STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2091–2097 (DOI: 10.1002/sim.856)

Projecting the standard error of the Kaplan–Meier estimator

Alan B. Cantor*,†

H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, U.S.A.

SUMMARY

Clinical studies in which a major objective is to produce Kaplan–Meier estimates of survival probabilities should be designed to produce those estimates with a desired prespecified precision as measured by their standard errors. By considering the Peto and Greenwood formulae for the estimated standard error of the Kaplan–Meier estimate and replacing their constituents with expected values based on the study's design parameters, formulae for projected standard errors can be produced. These formulae are shown, through simulations, to be quite accurate. Copyright © 2001 John Wiley & Sons, Ltd.

INTRODUCTION

It is now well established that clinical trials whose objectives call for some type of formal statistical hypothesis testing should have a sample size which is justified by the desire to have adequate (often 0.80 or 0.90) power for alternatives of a given magnitude. For studies whose objectives involve estimation rather than hypothesis testing, the sample size should be adequate to produce estimates of desired precision. This precision is generally expressed in terms of the estimated standard error of the estimate or the width of an estimated (perhaps 95 per cent) confidence interval for the true value of the quantity being estimated.

When a mean or proportion is being estimated, the relationship between the standard error of the estimate and the sample size is fairly simple, thus allowing for easy calculation of the sample size needed to ensure a desired standard error. However, when a survival probability at a specified time is being estimated and some survival times are right censored, the situation is much more complicated. By far the most common method of obtaining survival function estimates is the method of Kaplan and Meier [1]. A formula due to Greenwood [2] is usually used to estimate the standard error of the Kaplan–Meier estimate although Peto *et al.* [3] provide a simpler (and more conservative) alternative. In both cases, the estimated standard error depends not only on the number of patients in the study, but also on the accrual time, follow-up time, the actual survival function, and losses to follow-up. Perhaps because of the complexity of this relationship, clinical trial planners do not generally attempt to justify a sample size for a trial by the desire to achieve a sufficiently small standard error for a survival rate estimate.

*Correspondence to: Alan B. Cantor, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, U.S.A.
†E-mail: abcantor@moffitt.usf.edu

Received January 2000
Accepted September 2000

By the projected variance (or standard error) estimate we mean a value, based on a study's planning parameters and calculated prior to the collection of data, that is a prediction of the value of that variance (or standard error) estimate after the study is completed. Such a projection would allow planners to design studies which would produce Kaplan–Meier estimates of desired precision. At this time, methods of projecting the standard error of Kaplan–Meier estimates over time have not been reported and such calculations are not done by planners of clinical trials. In the discussion that follows, we will distinguish estimated variances and standard errors, calculated from study data, from their projections, calculated before data are collected and based on the study parameters, by using the 'hat' notation (as in $\widehat{\mathrm{var}}(t)$) for the former and 'wave' notation (as in $\widetilde{\mathrm{var}}(t)$) for the latter.

In this paper, methods of projecting both the Peto and Greenwood standard error estimates of Kaplan–Meier survival function estimates will be developed and their properties discussed. These methods will assume uniform accrual of patients to a clinical trial at rate r per unit of time over time T, additional follow-up for a time τ, and losses to follow-up at a rate given by $\theta(t)$. The probability of not being lost to follow-up at time t is given by $U(t) = \exp[-\int_0^t \theta(t)\,\mathrm{d}t]$. Further, it will be assumed that mortality is subject to the hazard function $\lambda(t)$, so that the actual survival function is given by $S(t) = \exp[-\int_0^t \lambda(t)\,\mathrm{d}t]$ and that loss to follow-up and mortality are independent events. We also assume that both loss to follow-up and mortality are independent of the time at which a patient entered the study.

THE PETO STANDARD ERROR

The Peto formula for the estimated standard error of the Kaplan–Meier estimate at time t will be considered first. This standard error estimate is given by $\widehat{\mathrm{SE}}_1(t) = [\widehat{\mathrm{var}}_1(t)]^{1/2}$ where

$$\widehat{\mathrm{var}}_1(t) = \frac{\hat{S}^2(t)[1 - \hat{S}(t)]}{N(t)} \tag{1}$$

Here $\hat{S}(t)$ is the Kaplan–Meier estimator. For a projection, it can be replaced in (1) by the expected value, $S(t)$ of $\hat{S}(t)$. $N(t)$ is the number of patients at risk at time t. For a patient to be at risk at time t, each of the following three events must be true:

1. The patient's survival time must exceed t. The probability of this is $S(t)$.
2. The patient's time of loss to follow-up must exceed t. That probability is $U(t)$.
3. The patient must have entered the study at least t units of time prior to $T + \tau$. The probability of this is 1 if $t < \tau$ and $(T + \tau - t)/T$ if $t \geqslant \tau$.

