Use of Irwin's Restricted Mean as an Index for Comparing Survival in Different Treatment Groups—Interpretation and Power Considerations

Theodore G. Karrison, PhD
Department of Medicine, University of Chicago, Chicago, Illinois

ABSTRACT: In the analysis of survival data from clinical trials and other studies, the censoring generally precludes estimation of the mean survival time. To accommodate censoring, Irwin (1949) proposed, as an alternative, estimation of the mean lifetime restricted to a suitably chosen time T. In this article we consider the use of Irwin's restricted mean as an index for comparing survival in different groups, using as an example published data from a randomized clinical trial in patients with primary biliary cirrhosis. Irwin's method, originally based on the actuarial survival estimator, is extended to incorporate covariates into the analysis through the use of piecewise exponential models.

For comparing two survival curves, the logrank test is known to be optimal under proportional hazards alternatives. However, comparison of restricted means may outperform the logrank test in situations involving nonproportional hazard rates. We examine the size and power of these two procedures under various proportional and nonproportional hazards alternatives, with and without covariate adjustment. For survival curves that separate early in time the censored data generalization of the Wilcoxon test is known to exhibit high power, and we examine how the comparison of restricted means performs relative to this procedure also. *Controlled Clin Trials 1997;18:151–167* © Elsevier Science Inc. 1997

KEY WORDS: *Restricted mean, censoring, covariates, power*

INTRODUCTION

In the analysis of survival data from clinical trials and other studies the censoring generally precludes estimation of the mean survival time. In order to accommodate censoring, Irwin [1] proposed, as an alternative, estimation of the expectation of life limited (restricted) to a suitably chosen time T, based on the actuarial estimator of the survival curve. In this article we consider the use of Irwin's restricted mean as an index for comparing survival in different treatment groups, using as an example data from a study conducted at the Mayo Clinic comparing D-penicillamine to placebo in patients with primary biliary cirrhosis [2]. We discuss various

Address reprint requests to: Theodore G. Karrison, PhD, Department of Medicine MC6098, University of Chicago, 5841 S. Maryland Avenue, Chicago, IL 60637.

Received November 27, 1995; accepted April 1, 1996.

Controlled Clinical Trials 18:151–167 (1997)
© Elsevier Science Inc. 1997
655 Avenue of the Americas, New York, NY 10010

0197-2456/97/$17.00
PII S0197-2456(96)00089-X

ways of interpreting the restricted mean, including standardization to a "normal" population in a manner akin to relative survival estimation [3]. We also extend the method to incorporate covariates into the analysis through the use of a piecewise exponential model [4–6]. The piecewise exponential model is similar to the Cox [7] proportional hazards regression model in that the covariates are assumed to influence the hazard (death) rate through a multiplicative effect, the same for both treatments. However, unlike the Cox model where the proportional hazards assumption is made with respect to treatment effects also, in the form of the piecewise exponential model adopted here the underlying hazard rates in the two treatment groups are left completely unconstrained and the restricted mean (adjusted for covariates) is used as the basis for group comparisons.

For comparing two survival curves, the logrank test [8,9]—equivalent to the score test from the Cox regression model—is known to be optimal under proportional hazards alternatives and so one can expect some loss in efficiency for a test based on comparison of restricted means in such cases. However, comparison of restricted means may outperform the logrank test in situations involving nonproportional hazard rates. We examine the size and power of these two procedures under various proportional and nonproportional hazards alternatives, for both the unadjusted case and adjusted for a single binary covariate. For early differences in survival the censored data generalization of the Wilcoxon procedure is more powerful than the logrank test [10], and we examine how comparison of restricted means performs relative to this procedure also.

The next section presents Irwin's restricted mean and the extension to covariates via the piecewise exponential model, following closely the exposition given in Karrison [6]. We discuss some related work by Pepe and Fleming [11] and present guidelines for choosing the point of restriction *T*. We then illustrate the methodology using the data from the Mayo Clinic study. Following the example, we derive some power formulas and present the results of a companion simulation study to compare the size and power of the restricted means analysis with the logrank and generalized Wilcoxon tests. The final section of the article summarizes the main ideas and results. An appendix provides details regarding the power calculations.

RESTRICTED MEAN LIFE

Homogeneous Case

Let t_1^0, t_2^0, ..., t_n^0 denote the true survival times from a sample of size n. For the moment we ignore any covariate information and regard the t_i^0s as independent and identically distributed random variables with survival function $S(t) = P(t_i^0 > t)$. In the case of right-censored data the t_i^0s are not always observed; rather, the observed survival time for the i^{th} individual is $t_i = min(t_i^0, c_i)$, where c_i is the i^{th} individual's censoring time, arising either because the subject was still alive at the time of analysis or was lost to followup. Along with t_i we observe the indicator variable δ_i taking the value 1 if t_i corresponds to a death and 0 if t_i corresponds to a censored observation. Thus the data consist of the n pairs (t_1,δ_1), (t_2,δ_2), ..., (t_n,δ_n).

