The Price of Kaplan–Meier

Paul MEIER, Theodore KARRISON, Rick CHAPPELL, and Hui XIE

Miller has studied the asymptotic efficiency of the nonparametric, Kaplan–Meier survival estimator relative to parametric estimates based on the exponential and Weibull distributions. He concluded that in certain cases, the asymptotic efficiency is low and recommended that analysts give more consideration to parametric estimators, particularly for estimation of small tail probabilities. In this article we revisit this issue and examine the performance of the nonparametric procedure for estimation not only of a point on the survival curve, but also of the mean (or restricted mean) lifetime. In addition to the exponential and Weibull families, we consider the performance of the Kaplan–Meier procedure relative to a more flexible parametric model proposed by Efron. We find that the reduction in efficiency of the Kaplan–Meier survival estimate becomes negligible fairly quickly as the number of parameters in the parametric model increases. Moreover, for estimation of the mean or restricted mean, the loss in efficiency, even relative to the exponential distribution, is small or nil. We conclude that a parametric estimate of the survival curve may be necessary in certain extreme situations, such as when the sample size is very small. In these cases, careful attention must be given to considering the degree of fit, although with sparse data, this must be assessed from outside sources. For certain functionals of the survival curve, such as the mean or restricted mean, the nonparametric approach is unbiased and entails little or no loss in efficiency, and therefore would generally be preferred over a parametric-based estimate.

KEY WORDS: Efficiency; Kaplan–Meier estimator; Mean squared error; Parametric modeling; Restricted mean.

1. INTRODUCTION

Twenty-three years ago, after the publication of "Nonparametric Estimation from Incomplete Observations" by Kaplan and Meier (1958), Rupert Miller wrote a textbook titled *Survival Analysis* (Miller 1981), which gave a thorough review of the Kaplan–Meier method as well as various parametric distributions and procedures for survival analysis. Two years later he published an article titled "What Price Kaplan–Meier?" (Miller 1983), in which he "explained" why the Kaplan–Meier method is inefficient and recommended, when possible, a parametric analysis—using, say, the exponential or Weibull distribution—instead. That article has been widely quoted, but we argue that although Miller is logically correct, there are compelling reasons for preferring the nonparametric approach.

In his introduction, Miller (1983, p. 1077) wrote:

> The product-limit [or Kaplan–Meier] estimator is attractive because it is easy to compute and understand. It has an asymptotic normal distribution with an estimated variance that is easily computed by Greenwood's formula. For the underlying probability structure, no assumptions are required other than the basic one of independence between the survival and censoring variables. In fact, the product-limit estimator is so seductive that there is a danger of becoming mentally lazy and not considering parametric modeling. Is there a price to be paid for this easy living?

Miller considered point estimates of the survival function when the true underlying distribution is exponential or Weibull, deriving asymptotic efficiencies for the Kaplan–Meier estimator relative to the parametric ones. He noted that "the asymptotic efficiencies of the Kaplan–Meier estimator are low, especially for high censoring proportions or for surviving fractions that are close to one or zero." He argued for the parametric approach, even though he agreed that the exponential (or Weibull) distribution may not be quite right:

> The efficiencies have been computed on the 'home turf' of the maximum likelihood estimator, that is, under the assumption that the parametric family of distributions has been correctly selected. But in practice where is the oracle that informs the statistician of the correct choice? In the limit, as $n \to \infty$, the mean square error for the Kaplan–Meier estimator will tend to zero, unlike that for the maximum likelihood estimator (Miller 1983, p. 1080).

Miller closed by alluding to studies under way of the robustness of the parametric approach when the assumed underlying model is incorrect, but unfortunately he died 2 years after the article was published. Such studies have since been undertaken by several authors, and we discuss the findings from two such studies (Klein and Moeschberger 1989; Aranda-Ordaz 1987). We also consider another aspect of the problem not addressed by Miller—one that we feel is more relevant—namely, estimation of a functional of the survival distribution, that is, the mean or, more suitably for censored data, the restricted mean (Irwin 1949; Kaplan and Meier 1958; Meier 1975; Karrison 1987).

We begin with a discussion of the noncensored case in Section 2. This is familiar territory, and we find that although the efficiency of the Kaplan–Meier estimator [equivalent to that of the empirical distribution function (edf)] can indeed be low in the exponential case for estimation of a point on the survival curve, the efficiency is essentially 1.0 for estimation of the mean. We then consider censored data in Section 3, first for point estimation and then for the restricted mean. The former includes a description of robustness findings reported in the literature, as well as a comparison of a more flexible (relative to other parametric models) modeling approach proposed by Efron (1988) with the Kaplan–Meier estimator. We provide a discussion and concluding remarks in Section 4.