Multiplying the product of the above probabilities by the number of patients on the study, we obtain the expected value of $N(t)$. Replacing $\hat{S}(t)$ and $N(t)$ of (1) by their expected values, we obtain as the projected value of $\mathrm{var}_1(t)$

$$\widetilde{\mathrm{var}}_1(t) = \begin{cases} \dfrac{S(t)[1 - S(t)]}{rTU(t)} & \text{if } t < \tau \\ \dfrac{S(t)[1 - S(t)]}{r(T + \tau - t)U(t)} & \text{if } t \geqslant \tau \end{cases} \tag{2}$$

Note that the numerators of (2) are maximized when $S(t)=0.5$, so that the use of that value will always produce a conservative projection. The conservatism is moderate if $0.3<S(t)<0.7$. Of course, this mirrors the situation when projecting the standard error of a proportion.

In many cases, it will be reasonable to consider $\theta(t)$ to be a constant, θ. In that case (2) becomes

$$\widetilde{\mathrm{var}}_1(t)=\begin{cases}\dfrac{S(t)[1-S(t)]}{rT\exp(-\theta t)} & \text{if } t<\tau\\ \dfrac{S(t)[1-S(t)]}{r(T+\tau-t)\exp(-\theta t)} & \text{if } t\geqslant\tau\end{cases} \tag{3}$$

THE GREENWOOD STANDARD ERROR

The Greenwood formula for the estimated standard error of the Kaplan–Meier estimate at time t is given by $\widehat{\mathrm{SE}}_2(t)=[\widehat{\mathrm{var}}_2(t)]^{1/2}$ where

$$\widehat{\mathrm{var}}_2(t)=\hat{S}^2(t)\sum_{t_i\leqslant t}\frac{d_i}{N(t_i)[N(t_i)-d_i]} \tag{4}$$

Here t_i is the ith ordered study time, $d_i=0$ if t_i is a time at which the patient is still living (censored) and 1 otherwise, and $N(t)$ is the number at risk at time t.

In order to project the value of (4), we begin by replacing the observed times, t_i, in (4) by a partition of the interval $[0,t]$ given by $0=x_0<x_1<\cdots<x_n=t$ where each $x_i-x_{i-1}=\Delta x$. The d_i can now be thought of as the number of deaths in the interval $[x_{i-1},x_i)$ and $N(x_i)$ as the number at risk at time x_i. We next replace each of these by approximations of their expected value. The expected value of d_i is approximated by $\lambda(x_i)N(x_i)\Delta x$. The expected value of $N(x_i)$ is $rTS(x_i)U(x_i)$ if $x_i<\tau$ and $rS(x_i)U(x_i)(T+\tau-x_i)$ if $x_i\geqslant\tau$. Thus the projected value of the right side of (4) is given by

$$S^2(t)\left\{\frac{1}{rT}\sum_{x_i<t^*}\frac{\lambda(x_i)\Delta x}{S(x_i)U(x_i)[1-\lambda(x_i)\Delta x]}+\frac{1}{r}\sum_{\tau\leqslant x_i\leqslant t}\frac{\lambda(x_i)\Delta x}{S(x_i)U(x_i)(T+\tau-x_i)[1-\lambda(x_i)\Delta x]}\right\} \tag{5}$$

where $t^*=\min\{t,\tau\}$ and the second summation is taken as zero if $t\leqslant\tau$.