With censored data the sample mean, $\bar{t} = 1/n\sum_{i=1}^{n} t_i$, obviously underestimates the true mean survival time; however, since the survival time is a nonnegative random variable, it is easy to show that the true mean μ can be written

$$\mu = \int_0^\infty S(t)dt. \tag{1}$$

One might therefore consider using Eq. (1) to estimate μ when some of the t_is are censored by substituting the actuarial [12] or Kaplan-Meier [13] estimate of the survival function into the equation. The problem, however, is that when censoring is heavy, $S(t)$ is often ill-determined, even undefined, beyond a certain range. In most clinical trials the follow-up period is much too short to provide reliable information about the tail of the survival curve.

In consideration of this problem, Irwin [1] proposed estimation of the mean lifetime restricted to a suitably chosen time, T, i.e.,

$$\mu(T) = \int_0^T S(t)dt. \tag{2}$$

Thus, Irwin's restricted mean is simply the area under the survival curve up to the point of restriction. More precisely, $\mu(T)$ is the mean of a new random variable taking the value t^0 if $t^0 \leq T$ and T if $t^0 > T$. It is also equivalent to the "total time on test" statistic used in connection with certain life-testing problems [14].

Irwin used the actuarial estimator for $S(t)$ and estimated $\mu(T)$ by applying the trapezoidal rule for numerical integration to Eq. (2), as follows. Divide the interval $(0,T]$ into J, say, subintervals, $(\tau_0,\tau_1]$, $(\tau_1,\tau_2]$, ..., $(\tau_{J-1},\tau_J]$ where $\tau_0 = 0$ and $\tau_J = T$. (Here $(\tau_{j-1},\tau_j]$ denotes an interval open on the left and closed on the right.) Let $\Delta_j = \tau_j - \tau_{j-1}$ and let $\hat{S}(\tau_j)$ be the actuarial estimate of S at τ_j. Applying the trapezoidal rule to Eq. (2) gives

$$\hat{\mu}(T) = \sum_{j=0}^{J} a_j\hat{S}(\tau_j), \tag{3}$$

where $a_j = (\Delta_i + \Delta_{j+1})/2$ $(\Delta_0 = \Delta_{J+1} \equiv 0)$. The variance of $\hat{\mu}(T)$ is then

$$V(\hat{\mu}(T)) = \sum_{j=1}^{J} a_j^2 V(\hat{S}(\tau_j)) + 2\sum_{j<k} a_j a_k Cov(\hat{S}(\tau_j), \hat{S}(\tau_k)), \tag{4}$$

where the variance and covariance terms in Eq. (4) can be approximated using Greenwood's [15] formula. For comparing restricted mean life in, say, two different treatment groups, Eq. (3) can be applied separately to the two groups, giving $\hat{\mu}_g(T)$, $g = 1,2$. The difference, $\widehat{\Delta\mu}(T) = \hat{\mu}_1(T) - \hat{\mu}_2(T)$, has standard error $SE(\widehat{\Delta\mu}(T)) = \sqrt{V(\hat{\mu}_1(T)) + V(\hat{\mu}_2(T))}$ and can be assessed in the usual way by referring $z = \widehat{\Delta\mu}(T)/SE(\widehat{\Delta\mu}(T))$ to tables of the standard normal distribution.

In some closely related work Pepe and Fleming [11] consider a class of statistics based on an integrated weighted difference between the Kaplan-Meier survival curves, i.e.,

$$WKM = \int_0^{T_c} \hat{w}(t)(\hat{S}_1(t) - \hat{S}_2(t))dt,$$

where $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$ are the Kaplan-Meier estimators, $T_c = sup\{t{:}\hat{C}_1(t)\wedge\hat{C}_2(t) > 0\}$, $\wedge$ denotes minimum, and $\hat{C}_g(\cdot)$ is the Kaplan-Meier estimator of the censoring survival function in group g. Note that if the weight function $\hat{w}(t)$ is identically one, the WKM statistic is a difference in mean lifetimes restricted to T_c; however, since $\hat{S}_g(t)$ can be unstable for t near the end of the follow-up period, Pepe and Fleming recommend choosing a weight function that downweights the contributions

Table 1 Required Number of Subjects Remaining at Risk To Achieve Standard Errors of ≤10%, ≤7.5%, and ≤5%

Estimated Survival	Standard Error of Ŝ(t)		
Ŝ(t)	≤10%	≤7.5%	≤5%
50%	13	23	50
40%	10	18	39
30%	7	12	26
20%	4	6	13
10%	1	2	4

of $\hat{S}_1(t) - \hat{S}_2(t)$ over later time periods if censoring is heavy. (Pepe and Fleming provide constraints on the weight function to ensure stability of the WKM statistic.) In essence, Pepe and Fleming suggest going out along the survival curves as far as possible, but downweighting toward the end of the observation period. This approach is attractive, for it ensures stability of the statistic while avoiding the arbitrariness in choosing the point of restriction, although one must still choose $\hat{w}(t)$. However, it produces essentially only a test statistic. Under stochastic ordering, in which $\hat{S}_1(t) \geq \hat{S}_2(t)$ for all t, the WKM statistic would provide an estimated lower bound for the corresponding difference in restricted means, but in general $\int_0^{T_c} \hat{w}(t)\hat{S}_g(t)dt$ is not an interpretable index.