2. NONCENSORED CASE

2.1 Point Estimate of Survival Function

Miller first considered the case of an underlying exponential distribution with exponential censoring and derived a formula for the asymptotic efficiency of the Kaplan–Meier estimator relative to the maximum likelihood estimator (MLE). If there are no censored data, then the relative efficiency can be obtained more simply, as follows. Let $T_1, T_2, \ldots, T_n$ denote independent and identically distributed survival times with distribution function $F(t)$ and survival function $S(t) = 1 - F(t)$. For uncensored data, the Kaplan–Meier estimator is just the complement of the edf, that is, $\hat{S}_{\mathrm{KM}}(t)$ is the proportion of pa-

Paul Meier is Special Lecturer, Columbia University, Department of Statistics, and Mailman School of Public Health, Department of Biostatistics and International Center for Health Outcomes and Innovation Research (InCHOIR) New York, NY 10032. Theodore Karrison is Research Associate (Associate Professor), Department of Health Studies, University of Chicago, Chicago, IL 60637. Rick Chappell is Professor, Departments of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53792. Hui Xie is Assistant Professor, Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118. The authors are grateful to the reviewers for their many helpful comments.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Review Article
DOI 10.1198/016214504000001259

tients with survival times exceeding t, and it has binomial variance $S(t)(1 - S(t))/n$. For the exponential survival distribution, $S_\lambda(t) = \exp(-\lambda t)$, and the likelihood function is given by

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda t_i), \tag{1}$$

where the t_i are the observed random survival times. From (1), the MLE for λ is easily found to be $\hat{\lambda} = 1/\bar{t}$, the reciprocal of the mean survival time, with asymptotic variance, given by the inverse of the Fisher information, equal to $\text{var}(\hat{\lambda}) = \lambda^2/n$. The parametric estimate of the survival rate at time t is then $\hat{S}_\lambda(t) = \exp(-\hat{\lambda} t)$, and the asymptotic variance of $\hat{S}_\lambda(t)$ can be computed using the delta method,

$$\text{var}(\hat{S}_\lambda(t)) = \left(\frac{dS_\lambda(t)}{d\lambda}\right)^2 \text{var}(\hat{\lambda}) = t^2(\exp(-\lambda t))^2 \frac{\lambda^2}{n}. \tag{2}$$

The asymptotic efficiency, *Eff*, of the edf (i.e., the Kaplan–Meier estimator) relative to the parametric estimator is then the ratio of (2) to the binomial variance

$$Eff = \frac{(\lambda t)^2}{\exp(\lambda t) - 1}. \tag{3}$$

At surviving fractions of .50, .25, and .10, the efficiencies are .48, .64, and .59. These are the same as the values given in the last row of Miller's table 1.

Thus the edf itself entails a loss in efficiency when considering a point estimate on the curve. This should not be unexpected, because it is ignoring information about the shape of the distribution function that the parametric estimator is exploiting. Of course, to be comfortable about using that information, we need to be reasonably sure that our parametric assumptions are correct.

2.2 Estimate of Mean Survival Time

When comparing two distributions, we ordinarily compare the means, $\bar{X}_1 - \bar{X}_2$. But when the life table and its standard error (SE) were first computed (Greenwood 1926), the SE of the mean was not considered; only the SE of the estimated survival function itself was included. Consequently, instead of comparing means, many people calculated the SE of the survival function at various places along the time scale and observed whether the rates were significantly different at some place. If this was indeed so, then they would declare the difference in survival to be "significant." (Of course, this did not allow for the selection of the point at which the difference, or z-statistic for the difference, was a maximum.) Some early work on this problem was done by Berger and Gold (1961) (see also Forsythe and Frey 1970), but was largely ignored in favor of alternative approaches comparing the overall survival curves, such as the Gehan–Wilcoxon test (Gehan 1965) and the log-rank test (Peto and Peto 1972).

Now it can be shown through a simple integration by parts that the mean, μ, is the area under the survival function, and that the sample mean is the area under the complement of the edf. Furthermore, the sample mean is also the MLE for the exponential distribution, so that in this case the efficiency of the estimator derived from the edf (i.e., the Kaplan–Meier curve) is in fact 1.0! So the type of index that one seeks to estimate matters a great deal.

The comparison does not end here, however, because although the estimated mean is the same in both instances, the estimated variances are different. For the parametric estimator, because $\hat{\mu} = 1/\hat{\lambda}$, the estimated variance (again applying the delta method) is

$$\hat{V}_\lambda(\hat{\mu}) = \hat{\mu}^2/n. \tag{4}$$

However, for the Kaplan–Meier estimator, the estimated variance in the uncensored case (see the Appendix) is given by

$$\hat{V}(\hat{\mu}) = \frac{1}{n^2} \sum_{i=1}^{n} (t_i - \bar{t})^2. \tag{5}$$

If the data are truly exponential, then $\sigma^2 = \mu^2$, and thus (4) and (5) have similar expectations. However, distributions different from the exponential can have lower or higher variances. For example, a uniform distribution with the same mean would have one-third the exponential variance. (Note that the Kaplan–Meier variance is the correct variance; the maximum likelihood result applies only to the exponential.)

Actually, in almost all cases we use the asymptotic normality of the mean (i.e., the central limit theorem) to analyze the data. Thus we find Fisher (1969), arguing that "even if the original distribution were not exactly normal, that of the mean usually tends to normality, as the size of the sample is increased; the method is therefore applied widely and legitimately to cases in which we have not sufficient evidence to assert that the original distribution was normal. . . ." Later others introduced a number of transformations to reduce skewness and stabilize the variance, such as the square root and logarithmic transformations. The data analyzed by Fisher on the difference in hours of sleep gained by a soporific (see Fisher 1969, table 27) are quite clearly skewed to the right, but Fisher carried out the analysis without any comment on it. He depended on the central limit theorem, which had been adopted by Gauss (1821) much earlier. (We hope that these are adequate "oracles.") A further problem with the exponential is that other distributions are very close to it. For example, the lognormal imitates it quite closely at the start, and compares well with it up to a surviving fraction of .4 (see Fig. 1). In that case, of course, after a logarithmic transformation, the MLE is once again the sample mean, with the usual form for the variance, $\frac{1}{n^2}\sum_{i=1}^{n}(y_i - \bar{y})^2$, where $y_i = \log t_i$.