Taking limits as $\Delta x\to 0$, we get as the projected variance estimate

$$S^2(t)\left\{\frac{1}{rT}\int_0^{t^*}\frac{\lambda(u)\,\mathrm{d}u}{S(u)U(u)}+\frac{1}{r}I[t>\tau]\int_\tau^t\frac{\lambda(u)\,\mathrm{d}u}{S(u)U(u)(T+\tau-u)}\right\} \tag{6}$$

In the special case of constant hazard and loss rates, that is, $\theta(t)\equiv\theta$ and $\lambda(t)\equiv\lambda$, this becomes

$$\exp(-2\lambda t)\left\{\frac{\lambda}{rT}\int_0^{t^*}\exp[(\lambda+\theta)u]\,\mathrm{d}u+\frac{\lambda}{r}I[t>\tau]\int_\tau^t\frac{\exp[(\lambda+\theta)u]\,\mathrm{d}u}{T+\tau-u}\right\} \tag{7}$$

Table I. Values of $E(x)$ for $x = 0.01$ to 2.99 by 0.01.

x	0	1	2	3	4	5	6	7	8	9
0.0	.	4.03793	3.35471	2.95912	2.68126	2.46790	2.29531	2.15084	2.02694	1.91874
0.1	1.82292	1.73711	1.65954	1.58890	1.52415	1.46446	1.40919	1.35778	1.30980	1.26486
0.2	1.22265	1.18290	1.14538	1.10988	1.07624	1.04428	1.01389	0.98493	0.95731	0.93092
0.3	0.90568	0.88151	0.85834	0.83610	0.81475	0.79422	0.77446	0.75544	0.73711	0.71944
0.4	0.70238	0.68591	0.67000	0.65461	0.63973	0.62533	0.61139	0.59788	0.58478	0.57209
0.5	0.55977	0.54782	0.53622	0.52495	0.51400	0.50336	0.49302	0.48296	0.47317	0.46365
0.6	0.45438	0.44535	0.43656	0.42800	0.41965	0.41152	0.40359	0.39585	0.38831	0.38095
0.7	0.37377	0.36676	0.35992	0.35324	0.34671	0.34034	0.33412	0.32803	0.32209	0.31628
0.8	0.31060	0.30504	0.29961	0.29430	0.28910	0.28402	0.27905	0.27418	0.26941	0.26475
0.9	0.26018	0.25571	0.25134	0.24705	0.24285	0.23874	0.23471	0.23076	0.22689	0.22310
1.0	0.21938	0.21574	0.21217	0.20867	0.20524	0.20187	0.19857	0.19534	0.19216	0.18905
1.1	0.18599	0.18299	0.18005	0.17717	0.17433	0.17156	0.16883	0.16615	0.16352	0.16094
1.2	0.15841	0.15592	0.15348	0.15108	0.14873	0.14641	0.14414	0.14191	0.13972	0.13757
1.3	0.13545	0.13337	0.13133	0.12933	0.12735	0.12542	0.12351	0.12164	0.11980	0.11800
1.4	0.11622	0.11447	0.11276	0.11107	0.10941	0.10778	0.10617	0.10460	0.10304	0.10152
1.5	0.10002	0.09854	0.09709	0.09567	0.09426	0.09288	0.09152	0.09019	0.08887	0.08758
1.6	0.08631	0.08506	0.08383	0.08261	0.08142	0.08025	0.07909	0.07796	0.07684	0.07574
1.7	0.07465	0.07359	0.07254	0.07151	0.07049	0.06949	0.06850	0.06753	0.06658	0.06564
1.8	0.06471	0.06380	0.06290	0.06202	0.06115	0.06029	0.05945	0.05862	0.05780	0.05700
1.9	0.05620	0.05542	0.05465	0.05390	0.05315	0.05241	0.05169	0.05098	0.05027	0.04958
2.0	0.04890	0.04823	0.04757	0.04692	0.04627	0.04564	0.04502	0.04440	0.04380	0.04320
2.1	0.04261	0.04204	0.04147	0.04090	0.04035	0.03980	0.03927	0.03874	0.03821	0.03770
2.2	0.03719	0.03669	0.03620	0.03571	0.03523	0.03476	0.03430	0.03384	0.03339	0.03294
2.3	0.03250	0.03207	0.03164	0.03122	0.03081	0.03040	0.03000	0.02960	0.02921	0.02882
2.4	0.02844	0.02806	0.02769	0.02733	0.02697	0.02662	0.02627	0.02592	0.02558	0.02525
2.5	0.02491	0.02459	0.02427	0.02395	0.02364	0.02333	0.02303	0.02273	0.02243	0.02214
2.6	0.02185	0.02157	0.02129	0.02101	0.02074	0.02047	0.02021	0.01994	0.01969	0.01943
2.7	0.01918	0.01893	0.01869	0.01845	0.01821	0.01798	0.01775	0.01752	0.01730	0.01707
2.8	0.01686	0.01664	0.01643	0.01622	0.01601	0.01581	0.01560	0.01540	0.01521	0.01502
2.9	0.01482	0.01464	0.01445	0.01427	0.01409	0.01391	0.01373	0.01356	0.01338	0.01322

