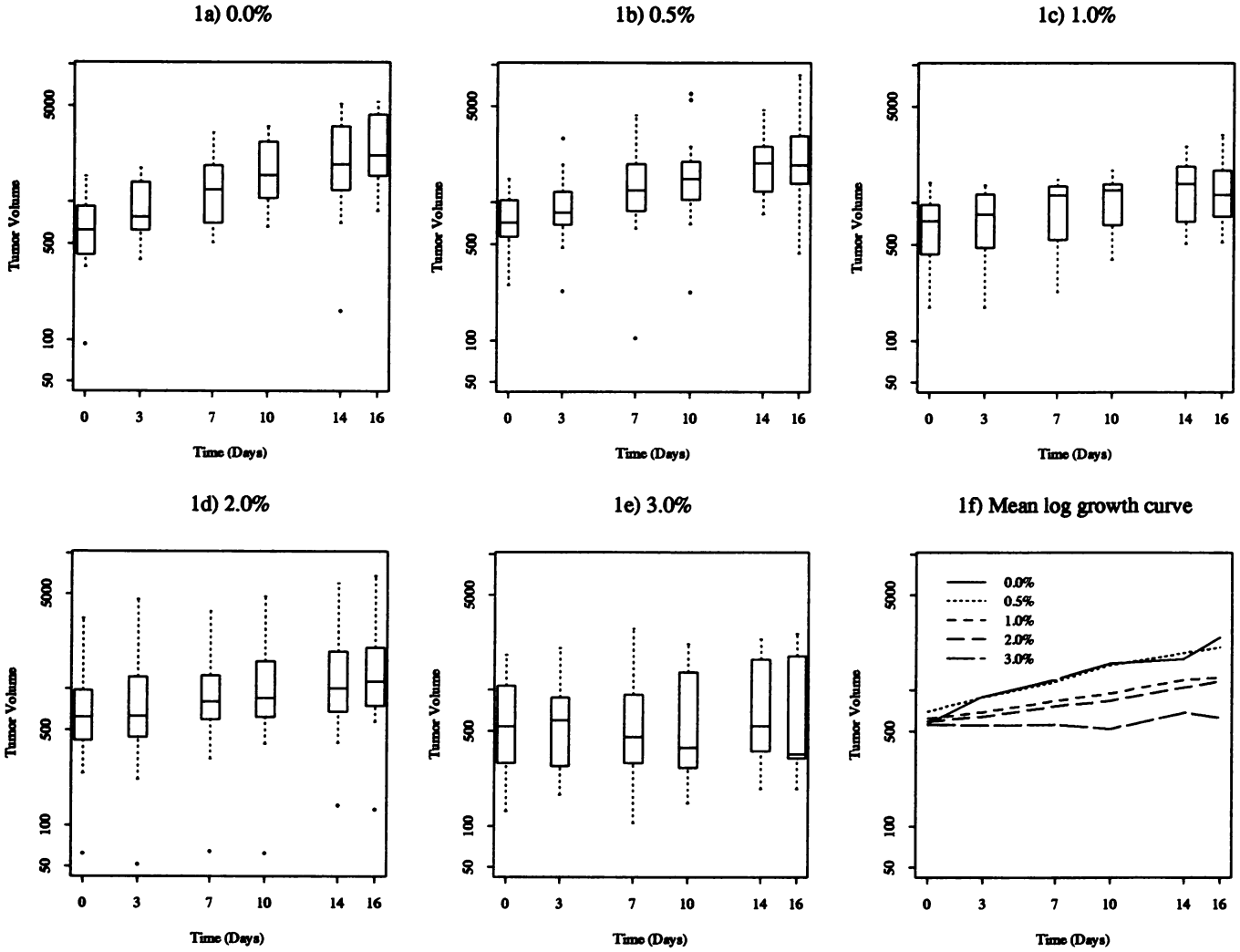

Fig. 1. Boxplots (30) of mean tumor volume (log scale) *versus* time since the beginning of DFMO therapy, by dose group (*a–e*), and mean log growth curves for all five dose groups (*f*).

a simpler description of the dependence of the slope on the dose. We tried several possibilities, including a linear model, a quadratic model, and a linear model with an additional parameter for active *versus* placebo. None fit as well as the model with arbitrary slopes. Thus we concluded that, within the ability of these data to resolve such questions, (*a*) tumor growth rates differ between dose groups, and (*b*) the relationship between DFMO dose and growth rate is negative and monotonic, but (*c*) a simple linear or quadratic function cannot adequately describe the dependence of slope on dose.

The slope in a log-linear growth model can be interpreted as the proliferation rate minus the death rate. The fact that the slope decreases with increasing dose suggests that DFMO affects one or both of these rates in a dose-dependent manner. Interestingly, analysis of a parallel experiment involving the hormone-dependent cell line MCF-7 showed a nonmonotone dose effect, raising questions about the mechanisms of DFMO action in these cell lines.