For evaluation and comparison of restricted means, one must generally choose a point of restriction less than T_c. As a guideline, one could consider choosing the largest time point t such that the standard error (SE) of the survival estimate at time t in each treatment group is within reasonable limits, say between 5% and 10%, depending on the relative magnitude of $\hat{S}(t)$. For this purpose, we suggest using the formula proposed by Peto et al. [16]

$$SE(\hat{S}(t)) \doteq \hat{S}(t)\sqrt{(1 - \hat{S}(t))/n_t},$$

where n_t is the number of patients remaining at risk at time t. This formula, though conservative, "deals appropriately with the increasing uncertainty that should properly be expected as one goes along the long flat region with which many life tables finish" [16]. Table 1 shows the required number of subjects remaining at risk in order for the SE to be ≤10%, ≤7.5%, and ≤5% for various values of $\hat{S}(t)$. It is seen that, if toward the end of the observation period $\hat{S}$ is still fairly high, i.e., approximately 40% or 50%, an n_t of 15 would ensure that the SE is between 7.5% and 10%. For $\hat{S}$ in the neighborhood of 20% or 30%, choosing t such that n_t is at least 10 would provide reasonable precision, and for $\hat{S}$ of 10% or below, an n_t of 3 or 4 in each group would yield an SE close to 5%.

Extension to Covariates

Suppose now that for each individual we also observe a $p \times 1$ vector of covariates $z_i = (z_{i1}, z_{i2}, ..., z_{ip})'$. Let (t_i, δ_i, z_i) denote the survival time, indicator variable, and vector of covariates for the i^{th} individual. A model that is particularly well-suited for this problem is the piecewise exponential (PE) model discussed in Holford [4] and Friedman [5]. In the PE model, the time axis is again divided into subintervals

$(0,\tau_1]$, $(\tau_1,\tau_2]$, ..., $(\tau_{J-1},\tau_J]$. The model employed here assumes an underlying, piecewise-constant hazard rate for each treatment group with proportional hazards covariate effects

$$\lambda_g(t;\mathbf{z}) = \lambda_{jg}e^{\boldsymbol{\beta}'\mathbf{z}}, \quad t \in (\tau_{j-1},\tau_j], \quad j = 1,2,\ldots,J, \quad g = 1,2 \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the $p\times1$ vector of regression coefficients. An attractive feature of this model is that the λ_{jg} parameters provide considerable flexibility in accommodating to the shape of the underlying survival curves, much more so than standard parametric approaches. Although the proportional hazards assumption is made with respect to covariate effects, this assumption is avoided with respect to treatment effects by virtue of the two sets of λs, one for each treatment group. If we now let R_{jg} denote the set of individuals entering interval j from treatment group g (the risk set), D_{jg} the set of individuals dying in interval j from treatment group g, and t_{ij} the portion of the j^{th} interval in which individual i is observed as surviving, i.e.,

$$t_{ij} = \begin{cases} t_i - \tau_{j-1}, & \text{if } i \text{ dies or is censored in interval } j \\ \Delta_j = \tau_j - \tau_{j-1}, & \text{if } i \text{ survives through interval } j, \end{cases}$$

then the likelihood function can be written

$$L(\boldsymbol{\lambda},\boldsymbol{\beta}) = \prod_{g=1}^{2}\prod_{j=1}^{J}\left\{\prod_{i\in R_{jg}} exp(-\lambda_{jg}e^{\boldsymbol{\beta}'\mathbf{z}_i}t_{ij}) \prod_{i\in D_{jg}} \lambda_{jg}e^{\boldsymbol{\beta}'\mathbf{z}_i}\right\}, \tag{6}$$

where $\boldsymbol{\lambda}$ is the 2J-dimensional vector $(\lambda_{11}, \lambda_{21}, \ldots, \lambda_{J1}, \lambda_{12}, \lambda_{22},\ldots, \lambda_{J2})'$. Maximum likelihood estimates $(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}})$ of the parameters can be obtained using a variant of the Newton-Raphson algorithm that only requires inversion of a $p \times p$ matrix, and the inverse of the observed Fisher information matrix provides asymptotic variances and covariances (see Karrison [6]).

The estimated survival curve for group g at a fixed value, $\mathbf{z}$, of the covariate vector is then

$$\hat{S}_g(\tau_k;\mathbf{z}) = exp\left(-e^{\hat{\boldsymbol{\beta}}'\mathbf{z}}\sum_{j=1}^{k}\hat{\lambda}_{jg}\Delta_j\right) \tag{7}$$

and the corresponding estimate of restricted mean life, using the trapezoidal rule, is

$$\hat{\mu}_g(T;\mathbf{z}) = \sum_{j=0}^{J}a_j\hat{S}_g(\tau_j;\mathbf{z}). \tag{8}$$

Finally, an adjusted comparison of restricted means between the two treatment groups is given by

$$\widehat{\Delta\mu}(T;\mathbf{z}) = \hat{\mu}_1(T;\mathbf{z}) - \hat{\mu}_2(T;\mathbf{z}). \tag{9}$$

An approximation to the variance of $\hat{S}_g(\tau_k;\mathbf{z})$, $\hat{\mu}_g(T;\mathbf{z})$, and $\widehat{\Delta\mu}(T;\mathbf{z})$ can be obtained by the delta method [17].