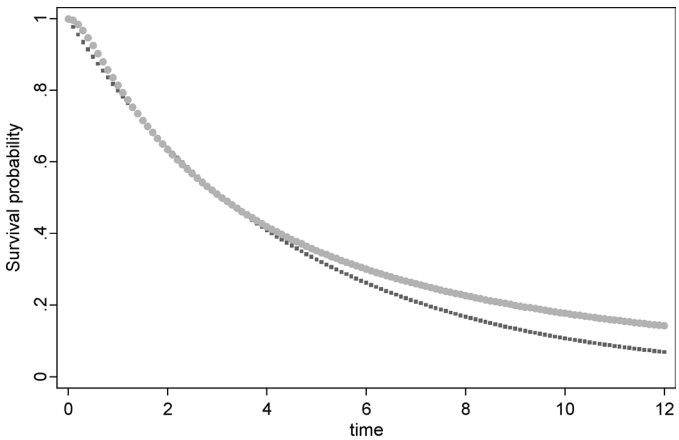

Figure 1. Exponential [$\mu = \exp(1.50)$] (•) and Lognormal ($\mu = 1.13$, $\sigma = 1.26$) (•) Survival Distribution Functions.

In summary, it seems clear to us that although the parametric estimator may offer an advantage for estimating the survival rate at a given point in time, the nonparametric approach is preferred for estimating the mean when the data are complete.

Now let us see what censoring does.

3. CENSORED CASE

3.1 Point Estimate of Survival Function

The observed random variables are $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$, where the C_i are independent and identically distributed censoring times assumed to be independent of the survival times. The Kaplan–Meier estimator is then given by $\hat{S}_{\mathrm{KM}}(t) = \prod_{t_{(i)} \le t}(1 - d_{(i)}/n_{(i)})$, where $t_{(i)}$ is the time of the ith ordered death, $d_{(i)}$ is the number of deaths, and $n_{(i)}$ is the number of patients at risk at time $t_{(i)}$. Computation of the MLE, at least for the censored exponential case, is comparatively simple: $\hat{S}_\lambda(t) = \exp(-\hat{\lambda}t)$, where $\hat{\lambda} = (\sum_{i=1}^n \delta_i / \sum_{i=1}^n t_i)$, the number of deaths per person-years (depending on the time units) of follow-up. The efficiency results were derived by Miller (1983) and essentially parallel those for the uncensored case. The efficiency of the product-limit estimator depends on both the point at which the survival function is being estimated and the degree of censoring. For example, in the exponential case (with exponential censoring), the efficiency at a modest level of censoring is .56 at a surviving fraction of .50 and .46 at a surviving fraction of .10; whereas for relatively heavy censoring, these efficiencies are .64 and .21. In the Weibull case, $S(t) = \exp(-(\lambda t)^\alpha)$, with exponential censoring and estimation of both the index (α) and scale (λ) parameters, the efficiencies are somewhat greater, but still not high. We acknowledge that for point estimation, the efficiency of the nonparametric approach can be low relative to a fully parametric procedure, and this will translate into higher mean squared error (MSE) when the assumed form of the distribution is correct. But if the fitted model is incorrect, then the bias could be substantial. Although hazard plotting and other graphical methods can guide the choice of parametric distribution, one cannot of course be sure that the right one has been selected or easily adjust the estimated standard error for this uncertainty.

Klein and Moeschberger (1989) conducted a rather exhaustive robustness study for small and small-to-moderate-sized samples. They examined the results of fitting exponential, Weibull, normal, lognormal, exponential power, log-logistic, Pareto, and Gompertz models to data arising from various true underlying hazard rates corresponding to each of these families as well as a bathtub-shaped hazard function. They compared the MSE from the parametric fit with that obtained using the Kaplan–Meier estimator for small and moderate-sized samples with varying degrees of censoring. They considered the MSE integrated over the 0–95th percentiles of the distribution, as well as at selected points along the curve. As might be expected, the exponential MLE was not robust and thus usually performed poorly relative to the Kaplan–Meier estimator when the data were not drawn from a distribution with constant underlying hazard. The results for the other models were mixed, but in most cases the parametric estimator outperformed the nonparametric estimator, and the authors concluded that "a statistician armed with the Weibull, log-logistic and exponential power distribution MLEs can provide better estimates ... than one armed only with the Kaplan–Meier estimator."

Two points about these findings should be noted, however. First, the Weibull distribution did not perform well when the underlying distribution was normal or lognormal and the sample size was 50, or when the data were generated from a log-logistic distribution; the exponential power distribution did not perform well for normal, lognormal, or log-logistic data; and the log-logistic distribution did poorly when the underlying distribution was exponential power and the sample size was 50. Hence the suggested three-fold modeling strategy and the need to carefully assess fit. Second, at the sample sizes studied, it is likely that the variance dominated the squared bias. As the sample size increases, the advantages of parametric modeling will decrease, because the Kaplan–Meier estimator is essentially unbiased, whereas the parametric estimator is not (under the wrong model). Because censoring decreases the effective sample size, the relative performance of the Kaplan–Meier estimator would also be expected to decrease with increased censoring, as Klein and Moeschberger also noted. The issue is complicated, however, by the fact that in their simulations, if the longest observation $t_{\max}$ was censored, then the value $\hat{S}_{\mathrm{KM}}(t_{\max})$ was imputed for $t > t_{\max}$ in the Kaplan–Meier estimator. [An alternative choice, suggested in Efron (1967), is to define $\hat{S}_{\mathrm{KM}}(t) = 0$ for $t > t_{\max}$.] This leads to a bias in the Kaplan–Meier estimator in the tail of the distribution, affecting the MSE in this region as well as the integrated MSE, an effect that becomes more pronounced as the degree of censoring increases. In our view it is preferable to simply leave the estimator undefined in such cases.