Comparison with Other Methods of Analysis. We also analyzed the BT-20 data using the other multivariate methods and the methods gleaned from the cancer literature. First we executed ANOVA and Kruskal-Wallis tests comparing the dose groups at each time point. The groups differed significantly at days 10 and beyond by ANOVA and days 7 and beyond by Kruskal-Wallis. A logrank test comparing the doubling time distributions was highly significant, as were similar tests for tripling and quadrupling times. The MANOVA dose-by-time interaction test ($F = 3.8$ with 24 and 222 d.f.) gave a P value of 5×10^{-8}. The growth curve slope ANOVA was also significant, with $F = 15.1$ on 4 and 62 d.f. ($P = 1 \times 10^{-8}$). In short, all the methods conclude that the dose groups are significantly different.

Comparison of Type I Error Rates. For the multivariate methods and ANOVA at the final measurement, statistical theory tells us that the type I error rates are close to 5% as long as model assumptions are nearly correct. We estimated the error rates of the other tests (ANOVA or Kruskal-Wallis at all measurement times, and logrank on doubling times) by a Monte Carlo experiment. We generated data under the assumptions that (*a*) the true, underlying growth curves are all equal, with intercept and slope equal to estimates from the 0 dose group under the RE/AR model (from Table 3) and (*b*) the true underlying variance matrix is RE/AR (Equation A.13), with parameters equal to the estimates from the BT-20 data. We simulated 1000 independent data sets having the design of the BT-20 experiment. Each dataset consisted of 5 groups of 15 animals, each animal having tumor volumes measured on days 0, 3, 7, 10, 14, and 16. We applied all the tests to each simulated dataset at each sample size, by taking the first n units from each group, $n = 2, 3, ..., 15$. We estimated type I error rates as the fraction of simulated datasets where significance was attained.

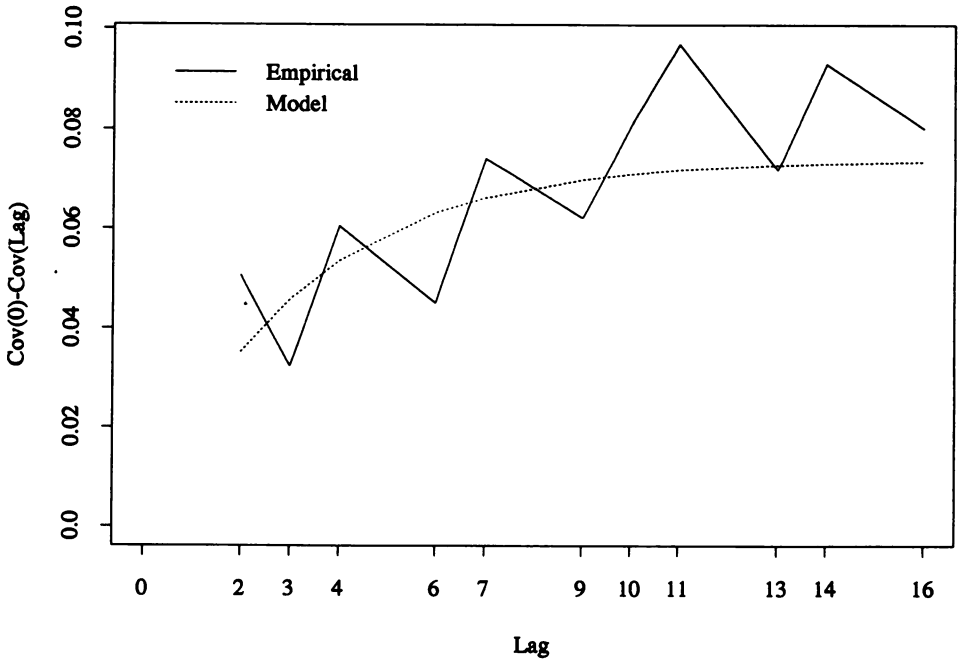

Fig. 2. Empirical and model semivariograms based on the full regression model.

Fig. 3 plots the type I error rates *versus* the sample size per group n. The solid lines indicate the target rate (5%) ± two Monte Carlo SEs (SE = $\sqrt{0.05 \times 0.95/1000}$); thus a symbol lying outside the two outer solid lines differs significantly from the target value. Although we expected the logrank test to have an error rate near 5%, we were concerned that it would be off somewhat because of the small sample sizes and discreteness in the doubling time distribution. Fig. 3 shows that the error rate of the test is never far from 5% and improves rapidly with increasing n. On the other hand, the type I error rate of testing at each time point (by either ANOVA or Kruskal-Wallis) considerably exceeds 5%.