Eq. (9) provides a comparison of restricted mean life in the two groups adjusted to a fixed value $\mathbf{z}$ of the covariate vector. If an overall adjusted comparison is desired, one can, for example, average Eq. (9) over the observed marginal covariate distribution arising form both treatment groups, i.e.,

$$\widehat{\Delta\mu_D}(T) = \frac{1}{n_1 + n_2} \sum_{i=1}^{n1+n2} \widehat{\Delta\mu}(T;\mathbf{z}_i), \tag{10}$$

where n_1 and n_2 are the sample sizes in groups 1 and 2, respectively. The "D" subscript is used here because Eq. (10) is analogous to estimates obtained by the method of "direct" adjustment, but with each "stratum" formed by an individual covariate vector. (Lane and Nedler [18] consider estimates of this form in connection with multiple logistic regression.) Again, the delta method can be used to obtain the asymptotic variance of $\widehat{\Delta\mu_D}(T)$ and a test of significance performed by referring $z = \widehat{\Delta\mu_D}(T)/SE(\widehat{\Delta\mu_D}(T))$ to tables of the standard normal distribution. Note that in a randomized trial, the adjusted difference [Eq. (10)] and the unadjusted difference estimate the same quantity, namely, the true difference in restricted means in the whole population, but $\widehat{\Delta\mu_D}(T)$ is more precise by virtue of the covariate adjustment.

EXAMPLE: PRIMARY BILIARY CIRRHOSIS TRIAL

A randomized, double-blind, placebo controlled clinical trial of the immunosuppressive agent D-penicillamine for the treatment of primary biliary cirrhosis (PBC) was performed at the Mayo Clinic between January 1974 and July 1986. Dickson et al. [2] give an early report of the results of this study in patients with histologically advanced disease. The full dataset appears in Appendix D.1 of the text by Fleming and Harrington [19] who fit various Cox proportional hazards models to the data to compare the survival rates in the two groups and to assess and adjust for the effects of covariates. The randomized trial enrolled a total of 312 patients; 158 were assigned to D-penicillamine and 154 to placebo. A total of 125 of the 312 patients died over the course of followup, which ranged from 2 to a little over 12 years, with all but 11 deaths attributable to PBC. Figure 1, the Kaplan-Meier survival curves for the two groups, shows little difference in the overall survival rates (logrank $p = 0.750$). Correspondingly, the estimated relative death rate (placebo/D-penicillamine) is 0.94, with a 95% confidence interval ranging from 0.66 to 1.34.

For evaluation and comparison of restricted means, we chose $T = 10$ years as the point of restriction. At this time point the survival rates are approximately 40% and 15 patients remained at risk in each of the two treatment groups. Thus, from Table 1 the standard error of $\hat{S}(10)$ is between 7.5% and 10%. The estimated mean life restricted to 10 years is 7.13 ± 0.28 years (estimate ± SE) in the D-penicillamine group and 7.29 ± 0.30 years in the placebo group. The difference, -0.15 ± 0.41 years, is small and not statistically significant. [These results were obtained by fitting a PE model to the data as given by Eq. (5), but without including any covariates. They may therefore differ slightly from those based on the traditional actuarial approach (see, e.g., Kaplan and Meier [13], pp. 471–475). Interval widths were 0.5 years of the first 5 years and 1 year thereafter.] How might we interpret these indices? If the observed survival curves at the point of restriction T were near zero, then these estimates would be close to the overall (unrestricted) means, but this is not the case. However, dividing each restricted mean by T yields the percentage of total possible life-years achieved over the period in question, since if there were no mortality the restricted mean would be equal to T. Thus, we can say that the D-penicillamine group achieved 7.13/10 = 71.3% of total possible life-years over this period compared to 7.29/10 = 72.9% in the placebo group.

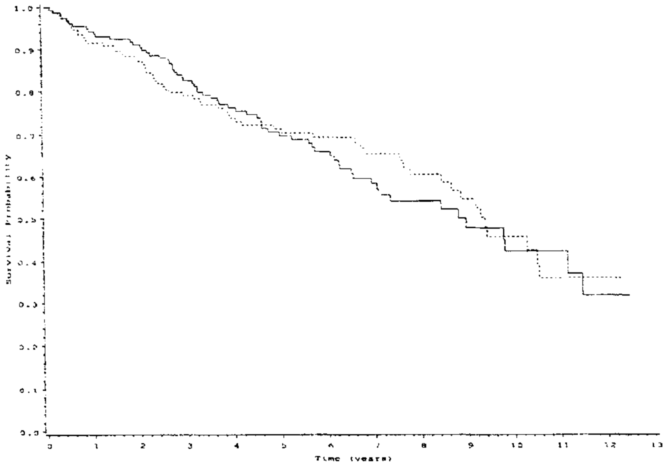

Figure 1 Survival curves for patients in the Mayo Clinic primary biliary cirrhosis trial. Solid line: D-penicillamine group; dotted line: placebo group.