Aranda-Ordaz (1987) examined the robustness of the Kaplan–Meier estimator relative to exponential and Weibull MLEs under corresponding exponential and Weibull mixture models for the true survival distribution, when the mixing is ignored in the model fitting. Both small and large sample sizes and censoring proportions of 0, .25, and .50 were considered. The mixing proportion was varied from .05 (reflecting mild contamination with outliers) to .40 (heavy mixing). The MSEs for estimation of the survival curve for the parametric and Kaplan–Meier estimators were then compared.

The MSE following exponential fitting was lower than that for Kaplan–Meier in all situations except the large-sample case ($n = 500$) without censoring and at the higher surviving fractions and mixing proportions. This is in accord with the expectation that the bias would eventually dominate the MSE as the effective sample size increases. Interestingly, under Weibull mixtures, the results were less in favor of parametric fitting. At $n = 500$ and moderate levels of mixing the Kaplan–Meier estimator almost always performed better. But at $n = 25$ and $n = 50$ and 5% contamination, the parametric procedure was clearly preferable. In other situations the results were mixed, although in most of the configurations the MSE from the parametric approach was less than that for Kaplan–Meier. It should be noted that in the simulations the index parameter for the Weibull distribution was fixed. As Aranda-Ordaz mentioned, if this parameter was also estimated, then the relative performance of the parametric procedure would decrease.

Suppose that we take a middle-ground approach, such as the "partial logistic regression" technique proposed by Efron

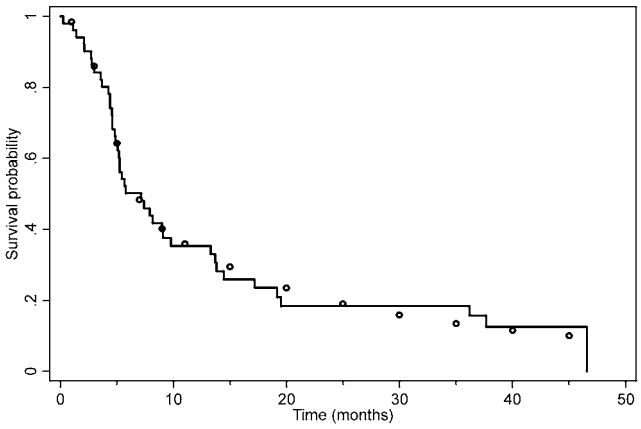

Figure 2. Cubic-Linear Spline Estimator (○) versus Kaplan–Meier (——), Treatment Arm A (from Efron 1988).

(1988), which he described as "parametric regression modeling on censored data in a flexible way that provides both estimates and standard errors." The example that Efron used involves a moderate degree of censoring occurring mostly toward the end of the observation period. The data are from a randomized clinical trial in head and neck cancer patients comparing radiation therapy (arm A) with radiation therapy and chemotherapy (arm B). A total of 51 patients were entered onto arm A of the trial and 45 onto arm B. This is a good example, because the data were analyzed with alternative models and because Efron gave the individual survival times for censored and noncensored observations. Efron suggested dividing the time axis into intervals and fitting a logistic regression model to the discrete hazard rates. For arm A of the trial, Efron's preferred model is a cubic-linear spline model, for which the logit is a cubic function of time before 11 months and a linear function thereafter. This model has five parameters when the join point is considered, and Efron derived the standard error for the estimated hazard rates that includes an additional error term for estimating the join. He showed rather nicely that this estimator provides a good fit to the data and is noticeably less variable than nonparametric estimates of the hazard rate, which are known to behave poorly.

The estimated survival function using Efron's method is obtained as a product of the conditional probabilities and is shown in Figure 2 for arm A, along with the Kaplan–Meier curve. The two curves are seen to be quite close. What is surprising, however, is the fact that the SEs for the two sets of estimates are also fairly similar (Table 1), with the Kaplan–Meier estimate exhibiting no more than a 25% increase (usually much lower) and in some cases a decrease in the SE relative to that for the cubic-linear spline model. (As in Efron 1988, here we compare SEs for the estimated log survival function rather than the survival function itself, to remove the factor $\hat{S}(t)$ from the standard error calculation.) Thus fitting just five parameters in the parametric model eliminates much of the gain in efficiency it achieves over the nonparametric approach. Even a four-parameter cubic model, which according to Efron did not fit the data adequately, resulted in only slightly smaller SEs. Note that although it is impossible to estimate the hazard rate fully nonparametrically, the log survival curve is an integral of the hazard function (i.e., the negative of the cumulative hazard). The effect of the integration or smoothing is to bring the efficiency of the nonparametric estimator relative to the parametric procedure closer to 1. Incidentally, Efron provided his own version of the life table estimate, noting that the SEs for the parametric estimates were only slightly smaller than those for the life table estimates. Figure 3 shows a comparison of Efron's life table estimates with the Kaplan–Meier estimates for both treatment arms in the cancer study. The two curves are virtually identical, so that Efron's report that the life table estimate does quite well means that the Kaplan–Meier estimate also does quite well.