Comparison of Powers. Among tests having equal type I error rates, the most powerful is generally preferable. Thus we compared our tests by computing their powers under the BT-20 design for n = 2, ..., 15. This time we assumed that (*a*) the true, underlying growth curves differed, with growth parameters equal to the RE/AR parameter estimates for the BT-20 data (Table 3), and (*b*) the true underlying variance matrix is RE/AR (equation A.13), with parameters equal to the estimates from the RE/AR model for the BT-20 data. We computed powers for three tests (ANOVA of data from the final day, MANOVA, and ANOVA on the slopes) by noncentral F approximation (22). We computed the power of the LR test in the RE/AR model by a noncentral χ^2 approximation (28), and the power of the logrank test on doubling times by Monte Carlo simulation.

Fig. 4 plots the power of each test as a function of n, the sample size per group. The methods that use all the data and exploit the underlying model—in this case the RE/AR model and the ANOVA on slopes—have greatest power. The MANOVA dose-by-time interaction test and the logrank test on doubling times are less powerful, because, although they use all the data, they do not exploit the linearity of the log-volume curves. The least powerful method is ANOVA on data from the final day, which uses only a small fraction of the data and totally ignores the shape of the curves. The minimum sample sizes required for 90% power reflect these differences: For RE/AR, the minimum n is 3, for the slope ANOVA it is 4, for MANOVA and the logrank test it is 7, and for ANOVA on the final day it is 11.

Although our calculations suggest that n = 3/group is adequate, in practice we would not run such a small study. First, the power approximation for the RE/AR model assumes that the variance parameters are known *a priori*, which is never the case. The power of RE/AR is therefore somewhat overstated, although we suspect that the effect is to underestimate sample size by only one or two animals per group. Second, our computations assume that all tumors grow and no animals die prematurely, whereas in reality such data losses are common and need to be provided for. Finally, n = 3/group may not give sufficient power for other outcomes of interest. For example, in the BT-20 experiment tumor polyamine levels were an important end point. Because these can be measured only at sacrifice, there is no alternative to a univariate analysis for this end point, and consequently a larger sample size is necessary.

These comparisons do not imply that RE/AR is best in every situation. Although our diagnostic analyses suggest that log-linear growth and RE/AR covariance are reasonable assumptions for the BT-20 data, if they were not, the power advantage of RE/AR could be reduced or even reversed. However, it is generally true that the use of detailed model information leads to more powerful tests; thus it is best to use as much of this information as is available.

Table3 *Estimated slopes of log tumor growth: BT-20 DFMO experiment*

DFMO dose (%)	Estimated Slope ± SE
0.0	0.0906 ± 0.0064
0.5	0.0723 ± 0.0060
1.0	0.0445 ± 0.0058
2.0	0.0411 ± 0.0058
3.0	0.0137 ± 0.0061

DISCUSSION

Table 1 summarizes the methods we have discussed. We find the advantages of the multivariate methods (correct type I error probabilities, enhanced power, and the capacity to model data rather than just test significance) compelling. Among the multivariate models, the RE/AR model uses the data most efficiently but requires the most work to apply.

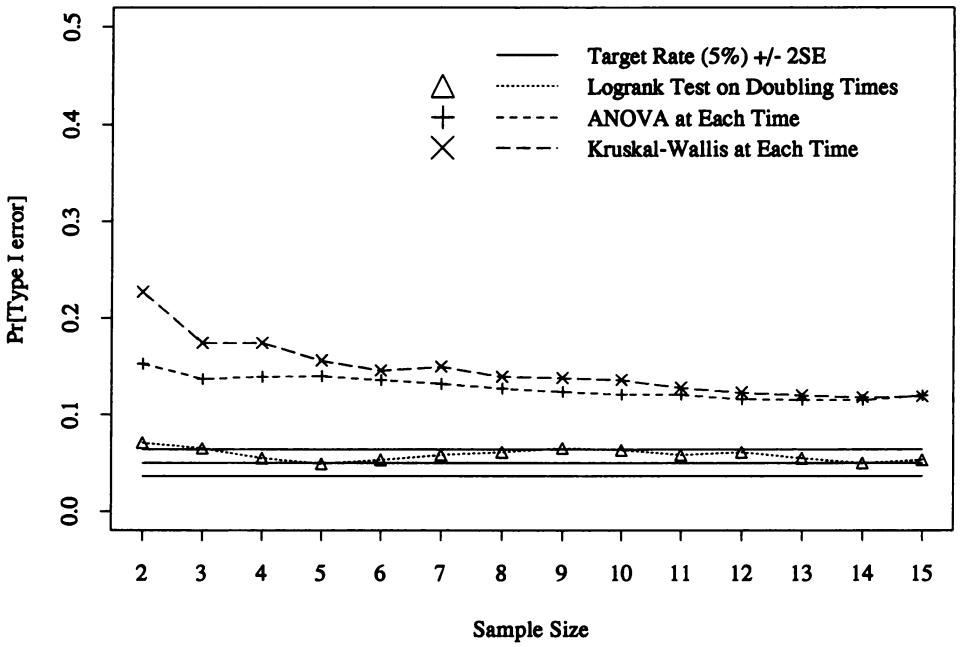

Fig. 3. Type I error rate as a function of sample size per group for three methods of analysis.