This can be written as

$$\exp(-2\lambda t)\left\{\frac{\lambda}{(\lambda+\theta)rT}\{\exp[(\lambda+\theta)t^*]-1\}-I[t\geqslant\tau]\frac{\lambda}{r}\exp[(\lambda+\theta)(T+\tau)]\right.$$
$$\left.\times\{E[(T+\tau-t)(\lambda+\theta)]-E[T(\lambda+\theta)]\}\right\} \tag{8}$$

where $E(x)=\int_x^\infty u^{-1}\exp(-u)\,\mathrm{d}u$.

To facilitate the evaluation of (8), Table I provides values of $E(x)$ that should be sufficient for most purposes. More extensive tables are available [4]. As an alternative, $E(x)$ can be estimated by expressing its integrand as a Taylor series.

While, in theory at least, one could solve for T, τ or r to determine the accrual time, follow-up time, or accrual rate needed to achieve a desired standard error, in practice it might be preferable to simply calculate the results of (8) over a range of values.

Table II. Projected Greenwood standard errors.

	λ						
t	0.23	0.25	0.27	0.29	0.31	0.33	0.35
1	0.0288	0.0296	0.0303	0.0309	0.0315	0.0320	0.0325
2	0.0347	0.0351	0.0354	0.0357	0.0358	0.0359	0.0360
3	0.0363	0.0362	0.0361	0.0358	0.0355	0.0351	0.0346
4	0.0358	0.0353	0.0347	0.0340	0.0333	0.0325	0.0316
5	0.0344	0.0335	0.0324	0.0314	0.0303	0.0292	0.0281
6	0.0332	0.0319	0.0306	0.0292	0.0279	0.0266	0.0253
7	0.0332	0.0316	0.0300	0.0284	0.0269	0.0254	0.0239
8	0.0357	0.0337	0.0317	0.0298	0.0280	0.0262	0.0246

As an example, consider a clinical trial designed to estimate survival probabilities, over time, for a cohort of patients subjected to an experimental treatment. Exponential mortality and loss rates are assumed. The mortality hazard is expected to be between 0.23 and 0.35 per year, corresponding to median survival between two and three years. A loss rate of about 0.03 per year is anticipated. Patients are to be accrued at a rate of 50 per year for four years with five years of additional follow-up. The projected Greenwood standard error estimates are summarized in Table II.

Now suppose the study planners wish to have an estimated standard error estimate for the five-year survival estimate not exceeding 0.025 if the median survival time is two years, that is, $\lambda = 0.35$. If the accrual rate and loss to follow-up rate are not subject to change, then the accrual time must be increased. Additional calculations show that the desired estimated standard error can be achieved with accrual times and follow-up times of 5.1 and 5.0 years, respectively. If a shorter study, at the cost of a longer accrual period, is desired, 6 years of accrual with 2.7 years of follow-up time will also achieve an estimated standard of 0.025 for the five-year survival estimate.

In some cases, the assumption of exponential survival may seem inappropriate. This would be true, for example, if a substantial proportion is expected to be cured; that is, the survival curve exhibits a positive horizontal asymptote. In this case, $\lambda(t)$ and $S(t)$ must be chosen to reflect this property and several such models have been discussed [5–10]. One of these, the Gompertz model, assumes that $\lambda(t) = \alpha \exp(\beta t)$ with $\alpha > 0$ and $\beta < 0$. It follows that $S(t) = \exp\{-\frac{\alpha}{\beta}[\exp(\beta t) - 1]\}$. Often an investigator will have some idea of the value of $S(t)$ for some specific value of t and of the proportion 'cured', that is, $\lim_{t\to\infty} S(t) = S(\infty)$. This allows calculation of values for α and β. Formula (6) can then be used with a module that produces estimates of the definite integrals involved.

As an example, as before, suppose we are planning a study with an accrual rate of 50/year for 4 years followed by 5 years of follow-up. We expect a 3 per cent rate of losses to follow-up. Now, however, we believe that the three-year survival rate will be about 50 per cent and that 30 per cent will be long-term survivors, or cures. If we assume a Gompertz model as described above, we have $\alpha = 0.344$ and $\beta = -0.286$. Applying the QUAD module that is part of SAS IML (version 8), we find the projected estimated standard error of the Kaplan–Meier estimators to be 0.066, 0.069, 0.070 and 0.072 at 4, 5, 6 and 7 years.