3.2 Estimate of Restricted Mean

If the data are subject to censoring, then it may not be possible, or even desirable, to estimate the mean survival time. For example, as discussed earlier, if the longest observation is censored, then the Kaplan–Meier estimator is not really defined beyond this point. To accommodate censoring, Irwin (1949) proposed estimating the mean lifetime restricted to a suitably chosen time T, that is,

$$\mu_T = \int_0^T S(t)\,dt. \tag{6}$$

Irwin used the actuarial survival estimator for $S(t)$ in (6) to estimate μ_T and derived a formula for its variance. Meier (1975) showed that substitution of the product-limit estimator, $\hat{S}_{\mathrm{KM}}(t)$, in (6) provides an unbiased estimate that is asymptotically normal. Karrison (1987) extended the method to incorporate covariates into the analysis under a piecewise exponential model.

Table 1. Estimated Survival Functions and Standard Errors for Head and Neck Cancer Trial, Treatment Arm A

	Estimated survival		*Estimated standard errors for log survival*		
Month	*Cubic-linear*	*Kaplan–Meier*	*Cubic-linear*	*Kaplan–Meier*	*% difference*
1	.985	.980	.031	.020	−36.2
3	.860	.842	.059	.061	3.3
5	.642	.641	.090	.106	17.5
7	.483	.501	.129	.141	9.4
9	.402	.397	.155	.176	13.4
11	.359	.354	.186	.194	4.1
15	.295	.259	.256	.248	−3.0
20	.235	.183	.286	.320	11.9
25	.191	.183	.309	.320	3.6
30	.159	.183	.319	.320	.3
35	.134	.183	.324	.320	−1.2
40	.115	.126	.338	.420	24.2
45	.100	.126	.436	.420	−3.7

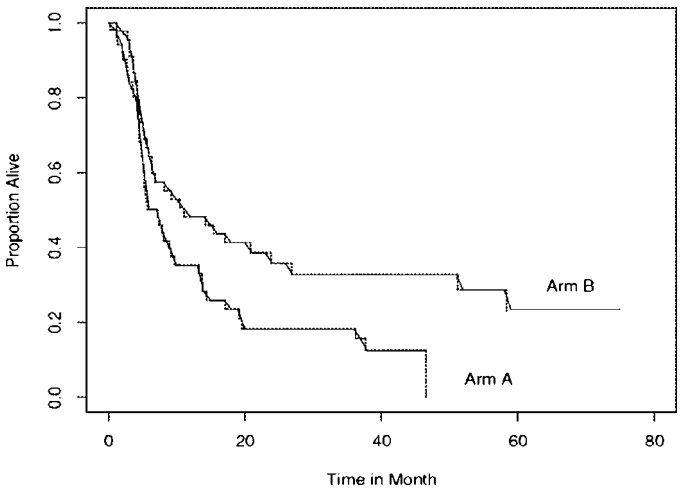

Figure 3. Comparison of Kaplan–Meier (·····) and Efron's Life Table Estimates (——) for Head and Neck Cancer Study, Treatment Arms A and B.

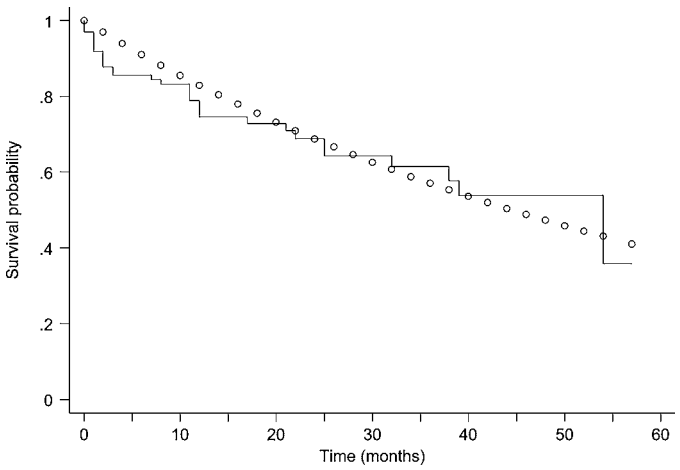

Figure 4. Survival After Diagnosis of SLE, Exponential (o) versus Kaplan–Meier (——).

Meier provided a formula for the variance of $\hat{\mu}_T$, but a simpler version was given by Klein and Moeschberger (2003). As before, let $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ denote the D ordered distinct death times, and let $d_{(i)}$ denote the number of deaths and $n_{(i)}$ denote the number of individuals remaining at risk at time $t_{(i)}$. Then

$$\hat{V}(\hat{\mu}_T) = \sum_{i=1}^{D} \left[\int_{t_{(i)}}^{T} \hat{S}_{\mathrm{KM}}(t)\, dt \right]^2 \frac{d_{(i)}}{n_{(i)}(n_{(i)} - d_{(i)})}. \tag{7}$$

Now in the exponential case,

$$\mu_T = \int_0^T \exp(-\lambda t)\, dt = \frac{1}{\lambda}\big(1 - \exp(-\lambda T)\big), \tag{8}$$

and the parametric estimate can be obtained by replacing λ in (8) with $\hat{\lambda}$. An estimate for the variance via the delta method is then

$$\hat{V}_\lambda(\hat{\mu}_T) = \left[\frac{\exp(-\hat{\lambda}T)(1 + \hat{\lambda}T) - 1}{\hat{\lambda}^2} \right]^2 V_\lambda(\hat{\lambda}). \tag{9}$$

If we have minimal censoring, then the restricted mean (6) with T set to the longest observed survival time will be either a valid estimate of the mean itself (if the longest observation corresponds to a death) or only a slight underestimate (if the longest observation corresponds to a censoring). In this situation, the relative efficiency should be close to 1.0, and the nonparametric estimate is preferable for the reasons discussed in Section 2.2.