Although we have concentrated on log-linear models, one can, and often should, use other shapes to describe the basic growth curve. For example, cancer treatments are often administered in pulses that reduce tumor volume transiently before a period of regrowth. In such cases the volume curves are not monotone and are better described by spline (piecewise polynomial) models (31). We have used this approach to analyze data from an *in vivo* study of androgen priming in prostate cancer (32).

The practical price of the superior properties of the multivariate methods is the greater expense of applying them. Table 2 summarizes the assumptions of the models, the diagnostics that address their adequacy, and the alternatives available when there are problems. Successful execution of the diagnostics and the modeling requires a level of programming skill and statistical judgment beyond what one can acquire in a typical elementary statistics course. Thus many investigators will need professional statistical assistance to apply these models.

Missing data, usually resulting from animal mortality or morbidity, is a common and potentially serious problem in growth analyses. If the missing data are ignorable, in the sense that the probability of the observed missingness pattern does not depend on the observed or unobserved data values, then it is appropriate to treat the missing data

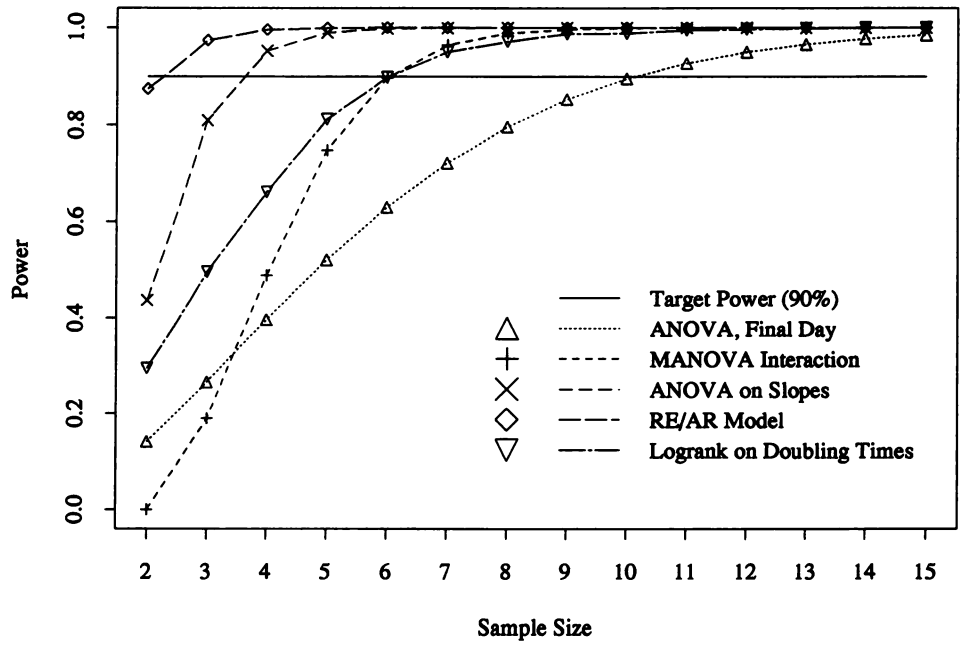

Fig. 4. Power as a function of sample size per group for five methods of analysis whose type I error rate is exactly or approximately 5%.

as though they were missing by design. If the missingness is not ignorable, then parameter estimates may be biased, and significance tests may have type I error rates exceeding their target values (33).

When missingness is ignorable, several analysis strategies are available. As has been indicated, implementations of the MANOVA and growth curve models in the major statistical packages require balanced data, and to obtain it all animals are deleted for which there are any missing observations. This can result in a considerable loss of efficiency if many animals have missing data. Vonesh and Carter (24) have proposed a method for fitting these models that uses all the data. As indicated above, RE/AR modeling automatically uses all available data.

When missingness is not ignorable, one must model both the growth data and the missingness pattern. Two approaches have been proposed. In the first (34, 35), one assumes that each animal has its own underlying slope and that the probability that the animal is lost depends on its slope. In the second (36), one assumes that the time of dropout can depend on previous and current values of tumor volume. Whichever model one uses, it is necessary to estimate the growth curve and missingness parameters simultaneously.

Unfortunately, inferences under nonignorable models can be sensitive to the assumed model; *i.e.*, erroneous assumptions about the underlying distributions, usually very difficult to detect, can cause serious errors in inferences (37). Further theoretical and empirical research is needed to elucidate the proper methods for modeling incompleteness in tumor growth data.