Table III. Accuracy of projected standard errors.

t (years)	$\widetilde{SE}_1$	Mean $\widehat{SE}_1$	$\widetilde{SE}_2$	Mean $\widehat{SE}_2$	SE
1	0.0327	0.0325	0.0325	0.0324	0.0334
2	0.0364	0.0362	0.0360	0.0359	0.0366
3	0.0353	0.0349	0.0346	0.0346	0.0346
4	0.0324	0.0319	0.0316	0.0315	0.0309
5	0.0289	0.0283	0.0281	0.0283	0.0278
6	0.0293	0.0285	0.0253	0.0252	0.0241
7	0.0312	0.0295	0.0239	0.0237	0.0237
8	0.0381	0.0337	0.0246	0.0237	0.0254

$\widetilde{SE}_1$ = projected Peto SE. $\widetilde{SE}_2$ = projected Greenwood SE.
Mean $\widehat{SE}_1$ = mean of 1000 Peto SEs. Mean $\widehat{SE}_2$ = mean of 1000 Greenwood SEs.
SE = sample SD of 1000 observations of $\hat{S}(t)$.

SIMULATIONS

In order to further explore the behaviour of the projected Peto and Greenwood standard error estimates, 1000 samples of survival data similar to the scenario described above, assuming exponential survival, were simulated. For each sample a constant hazard rate of 0.35 (corresponding to median survival of about two years) and a 3 per cent annual loss to follow-up rate was assumed. Patients were assumed to enter the study uniformly over a four-year interval at the rate of 50 per year and to be followed for five years after the conclusion of the accrual period. The projected Peto and Greenwood standard error estimates were calculated using formulae (3) and (8) and compared to the mean values of these standard error estimates as calculated from the data. Although it is incidental to the current discussion, the empiric standard errors of these sample survival estimates were also calculated in order to evaluate and compare the performance of the Peto and Greenwood standard error estimates. The results are seen in Table III.

While Table III does not comprise an exhaustive study, it does indicate that the formulae (3) and (8) provide a reasonably accurate projection, based on study planning parameters, of the standard errors of Kaplan–Meier estimates of survival probabilities as estimated by the formulae of Peto and Greenwood. Not surprisingly, the results show little difference from each other and from the sample standard errors until we reach the 'right hand tail' of the survival curve. At seven and eight years, the expected numbers at risk are 6.99 and 2.4, respectively, so we would probably not trust survival estimates or their standard errors. It is interesting to note, however, that the tendency of the Greenwood estimate to underestimate the standard error is not seen here prior to $t = 8$ years, while the conservatism of the Peto estimate appears to be more severe.

A computer program to calculate projected Peto and Greenwood standard errors of Kaplan–Meier estimates is available from the author. Requests can be made via e-mail at abcantor@moffitt.usf.edu.

ACKNOWLEDGEMENTS

The author would like to express his appreciation for helpful discussions with Emmanuel Lazaridis, and for suggestions by Jeffrey Krischer and two anonymous referees.

REFERENCES

1. Kaplan EL, Meier PL. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
2. Greenwood M. The natural duration of cancer. *Reports on Public Health and Medical Subjects* 1926; **33**:1–26.
3. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II analysis and examples. *British Journal of Cancer* 1972; **35**:1–39.
4. Abramowitz M, Stegun I (eds). *Handbook of Mathematical Functions*. Dover Publications: New York, 1964.
5. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society* 1949; **11**:15–44.
6. Gamel JW, Vogel RL. A model of long-term survival following adjuvant therapy for stage 2 breast carcinoma. *British Journal of Cancer* 1993; **68**:1167–1170.
7. Ghitany ME, Maller RA, Zhou S. Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis* 1994; **49**:218–241.
8. Tsodikov A, Loeffler M, Yakovlev A. A cure model with time-changing risk factor: An application to the analysis of secondary leukaemia. *Statistics in Medicine* 1998; **17**:27–40.
9. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**:1041–1046.
10. Cantor AB, Shuster JS. Parametric versus nonparametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine* 1992; **11**:931–937.