In the more typical study we will have moderate or even heavy censoring, and for this case we consider an example that goes way back—predating even Kaplan–Meier—to an article by Merrell and Shulman (1955) examining the prognosis of patients with systemic lupus erythematosus (SLE). A total of 99 patients were followed from the date of diagnosis until the date of death (31 observations) or date last known alive (68 observations). The Kaplan–Meier survival curve from this study is shown in Figure 4. The 4-year survival rate just barely exceeds 50%, at which time only four subjects remained at risk, with respective subsequent survival times of 53^+, 54, 56^+, and 57^+ months (where $+$ denotes censoring). Thus the survival rate at the limit of the observation period is far from 0, and estimating the overall mean survival time is out of the question. We could seek to estimate the restricted mean at $T = 57$ months, but because of the censoring, the SE of the survival rate is fairly high at this point (15.5%). To ensure that the SE of the survival estimate is less than 10%, the number of subjects remaining at risk should be at least 13 (see Karrison 1997). This occurs at 39 months, but for convenience we consider estimation of the restricted mean at $T = 36$ months, or 3 years.

The estimate of mean lifetime restricted to 36 months using the Kaplan–Meier curve is $\hat{\mu}_T = 26.7$ months with a SE, obtained from (7), of 1.50 months. (One interpretation of this index is that the SLE patients achieved $26.7/36 = 74.2\%$ of total possible life-years over this time period.) If one proceeds parametrically, note that an exponential model does not provide a very good fit to the data (see Fig. 4), although fitting a Weibull model with two parameters does not significantly increase the log-likelihood. In any case, substituting the estimated scale parameter, $\hat{\lambda} = .0156$, into (8) gives $\hat{\mu}_T = 27.5$ months for the estimated mean life restricted to 3 years. Of most interest here, however, is the fact that the SE based on (9) is 1.33, just slightly less than that for the nonparametric estimator. So in this example, only little loss in efficiency results from using the Kaplan–Meier procedure to estimate the restricted mean. Note that, analogous to shifting from estimating the hazard rate to the (log) survival curve discussed in Section 3.1, estimating the restricted mean involves a second integration, this time of the survival function itself.

The foregoing example suggests that with censored data, the efficiency of the Kaplan–Meier procedure for estimating the restricted mean should be quite good. We addressed this issue more systematically by performing a simulation study. Following Miller (1983), we generated 50 survival times, T_i, from an exponential distribution with scale parameter λ. We generated independent random censoring times, C_i, from an exponential distribution with scale parameter ν and took the observed survival time as $t_i = \min(T_i, C_i)$. The expected proportion of observations censored by this mechanism, the censoring proportion (CP), is then given by $\mathrm{CP} = \Pr(C_i < T_i) = \rho/(1 + \rho)$,

Table 2. Efficiency of the Estimation of Restricted Means by the Kaplan–Meier Method Relative to the MLE for the Exponential Distribution

	SF		
CP	.50	.25	.10
.50	.92	.97	.87
.25	.86	.93	.98
0	.83	.90	.94

NOTE: CP, censoring proportion; SF, surviving fraction.

where $\rho = \nu/\lambda$. For convenience, we set λ to 1 and considered CPs of 0, .25, and .50 ($\nu = 0$, $1/3$, and 1). Mean life was restricted to values of T corresponding to surviving fractions (SFs) of .50, .25, and .10 ($T = .693$, 1.39, and 2.30). Any observations with $t_i > T$ were then "administratively" censored at time T, so that even for the case with CP = 0, we are still dealing with a restricted mean.

We ran a total of 10,000 simulations for each combination of CP and SF. Computation was performed in S–PLUS. For each generated dataset, the restricted mean was estimated nonparametrically using (6), substituting the Kaplan–Meier estimator for the survival function, and parametrically using (8), substituting the MLE for the scale parameter. The variances of these estimates over the 10,000 simulations were then calculated and the relative efficiency derived as the ratio of the variance of the parametric estimate to the nonparametric estimate. The results are given in Table 2.

The efficiencies are seen to be fairly high, ranging from .83 to .98. For SFs of .50 and .25, the efficiency decreases with decreasing CP, whereas at SF = .10, this relationship does not hold. Overall, the results indicate that one does not give up very much by taking the nonparametric approach.

4. DISCUSSION

It is well known that estimation of the hazard function presents a number of difficulties, just as does estimation of the density function for uncensored data. Although several nonparametric techniques have been proposed in the literature, most based on some form of kernel smoothing (see, e.g., Tanner and Wong 1984; Efron 1988), the hazard function itself has rarely been used to draw inferences or to make comparisons, say, between two treatment groups. The attractiveness of the widely used Cox (1972) regression model for assessing the effects of treatments or covariates is that it avoids estimation of the underlying hazard rate.