A second problem with the analysis of tumor growth data involves selection of the transformation to normality and the regression model. Some statisticians claim that one should explicitly adjust the analysis if the data are used to select a transformation or model, whereas others argue that this is unnecessary (38). In this study, for example, our choice of log-linear models was based on analyses of the same data; therefore a more rigorous analysis would adjust estimates and tests to account for this selection. Yet it is common practice to ignore this problem, and there exist no practical methods for making such adjustments.

The methods we have presented are just a few of the many techniques available for analyzing tumor growth data. For example, the RE/AR model is a special case of the longitudinal-data linear model of Laird and Ware (39). This model accommodates more general random effects (such as random slopes) and serial correlation structures (including higher-order autoregressions). Although these more general correlation models may be valuable in many applications, we believe that the RE/AR model captures the most important features of tumor growth data.

All our multivariate models assume normality and linearity, but other methods are available when such assumptions are restrictive or unwarranted. One approach is to base significance tests on the distribution of multivariate rank statistics (40) or the randomization distribution (41). These tests are reliable under assumptions more general than ours, with some loss of efficiency if a normal model is actually appropriate. The GEE approach (42) involves specifying only the shape of the growth curve and the covariance matrix, not the underlying distribution. GEE is robust to errors in the assumed correlation structure and can handle arbitrary patterns of missing data; like the rank and randomization tests, it is reliable but potentially inefficient. Another approach involves modeling tumor growth curves with nonlinear rather than linear models (43–46). This is more difficult to apply than linear modeling but can give greater insight into the biological processes of tumor growth. None of the methods cited in this paragraph is available in production versions of major statistical packages, although some good programs are publicly available.

In summary, statisticians have developed an array of multivariate methods that can dramatically improve the analysis of tumor growth studies. Careful application of these methods will lead to more efficient and humane experiments and more valid and comprehensive data analyses.

TECHNICAL APPENDIX

A.1. The MANOVA model (21) states that

$$Y = X_M B_M + \epsilon \tag{A.1}$$

where Y is an $N \times p$ matrix of observed log volume data, X_M is an $N \times r$ design matrix, B_M is an $r \times p$ matrix of regression coefficients, and ϵ is an $N \times p$ error matrix the rows of which are independent and multivariate normal with mean 0 and variance-covariance matrix Σ. Here N refers to the number of animals, p to the number of times each animal is measured, and r to the number of predictors in the design matrix. The ith row of Y is the vector of log volumes for the ith animal, and the ith row of X_M is the design for the ith animal. The columns of B_M are the model coefficients, with one column for each of the p measurement times.

To apply the multivariate linear model in the BT-20 example, suppose for the moment that there is $n = 1$ animal in each of the five dose groups. Because each animal's tumor volume is measured at six times, $p = 6$; with five groups and one animal per group, $N = 5$. A simple model would assume a time effect (*i.e.*, the tumors grow) and a group effect (the volume depends on the dose). There is a dose-by-time interaction if the time effects differ by group. In terms of model (A.1), X_M is the 5×5 identity matrix and B_M is the 5×6 matrix where the row-i/column-j element is the expected tumor volume at the jth time for an animal in group i. When there are n animals per group, X_M is simply n copies of the 5×5 identity matrix stacked vertically.

To test for a dose-by-time interaction, we express the null hypothesis as a general multivariate linear hypothesis $C_M B_M U_M = 0$, where C_M is a "between-units" contrast matrix and U_M is a "within-units" contrast matrix. The hypothesis of parallel growth curves (*i.e.*, no dose-time interaction) has contrast matrices

$$C_M = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{pmatrix} \tag{A.2}$$

and

$$U_M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix} \tag{A.3}$$

The Hotelling-Lawley test of this hypothesis (21) can be executed with the GLM procedure in the SAS System. The power can be approximated with the noncentral F (22).

A.2. The multivariate growth curve model (23) states that

$$Y = X_G B_G P_G + \epsilon \tag{A.4}$$

where Y is an $N \times p$ matrix of observed log volume data, X_G is an $N \times g$ between-animals design matrix, B_G is a $g \times q$ matrix of regression coefficients, P_G is a $q \times p$ within-individuals design matrix, and ϵ is an $N \times p$ error matrix the rows of which are independent and multivariate normal with mean 0 and variance-covariance matrix Σ. Here N refers to the number of animals studied, p to the number of times each tumor is measured, g to the number of treatment groups, and q to the number of predictors in the within-individual design matrix. The ith row of Y is the vector of log volumes for the ith animal and the ith row of X_G is the between-animals design for animal i. The jth row of B_G is the vector of regression coefficients for animals in the jth treatment group.