Alternatively, to account for naturally occurring mortality, the restricted mean can be divided by the corresponding figure for an age, race, and sex-matched "normal" population. For this calculation the survival curve for the normal population is first obtained from published U.S. life-tables as described in Ederer et al. [3]. The area under this curve from 0 to T then provides the corresponding denominator term. For the PBC data this procedure yields values of 9.25 and 9.38 years in the D-penicillamine and placebo groups, respectively. (The published dataset does not include the race of the subjects. However, Table 1 of Dickson et al. [20] indicates that only 2% of the patients were nonwhite; therefore, for purposes of the above calculation, we used the U.S. life-tables for white males and white females.) This calculation indicates that over this period the D-penicillamine and placebo groups achieved 7.13/9.25 = 77.1% and 7.29/9.38 = 77.7%, respectively, of "expected" life-years. One could also simply report that in both treatment groups the PBC patients lost approximately 2.1 years over the 10-year follow-up period compared to the nondiseased population.

Because a model for the natural history of PBC was also of interest to the Mayo investigators, they fit various multivariate Cox proportional hazards models using information available from a set of 45 predictor variables [19,20]. The primary objective of their analysis was to develop a simple model based on inexpensive, noninvasive measurements that could be used for patient management and for decisions about the timing of liver transplantation. The final model chosen included age, log(albumin), log(bilirubin), log(prothrombin time), and the presence or absence of edema; Table 2 reproduces the results. The treatment effect adjusted for

Table 2 Estimated Covariate Effects and Adjusted Treatment Effect for PBC Trial (Cox Regression and Piecewise Exponential Models)

	Cox Model			PE Model	
	$\hat\beta$ ± SE	Z-statistic		$\hat\beta$ ± SE	Z-statistic
Age	0.035 ± 0.009	3.89		0.033 ± 0.009	3.63
log(Albumin)	−3.08 ± 0.72	−4.28		−3.23 ± 0.71	−4.54
log(Bilirubin)	0.88 ± 0.10	8.96		0.83 ± 0.10	8.44
Edema	0.79 ± 0.30	2.65		0.81 ± 0.30	2.72
log(Prothrombin time)	2.97 ± 1.02	2.92		3.22 ± 1.02	3.15
Treatment effect* $\hat\beta$ =	0.14 ± 0.18	0.73	$\hat\Delta\mu_D(10)$ =	0.06 ± 0.27	0.22

SE: standard error.

* As assessed by the regression coefficient for the Cox model and the directly adjusted difference in restricted means for the piecewise exponential model.

these covariates now slightly favors D-penicillamine, with an estimated relative death rate (placebo/D-penicillamine) of $e^{0.14}$ = 1.15 and a 95% confidence interval ranging from 0.80 to 1.65. Table 2 also shows the results obtained from fitting a PE model, as given in Eq. (5), using the same set of covariates. As before, interval widths of 0.5 years for the first 5 years and 1 year for the next 5 years were used. As might be expected the estimated covariate effects (and standard errors) from the PE model are all very close to those obtained from the Cox regression model. Here, however, the treatment effect is expressed in terms of the difference in restricted means. As mentioned above, an advantage of this comparison relative to that based on the Cox model is that, at least as far as treatment effects are concerned, the proportional hazards assumption is avoided. (Note that the Kaplan-Meier survival curves cross at about 4.5 years). The "directly" adjusted treatment difference also slightly favors the D-penicillamine group, as did the adjusted Cox analysis, but again is well within the range expected to occur by chance alone. With 95% confidence the difference in favor of D-penicillamine is at most 0.06 + 1.96(0.27) = 0.59 years, an effect that would not likely outweigh the drug's expense and associated toxicity.

POWER COMPARISONS

In this section we derive analytical power expressions for the comparison of restricted means in the unadjusted case and in the case of a single binary covariate. We then compare the power obtained from a comparison of restricted means with that of the logrank and generalized Wilcoxon tests under various proportional and nonproportional hazards alternatives. (We obtained power estimates for the logrank and generalized Wilcoxon tests using the methods and computer program developed by Lakotos [21].) In addition to the analytically derived power estimates, we present the results of a companion simulation study.

In the unadjusted (homogeneous) case, the estimated difference in restricted means can be written

$$\widehat{\Delta\mu}(T) = \sum_{j=0}^{J} a_j(\hat{S}_1(\tau_j) - (\hat{S}_2(\tau_j)), \tag{11}$$

where $\hat{S}_g(\cdot)$ denotes the estimated survival curve in group g based on a PE model with no covariates. The approximate expectation of Eq. (11) is just

$$E = \sum_{j=0}^{J} a_j(S_1(\tau_j) - (S_2(\tau_j)),$$

where $S_1(\cdot)$ and $S_2(\cdot)$ are the true survival probabilities. Letting V denote the variance of $\widehat{\Delta\mu}(T)$, the power of a two-sided, α-level test can be obtained by solving for $1 - \beta$ in

$$\frac{E}{\sqrt{V}} = z_{\alpha/2} + z_{\beta}, \tag{12}$$

where z_α is the standard normal deviate.