The survival curve is a more friendly target, because, being a function of the cumulative hazard function, it has already undergone one smoothing operation, that of integration. Although estimating the survival curve nonparametrically entails a loss in efficiency, application of Efron's partial logistic regression technique suggests that this loss becomes small fairly quickly as the number of parameters in the parametric model increases. Thus, when facing the bias–variance trade-off, the choice comes down to deciding whether a model with very few parameters adequately fits the data. How "adequate" the fit needs to be will depend on the sample size. Robustness studies indicate that for small to moderate sample sizes, particularly when censoring is heavy, the parametric approach can offer advantages relative to Kaplan–Meier in terms of MSE, although this advantage diminishes with increasing sample size. Miller's assertion that estimation of survival rates should be performed via parametric modeling has perhaps not been given sufficient consideration in practice, but its applicability may be more limited than a casual reading of Miller's article would imply.

There are further considerations beyond balancing bias with variance that weigh in favor of the Kaplan–Meier estimator. First, the variance is always estimable, whereas bias is not. Using diagnostics to assess the fit of a parametric model and then conditioning on the chosen model is problematic in the sense that there is no accounting for the (possibly erroneous) choice of distributions. Therefore, even in those situations of moderate misspecification in which a parametric fit has higher efficiency than the Kaplan–Meier curve, the parametric estimate of MSE (as equal to the variance) may be biased downward. This may lead to anticonservative conclusions, which in some situations can be disastrous. Second, certain features of a distribution (e.g., hazard peaks at specific times or multimodality) may not be fittable using existing parameterizations. This difficulty is surmountable but still inconvenient. Third, although the Kaplan–Meier curve's small-scale roughness accounts for much of its variability, the human eye tends to smooth it out. (If it did not, then one could easily connect the midpoints of the vertical drops to render a continuous curve.) An exception may be when the data are very sparse (see Lee et al. 2002 for an example), although even here the degree of fit of the model would need to be assessed from external sources. We also acknowledge that parametric models are useful for performing sample-size calculations, where exponential (or more general Weibull) survival is frequently assumed. When estimating an integral of the distribution function, resulting in a weighted average, the nonparametric approach is superior. In the uncensored case, the area under the Kaplan–Meier curve is the same as the sample mean and thus is fully efficient, at least in the exponential and normal cases. Furthermore, we have seen that for censored data, nonparametric estimation of the restricted mean involves a relatively minor loss in efficiency. Note that the restricted mean involves *two* integration steps, first of the hazard function and then of the survival curve itself, which drives the efficiency of the nonparametric estimation procedure closer to 1. Given its unbiasedness, we feel that it is hard to argue for a parametric alternative when considering functionals of the survival curve such as the mean or restricted mean.

APPENDIX: VARIANCE OF MEAN SURVIVAL TIME

Show that $\hat{V}(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n}(t_i - \bar{t})^2$ in the uncensored case. A convenient form for the variance of the nonparametric estimate of the restricted mean was given by Klein and Moeschberger (2003), eq. (4.5.2),

$$\hat{V}(\hat{\mu}_T) = \sum_{i=1}^{D}\left[\int_{t_i}^{T}\hat{S}(t)\,dt\right]^2 \frac{d_i}{n_i(n_i - d_i)}, \qquad \text{(A.1)}$$

where D is the total number of deaths, t_i is now the time of the ith ordered death, d_i is the number of deaths, and n_i is the number of patients at risk at time t_i. Note that in the uncensored case $D = n$ and $T = t_n$. Assuming no ties and working from right to left (adding the

horizontal slices of the survival curve rather than the vertical), we have

$$\begin{aligned}\hat{V}(\hat{\mu}) &= \left[(t_n - t_{n-1})\frac{1}{n}\right]^2 \frac{1}{2(1)} \\ &\quad + \left[(t_n - t_{n-2})\frac{1}{n} + (t_{n-1} - t_{n-2})\frac{1}{n}\right]^2 \frac{1}{3(2)} \\ &\quad + \left[(t_n - t_{n-3})\frac{1}{n} + (t_{n-1} - t_{n-3})\frac{1}{n}\right. \\ &\qquad \left. + (t_{n-2} - t_{n-3})\frac{1}{n}\right]^2 \frac{1}{4(3)} + \cdots \\ &\quad + \left[(t_n - t_1)\frac{1}{n} + (t_{n-1} - t_1)\frac{1}{n} + \cdots + (t_2 - t_1)\frac{1}{n}\right]^2 \\ &\qquad \times \frac{1}{n(n-1)} \\ &= \frac{1}{n^2}\left[(t_n - t_{n-1})^2 \frac{1}{2} + (t_n + t_{n-1} - 2t_{n-2})^2 \frac{1}{2(3)}\right. \\ &\qquad + (t_n + t_{n-1} + t_{n-2} - 3t_{n-3})^2 \frac{1}{3(4)} + \cdots \\ &\qquad + \big(t_n + t_{n-1} + \cdots \\ &\qquad \left. + t_2 - (n-1)t_1\big)^2 \frac{1}{(n-1)n}\right]. \end{aligned} \tag{A.2}$$

The result will be established if it can be shown that the term in square brackets is equal to

$$\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2 \Big/ n = \frac{(n-1)}{n}\sum_{i=1}^{n} t_i^2 - \frac{2}{n}\sum_{i<j} t_i t_j.$$

Thus we need to show that the coefficient of each squared term in (A.2) is $\frac{(n-1)}{n}$ and the coefficient of each cross-product term is $-\frac{2}{n}$. The proof makes repeated use of the fact that $\sum_{i=1}^{k} \frac{1}{i(i+1)} = \frac{k}{(k+1)}$, which can be shown by a simple induction argument.