In the BT-20 example, again take $n = 1$ animal in each of the five dose groups. Because each animal's tumor volume is measured on six occasions, $p = 6$, and with five groups and one animal per group, $N = 5$ and $g = 5$. Assuming tumor volume is log-linear in time, $q = 2$; *i.e.*, the tumor growth curve in each group is described by two parameters, a slope and intercept. Thus in Equation A.4, X_G is the 5×5 identity matrix, B_G is the 5×2 matrix the jth row of which is the intercept and slope for group j, and P_G is the within-animal design, in this case

$$P_G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 3 & 7 & 10 & 14 & 16 \end{pmatrix} \quad \text{(A.5)}$$

With n animals per group, X_G is merely n copies of the 5×5 identity matrix stacked vertically.

Model A.4 is not a special case of A.1, but it can be made so by appropriate reduction of the volume data. For example, instead of analyzing Y, one analyzes $Y^{(h)} = YH$, where

$$H = P_G^T (P_G^T P_G)^{-1} \quad \text{(A.6)}$$

The transformed data satisfy

$$Y^{(h)} = X_G B_G + \delta \quad \text{(A.7)}$$

where the rows of δ are independent and multivariate normal with mean 0 and variance-covariance matrix $\Gamma = H^T \Sigma H$. Practically, this means reducing each animal's data from a vector of log volumes to an animal-specific intercept and slope, which one then analyzes using MANOVA. Because it reduces the data this test involves some loss of information, but depending on the model and the reduction the loss can be small.

A general linear hypothesis for testing equality of slopes is $C_G B_G U_G = 0$ where C_G is the 4×5 contrast matrix

$$C_G = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{pmatrix} \quad \text{(A.8)}$$

and U_G is the 2×1 matrix

$$U_G = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{(A.9)}$$

With these contrast matrices the Hotelling-Lawley test reduces to a univariate ANOVA F test on the within-animal slopes. One can compute the exact power from the noncentral F (22).

A.3. The RE/AR regression model (25, 26) asserts that

$$Y_i = X_i \beta_R + \epsilon_i \quad \text{(A.10)}$$

where for animal i, Y_i is the column vector of p_i log tumor volumes, X_i is the $p_i \times q$ matrix of predictors, and ϵ_i is the vector of p_i random errors; β_R is a $q \times 1$ regression coefficient common to all animals. The error term is the sum of a random animal effect and an autoregressive process; hence we call this the RE/AR regression model.

In the BT-20 study we assume a linear model with dose-specific slopes and a common intercept. Because there are six tumor volumes per animal, $p_i = 6$. There are five treatment groups with a common intercept and a separate slope for each group; therefore $q = 6$—one parameter (β_0) is the intercept and the five others ($\beta_1^{(d)}$, d = 0,0.5,1,2,3) are the slopes:

$$\beta_R = (\beta_0, \beta_1^{(0)}, \beta_1^{(0.5)}, \beta_1^{(1)}, \beta_1^{(2)}, \beta_1^{(3)})^T \quad \text{(A.11)}$$

An animal in the dose group DFMO = 2% therefore has design matrix

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 3 & 0 \\ 1 & 0 & 0 & 0 & 7 & 0 \\ 1 & 0 & 0 & 0 & 10 & 0 \\ 1 & 0 & 0 & 0 & 14 & 0 \\ 1 & 0 & 0 & 0 & 16 & 0 \end{pmatrix} \quad \text{(A.12)}$$

The error term in this model is the sum of a random animal effect and an autoregressive process. We assume that the random effects are normal with mean 0 and variance τ^2. In an autoregressive process, the correlation of measurements at a distance of Δt time units is $\rho^{|\Delta t|}$, where ρ is the autocorrelation. An autoregressive process is parameterized by ρ and a scale parameter ς. Under the BT-20 design, the error term ϵ_i in Equation A.10 is multivariate normal with mean 0 and variance-covariance.

$$Var(\epsilon_i) = \tau^2 JJ^T + \varsigma^2/(1 - \rho^2) \times \begin{pmatrix} 1 & \rho^3 & \rho^7 & \rho^{10} & \rho^{14} & \rho^{16} \\ \rho^3 & 1 & \rho^4 & \rho^7 & \rho^{11} & \rho^{13} \\ \rho^7 & \rho^4 & 1 & \rho^3 & \rho^7 & \rho^9 \\ \rho^{10} & \rho^7 & \rho^3 & 1 & \rho^4 & \rho^6 \\ \rho^{14} & \rho^{11} & \rho^7 & \rho^4 & 1 & \rho^2 \\ \rho^{16} & \rho^{13} & \rho^9 & \rho^6 & \rho^2 & 1 \end{pmatrix} \quad \text{(A.13)}$$

where J is a 6×1 matrix of 1s.

We test hypotheses about the regression parameters using likelihood-ratio tests. With the model of Equations A.10–A.12, the hypothesis of equal slopes is $C_R \beta_R = 0$, where

$$C_R = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} \quad \text{(A.14)}$$

The power of this test can be approximated with the noncentral χ^2 (28).