In the unadjusted case, $\hat{\mu}_1(T)$ and $\hat{\mu}_2(T)$ are independent, so V is given by

$$V(\widehat{\Delta\mu}(T)) = V(\hat{\mu}_1(T)) + V(\hat{\mu}_2(T)).$$

The variance of each individual mean is

$$V(\hat{\mu}_g(T)) = \sum_{j=1}^{J} a_j^2 V(\hat{S}_g(\tau_j)) + 2\sum_{j<k} a_j a_k Cov(\hat{S}_g(\tau_j),\hat{S}_g(\tau_k)), \tag{13}$$

where

$$\begin{aligned} V(\hat{S}_g(\tau_k)) &= (\hat{S}_g(\tau_k))^2 \sum_{j=1}^{k} \Delta_j^2 V(\hat{\lambda}_{jg}) \\ &= (\hat{S}_g(\tau_k))^2 \sum_{j=1}^{k} \Delta_j^2 \frac{d_{jg}}{(\sum_{i\in R_{jg}} t_{ij})^2} \end{aligned} \tag{14}$$

and

$$Cov(\hat{S}_g(\tau_k),\hat{S}_g(\tau_l)) = \hat{S}_g(\tau_k)\hat{S}_g(\tau_l)\sum_{j=1}^{k} \Delta_j^2 \frac{d_{jg}}{(\sum_{i\in R_{jg}} t_{ij})^2},\ k < l. \tag{15}$$

(Here, d_{jg} denotes the number of individuals in treatment group g dying in interval j.) To obtain V for the power calculations, we replace $\hat{S}_g(\cdot)$, d_{jg}, and $\sum_{i\in R_{jg}} t_{ij}$ in Eqs. (14) and (15) with their approximate expectations. For $\hat{S}_g(\cdot)$, this is just the true survival rate, $S_g(\cdot)$. For d_{jg} and $\sum_{i\in R_{jg}} t_{ij}$, these quantities are

$$E(d_{jg}) = n_g \int_{\tau_{j-1}}^{\tau_j} [1 - H_g(t)]\lambda_g(t)S_g(t)dt \tag{16}$$

and

$$E\left(\sum_{i\in R_{jg}} t_{ij}\right) = n_g \int_{\tau_{j-1}}^{\tau_j} [1 - H_g(t)]S_g(t)dt, \tag{17}$$

respectively, where n_g is the total number of subjects in group g and $H_g(t)$ is the corresponding censoring distribution. The appendix provides details of the derivation for the particular censoring and survival configurations used in the power study (see below).

In the case of a single binary covariate taking the value $z = z_0$ for a proportion p of the observations and $z = z_1$ for the remaining observations, the directly adjusted difference [Eq. (10)] is

$$\widehat{\Delta\mu_D}(T) = p \sum_{j=1}^{J} a_j(\hat{S}_1(\tau_j;z_0) - \hat{S}_2(\tau_j;z_0)) + (1 - p) \sum_{j=1}^{J} a_j(\hat{S}_1(\tau_j;z_1) - \hat{S}_2(\tau_j;z_1)). \tag{18}$$

Again, to obtain the power we first calculate the approximate expectation of Eq. (18) by replacing the survival estimates with their true values. The variance of Eq. (18), using the delta method, is then

$$V(\widehat{\Delta\mu_D}(T)) = D'I(\hat{\lambda},\hat{\beta})^{-1}D, \tag{19}$$

where D is the $(2J+1)$ vector of derivatives of Eq. (18) with respect to the λs and β; and $I(\hat{\lambda},\hat{\beta})$ is the $(2J+1) \times (2J+1)$ Fisher information matrix. The elements of D can be expressed as functions of the survival estimates and $\hat{\beta}$ (see Karrison [22], (3.39), (3.40), and p 96). Thus, to calculate the approximate expectation of Eq. (19), the true survival rates and regression coefficient β can be substituted into the derivative vector. Calculation of the expected information matrix requires derivation of the expectation of quantities such as d_{jg}, $\Sigma_{i\in R_{jg}} e^{\hat{\beta}z_i} t_{ij}$, and $\Sigma_{i\in R_{jg}} z_i e^{\hat{\beta}z_i} t_{ij}$. For example,

$$E(d_{jg}) = n_{g0}\int_{\tau_{j-1}}^{\tau_j} [1 - H_g(t)]\lambda_g(t;z_0)S_g(t;z_0)dt + n_{g1}\int_{\tau_{j-1}}^{\tau_j} [1 - H_g(t)]\lambda_g(t;z_1)S_g(t;z_1)dt, \tag{20}$$

where n_{g0} and n_{g1} are the number of individuals in treatment group g with $z = z_0$ and $z = z_1$, respectively, and

$$E\left(\sum_{i\in R_{jg}} e^{\hat{\beta}z_i} t_{ij}\right) \doteq e^{\beta z_0} n_{g0}\int_{\tau_{j-1}}^{\tau_j} [1 - H_g(t)]S_g(t;z_0)dt + e^{\beta z_1} n_{g1}\int_{\tau_{j-1}}^{\tau_j} [1 - H_g(t)]S_g(t;z_1)dt. \tag{21}$$

The interested reader is referred to the appendix for further derivations and for details of the calculation under the assumed censoring and survival configurations used in the power study.