Now the coefficient of t_n^2 in (A.2) is

$$\begin{aligned}\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \cdots + \frac{1}{n(n-1)} &= \sum_{i=2}^{n} \frac{1}{(i-1)i} \\ &= \sum_{i=1}^{n-1} \frac{1}{i(i+1)} = \frac{(n-1)}{n},\end{aligned}$$

and the same is true for t_{n-1}^2. The coefficient of the general t_{n-i}^2 term, for $i \geq 2$, is

$$\frac{i^2}{i(i+1)} + \sum_{j=i}^{n-2} \frac{1}{(j+1)(j+2)} = \frac{i}{(i+1)} + \sum_{j=i+1}^{n-1} \frac{1}{j(j+1)}.$$

Because $\frac{i}{i+1} = \sum_{j=1}^{i} \frac{1}{j(j+1)}$, this expression reduces to $\sum_{j=1}^{n-1} \frac{1}{j(j+1)} = \frac{(n-1)}{n}$.

The general cross-product term in (A.2) ($i < j$, $i = 0, 1, 2, \ldots, n-2$) is

$$\begin{aligned}&2t_{n-i}t_{n-j}\left(\frac{-j}{j(j+1)} + \sum_{k=j}^{n-2} \frac{1}{(k+1)(k+2)}\right) \\ &\quad = 2t_{n-i}t_{n-j}\left(-\frac{1}{(j+1)} + \sum_{k=j+1}^{n-1} \frac{1}{k(k+1)}\right) \\ &\quad = 2t_{n-i}t_{n-j}\left(-\frac{1}{j+1} + \frac{(n-1)}{n} - \sum_{k=1}^{j} \frac{1}{k(k+1)}\right) \\ &\quad = 2t_{n-i}t_{n-j}\left(-\frac{1}{j+1} + \frac{(n-1)}{n} - \frac{j}{j+1}\right) \\ &\quad = 2t_{n-i}t_{n-j}\left(-\frac{1}{n}\right),\end{aligned}$$

and the proof is complete.

[Received June 2003. Revised January 2004.]

REFERENCES

Aranda-Ordaz, F. J. (1987), "Relative Efficiency of the Kaplan–Meier Estimator Under Contamination," *Communications in Statistics—Simulation and Computation*, 16, 987–997.

Berger, A., and Gold, R. Z. (1961), "On Comparing Survival Times," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 4, 67–76.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

Efron, B. (1967), "The Two-Sample Problem With Censored Data," in *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics IV*, New York: Prentice-Hall, pp. 831–853.

——— (1988), "Logistic Regression, Survival Analysis, and the Kaplan–Meier Curve," *Journal of the American Statistical Association*, 83, 414–425.

Fisher, R. A. (1969). *Statistical Methods for Research Workers*, Edinburgh, U.K.: Oliver and Boyd.

Forsythe, A. B., and Frey, H. S. (1970), "Tests of Significance From Survival Data," *Computers and Biomedical Research*, 3, 124–132.

Gauss, C. F. (1821), "Theory of the Combination of Observations Least Subject to Errors," translated by G. W. Stewart, in *Society for Industrial and Applied Mathematics Classics in Applied Mathematics*, Philadelphia: SIAM, 1995.

Gehan, E. A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, 52, 203–223.

Greenwood, M. (1926), "The Natural Duration of Cancer," in *Reports on Public Health and Medical Subjects*, No. 33, London: Her Majesty's Stationery Office, pp. 1–26.

Irwin, J. O. (1949), "The Standard Error of an Estimate of Expectational Life," *Journal of Hygiene*, 47, 188–189.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Karrison, T. (1987), "Restricted Mean Life With Adjustment for Covariates," *Journal of the American Statistical Association*, 82, 1169–1176.

——— (1997), "Use of Irwin's Restricted Mean as an Index for Comparing Survival in Different Treatment Groups—Interpretation and Power Considerations," *Controlled Clinical Trials*, 18, 151–167.

Klein, J. P., and Moeschberger, M. L. (1989), "The Robustness of Several Estimators of the Survivorship Function With Randomly Censored Data," *Communications in Statistics—Simulation and Computation*, 18, 1087–1112.

——— (2003), *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.), New York: Springer-Verlag.

Lee, A. W. M., Tung, S. Y., Kwong, D. L. W., Teo, P. M. L., Sze, W. M., Leung, T. W., Wu, P. M., Leung, S. F., Chappell, R., Peters, L. J., and Fowler, J. F. (2002), "Factors Affecting the Risk of Symptomatic Temporal Lobe Necrosis: Significance of Fractional Dose and Treatment Time," *International Journal of Radiation Oncology Biology Physics*, 53, 75–85.

Meier, P. (1975), "Estimation of a Distribution Function From Incomplete Observations," in *Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett*, ed. J. Gani, New York: Academic Press, pp. 67–87.

Merrel, M., and Shulman, L. E. (1955), "Determination of Prognosis in Chronic Disease, Illustrated by Systemic Lupus Erythematosus," *Journal of Chronic Diseases*, 1, 12–32.

Miller, R. G., Jr. (1981), *Survival Analysis*, New York: Wiley.

——— (1983), "What Price Kaplan–Meier?" *Biometrics*, 39, 1077–1081.

Peto, R., and Peto, J. (1972), "Asymptotically Efficient Rank Invariant Test Procedures" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 135, 185–206.

Tanner, M. A., and Wong, W. H. (1984), "Data-Based Nonparametric Estimation of the Hazard Function With Applications to Model Diagnostics and Exploratory Analysis," *Journal of the American Statistical Association*, 79, 174–182.