We fit the RE/AR model in a program we wrote in *S-Plus* (Statistical Sciences, Inc.). One can also fit this model in the SAS MIXED procedure (SAS Institute, Inc.) and BMDP program 5V (BMDP, Inc.).

REFERENCES

1. Damber, J-E., Bergh, A., Daehlin, L., Petrow, V., and Landström, M. Effects of γ-methylene progesterone on growth, morphology, and blood flow of the Dunning R3327 prostatic adenocarcinoma. Prostate, *20:* 187–197, 1992.
2. Jani, J. P., Mistry, J. S., Morris, G., Davies, P., Lazo, J. S., and Sebti, S. M. *In vivo* circumvention of human colon carcinoma resistance to bleomycin. Cancer Res., *52:* 2931–2937, 1992.
3. Milovanovic, S. R., Radulovic, S., Groot, K., and Schally, A. V. Inhibition of growth of PC-82 human prostate cancer line xenografts in nude mice by bombesin antagonist RC-3095 or combination of agonist [D-Trp^6]-luteinizing hormone-releasing hormone and somatostatin analog RC-160. Prostate, *20:* 269–280, 1992.
4. Yano, T., Korkut, E., Pinski, J., Szepeshazi, K., Milovanovic, S., Groot, K., Clarke, R., Comaru-Schally, A. M., and Schally, A. V. Inhibition of growth of MCF-7 MIII human breast carcinoma in nude mice by treatment with agonists of LH-RH. Breast Cancer Res. Treat., *21:* 35–45, 1992.
5. Yano, T., Pinski, J., Szepeshazi, K., Milovanovic, S. R., Groot, K., and Schally, A. V. Effect of microcapsules of luteinizing hormone-releasing hormone antagonist SB-75 and somatostatin analog RC-160 on endocrine status and tumor growth in the Dunning R-3327H rat prostate cancer model. Prostate, *20:* 297–310, 1992.
6. Yeh, M-Y., Roffler, S. R., Yu, M-H. Doxorubicin: monoclonal antibody conjugate for therapy of human cervical carcinoma. Int. J. Cancer, *51:* 274–282, 1992.
7. Kasprzyk, P. G., Song, S. U., Di Fiore, P. P., and King, C. R. Therapy of an animal model of human gastric cancer using a combination of anti-*erb*B-2 monoclonal antibodies. Cancer Res., *52:* 2771–2776, 1992.
8. Forssen, E. A., Coulter, D. M., and Proffitt, R. T. Selective *in vivo* localization of daunorubicin small unilamellar vesicles in solid tumors. Cancer Res., *52:* 3255–3261, 1992.
9. Ichikawa, T., Lamb, J. C., Christensson, P. I., Hartley-Asp, B., and Isaacs, J. T. The antitumor effects of quinoline-3-carboxamide linomide on Dunning R-3327 rat prostatic cancers. Cancer Res., *52:* 3022–3028, 1992.
10. Keller, R. P., Altermatt, H. J., Donatsch, P., Zihlmann, H., Laissue, J. A., and Hiestand, P. C. Pharmacologic interactions between the resistance-modifying cyclosporine SDZ PSC 833 and etoposide (VP 16-213) enhance *in vivo* cytostatic activity and toxicity. Int. J. Cancer, *51:* 433–438, 1992.
11. Sakaguchi, Y., Maehara, Y., Baba, H., Kusumoto, T., Sugimachi, K., and Newman, R. Flavone acetic acid increases the antitumor effect of hyperthermia in mice. Cancer Res., *52:* 3306–3309, 1992.
12. Schem, B-C., Mella, O., and Dahl, O. Thermochemotherapy with cisplatin or carboplatin in the BT$_4$ rat glioma *in vitro* and *in vivo*. Int. J. Radiat. Oncol. Biol. Phys., *23:* 109–114, 1992.
13. Song, C. W., Hasegawa, T., Kwon, H. C., Lyons, J. C., and Levitt, S. H. Increase in tumor oxygenation and radiosensitivity caused by pentoxifylline. Radiat. Res., *130:* 205–210, 1992.
14. Pop, L., Levendag, P., van Geel, C., Deurloo, I-K., and Visser, A. Lack of therapeutic