For the power study we assumed that the true survival times were drawn from two Weibull distributions with proportional hazards covariate effects,

$$S_1(t;z) = exp[-(\alpha_1 t)^{\gamma_1} e^{\beta' z}]$$

and

$$S_2(t;z) = exp\{-(\alpha_2 t)^{\gamma_2} e^{\beta' z}\}.$$

Four sets of parameter values were chosen to enable comparisons under the null hypothesis (no treatment difference) and under proportional hazards, early difference, and late difference alternatives. We then considered a hypothetical trial of 6 years duration with uniform enrollment over the first 2 years (giving potential censoring times between 4 and 6 years) and a moderate sample size of $n = 100$ patients per group. We included a single binary covariate taking the value $z = 0$ for 50 observations in each group and $z = 1$ for the remaining 50 observations. For the unadjusted (homogeneous) case, β was set to 0 and in the nonhomogeneous case β was set to 0.4, corresponding to about a 50% increase in the death rate for individuals with $z = 1$ compared to those with $z = 0$. Figure 2 shows the different configurations of survival distributions in the homogeneous case. Constant, half-year interval widths were used for estimation of the restricted mean. The power was calculated for two points of restriction, namely, $T = 5$ years and $T = 5.5$ years. Note from Figure 2 that the survival rates at 5 years are approximately 40%, so with uniform censoring over the interval (4, 6), the expected number of individuals remaining at risk in each of the two treatment groups is approximately 100 (0.4 × 0.5) = 20. From Table 1 the standard error of the survival estimates at this time

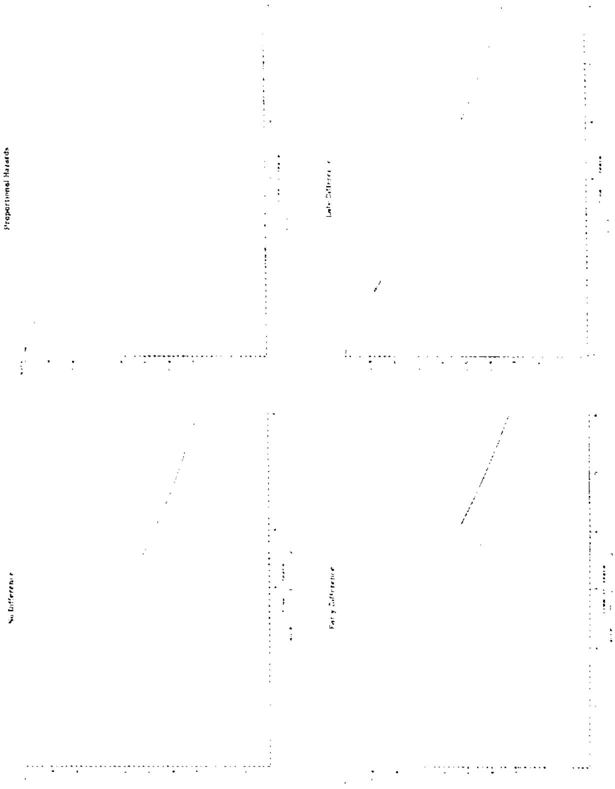

Figure 2 Hypothetical survival configurations used in the power study (homogeneous case; see Table 3 for parameter values).

point would be less than 7.5% using the Peto et al. formula. At T = 5.5 years, the expected number remaining at risk is approximately 100 (0.35 × 0.25) = 8.8; at this time point the standard error would be close to 10%.

For the companion simulation study, the observed survival time for each observation was the minimum of the survival time drawn from the appropriate Weibull distribution and a randomly generated censoring time drawn from the Uniform (4, 6) distribution, with the indicator variable set to 1 or 0 accordingly. For each generated dataset the survival curves were then compared by the three different methods, with testing performed at the two-sided, α = 0.05 level. The analysis based on restricted means was performed using a FORTRAN program; the logrank and generalized Wilcoxon tests were performed using the LIFETEST procedure of SAS [23]. Note that each of these latter tests can be expressed as a linear rank statistic (Kalbfleisch and Prentice [24], pp 144–149). For the restricted means comparison we evaluated the differences adjusted to z = 0, to z = 1, and using the direct adjustment procedure; because the results were similar we show only the latter, which correspond to the analytically derived power estimates. For the logrank and generalized Wilcoxon tests, the adjusted comparisons were based on a stratified version of the linear rank statistic. A total of R = 3000 simulations were performed for each configuration, giving a standard error due to simulation of approximately 0.4% in the null case and 0.9% in the nonnull case. Random numbers were generated using the linear congruential random number generator LCRANS available from the SUN math library.