gain when low dose rate interstitial radiotherapy is combined with cisplatin in an animal tumour model. Eur. J. Cancer, *28A:* 1471–1474, 1992.
15. Mayhew, E. G., Lasic, D., Babbar, S., and Martin, F. J. Pharmacokinetics and antitumor activity of epirubicin encapsulated in long-circulating liposomes incorporating a polyethylene glycol-derivatized phospholipid. Int. J. Cancer, *51:* 302–309, 1992.
16. Shoemaker, R. H., Smythe, A. M., Wu, L., Balaschak, M. S., and Boyd, M. R. Evaluation of metastatic human tumor burden and response to therapy in a nude mouse xenograft model using a molecular probe for repetitive human DNA sequences. Cancer Res., *52:* 2791–2796, 1992.
17. Miller, R. G. Simultaneous Statistical Inference, Ed. 2. New York: Springer-Verlag, 1981.
18. Kalbfleisch, J. D., and Prentice, R. L. The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons, Inc., 1980.
19. Gart, J. J., Krewski, D., Lee, P. N., Tarone, R. E., and Wahrendorf, J. Statistical Methods in Cancer Research, Vol. 3. Lyon, France: International Agency for Research on Cancer, 1986.
20. Manni, A., Badger, B., Martel, J., and Demers, L. Role of polyamines in the growth of hormone-responsive and -resistant human breast cancer cells in nude mice. Cancer Lett., *66:* 1–9, 1992.
21. Morrison, D. F. Multivariate Statistical Methods, Ed. 2. New York: McGraw-Hill Book Co., 1976.
22. Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. Power calculations for general linear multivariate models including repeated measures applications. J. Am. Stat. Assoc., *87:* 1209–1226, 1992.
23. Potthoff, R. F., and Roy, S. N. A generalized multivariate analysis of variance model useful especially in growth curve problems. Biometrika, *51:* 313–326, 1964.
24. Vonesh, E. F., and Carter, R. L. Efficient inference for random-coefficient growth curve models with unbalanced data. Biometrics, *43:* 617–628, 1987.
25. Chi, E. M., and Reinsel, G. C. Models for longitudinal data with random effects and AR(1) errors. J. Am. Stat. Assoc., *84:* 452–459, 1989.
26. Diggle, P. J. An approach to the analysis of repeated measurements. Biometrics, *44:* 959–971, 1988.
27. Jones, R. H. Serial correlation in unbalanced mixed models. Bull. Int. Stat. Inst. Proc. 46th Session, *4:* 105–122, 1987.
28. Kendall, M., and Stuart, A. The Advanced Theory of Statistics, Ed. 4, Vol. 2. London: Charles Griffin, 1979.
29. Emerson, J. D. Mathematical aspects of transformation. *In:* D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.), Understanding Robust and Exploratory Data Analysis, Chap. 8. New York: John Wiley & Sons, Inc., 1983.
30. Emerson, J. D., and Strenio, J. Boxplots and batch comparison. D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.), Understanding Robust and Exploratory Data Analysis, Chap. 3. New York: John Wiley & Sons, Inc., 1983.
31. Durrleman, S., and Simon, R. Flexible regression models with cubic splines. Stat. Med., *8:* 551–561, 1989.
32. English, H. F., Heitjan, D. F., Lancaster, S., and Santen, R. J. Beneficial effects of androgen primed chemotherapy in the Dunning R3327 G model of prostatic cancer. Cancer Res., *51:* 1760–1765, 1991.
33. Laird, N. M. Missing data in longitudinal studies. Stat. Med., *7:* 305–315, 1988.
34. Wu, M. C., and Bailey, K. R. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. Biometrics, *45:* 939–955, 1989.
35. Schluchter, M. D. Methods for the analysis of informatively censored longitudinal data. Stat. Med., *11:* 1861–1870, 1992.
36. Diggle, P. J., and Kenward, M. G. Informative dropout in longitudinal data analysis. (With discussion.) Appl. Stat., *43:* in press, 1994.
37. Little, R. J. A. Models for nonresponse in sample surveys. J. Am. Stat. Assoc., *77:* 237–250, 1982.
38. Hinkley, D. V., and Runger, G. The analysis of transformed data. J. Am. Stat. Assoc., *79:* 302–320, 1984.
39. Laird, N. M., and Ware, J. H. Random-effects models for longitudinal data. Biometrics, *38:* 963–974, 1982.
40. Koziol, J. A., Maxwell, D. A., Fukushima, M., Colmerauer, M. E. M., and Pilch, Y. H. A distribution-free test for tumor-growth curve analyses with application to an animal tumor immunotherapy experiment. Biometrics, *37:* 383–390, 1981.
41. Zerbe, G. O. Randomization analysis of the completely randomized design extended to growth and response curves. J. Am. Stat. Assoc., *74:* 215–221, 1979.
42. Liang, K. Y., and Zeger, S. L. Longitudinal data analysis using generalized linear models. Biometrika, *73:* 13–22, 1986.
43. Cressie, N., and Hulting, F. L. A spatial statistical analysis of tumor growth. J. Am. Stat. Assoc., *87:* 272–283, 1992.
44. Heitjan, D. F. Generalized Norton-Simon models of tumor growth. Stat. Med., *10:* 1075–1088, 1991.
45. Lindstrom, M. J., and Bates, D. M. Nonlinear mixed effects models for repeated measures data. Biometrics, *46:* 673–687, 1990.
46. Vonesh, E. F., and Carter, R. L. Mixed-effects nonlinear regression for unbalanced repeated measures. Biometrics, *48:* 1–17, 1992.