Table 3 shows the results of the power comparisons. Results for T = 5 and T = 5.5 years are based on independent runs. The logrank and generalized Wilcoxon results were obtained using the same datasets generated for the T = 5 restricted means comparison. The mean percent censored is the average of the means for T = 5 and T = 5.5. Upper entries are the analytically derived powers; lower entries are the power estimates obtained from the simulations. We note first that the theoretical and simulated powers agree quite well and all three procedures gave about the right size test. Under proportional hazards (PH) alternatives, for both the unadjusted and adjusted cases, the logrank test performed best as it should, although the analysis based on the comparison of restricted means was competitive. Power for the restricted means comparison was higher at T = 5.5 than at T = 5 years. The generalized Wilcoxon test was similar in power to the restricted means analysis under PH alternatives. With early differences, the restricted means comparison was much more powerful than the logrank test, and even outperformed the generalized Wilcoxon test. Here, although the area between the true survival curves is greater at 5 5 than at 5 years, the difference did not compensate for the increase in standard error and, consequently, the power of the restricted means comparison decreased with T. A tradeoff in power occurred for late difference alternatives, where the restricted mean exhibited less power than the logrank test. Because of the divergence of the survival curves the power of the restricted means analysis increased with T despite the increased standard error. As a result, at T = 5 years the restricted mean had lower power than the generalized Wilcoxon test, whereas at T = 5.5 years the power was slightly greater.

SUMMARY

This article was written with two main purposes in mind. The first was to show that the restricted mean can be a useful index for comparing survival in different

Table 3 Power Comparisons for Restricted Mean, Logrank, and Generalized Wilcoxon Tests

Survival	Weibull		Mean	Power (in %)			
Configuration	Parameters		% Cens.	RM		LR	GW
Homogeneous Case (β = 0)—Unadjusted				T=5	T = 5.5		
No difference	α_1 = 0.20	γ_1 = 1.25	37.1	5.0	5.0	5.0	5.0
	α_2 = 0.20	γ_2 = 1.25		4.7	5.2	4.9	5.0
Proportional	α_1 = 0.16	γ_1 = 1.25	38.0	73.6	76.7	80.5	76.1
hazards (PH)	α_2 = 0.24	γ_2 = 1.25		72.4	75.2	79.9	77.7
Non-PH	α_1 = 0.18	γ_1 = 1.50	40.0	83.6	77.9	41.3	71.2
Early difference	α_2 = 0.20	γ_2 = 0.75		82.8	79.9	41.3	71.4
Non-PH	α_1 = 0.18	γ_1 = 1.25	29.6	69.8	79.4	95.3	78.5
Late difference	α_2 = 0.28	γ_2 = 1.75		69.8	78.1	95.0	80.0
Nonhomogeneous Case (β = 0.4)—Adjusted							
No difference	α_1 = 0.18	γ_1 = 1.25	34.6	5.0	5.0	5.0	5.0
	α_2 = 0.18	γ_2 = 1.25		4.8	5.1	5.1	4.7
Proportional	α_1 = 0.14	γ_1 = 1.25	35.8	84.1	86.4	89.1	85.1
hazards (PH)	α_2 = 0.22	γ_2 = 1.25		83.1	85.7	88.4	85.3
Non-PH	α_1 = 0.16	γ_1 = 1.50	37.3	94.8	91.8	67.7	89.4
Early difference	α_2 = 0.18	γ_2 = 0.75		95.3	92.5	67.6	91.0
Non-PH	α_1 = 0.16	γ_1 = 1.25	27.8	73.2	82.1	96.0	79.1
Late difference	α_2 = 0.26	γ_2 = 1.75		71.6	81.9	94.9	77.9

Cens: censored; RM: restricted mean; LR: logrank; GW: generalized Wilcoxon; T is point of restriction. Top entry is analytically derived power, lower entry is from simulations. Constant half-year interval widths were used in the calculation of the RM.

groups, and the second was to examine the power of such comparisons relative to some commonly used tests. Although not as readily interpretable as the overall (unrestricted) mean, the restricted mean does provide a valid summary measure that can be expressed in a variety of ways. As illustrated by the analysis of the PBC trial, it can be expressed relative to T to provide an indication of the total possible life-years achieved or, perhaps better still (particularly for long-term studies), relative to the restricted mean of a suitably matched normal population. Under the "actuarial-like" approach taken here the estimate does not depend upon any assumed shape for the true survival curve. The method is closely related to the integrated WKM statistic proposed by Pepe and Fleming [11]. As noted earlier, the WKM method downweights the differences in survival with increasing t to ensure stability of the estimate under censoring. Comparison of restricted means involves truncation at an appropriate time point T prior to the end of the observation period with zero weight beyond T. We discussed some guidelines for choosing T based on Peto et al.'s estimate of the error in the survival curve; further work is needed in this area.

Covariates can be incorporated into the analysis by assuming a proportional hazards model; however, a nice feature of the method is that, with regard to the effect of most interest in a clinical trial, namely, the treatment effect, this assumption is easily avoided. As indicated from the results of the power study, compared with other test procedures this can lead to increased power for alternatives in which the treatment difference occurs early without sacrificing much power when the proportional hazards assumption does, in fact, hold. If survival differences do not occur until relatively late in the course of followup, however, then comparison of restricted means will entail a loss of power relative to the logrank test.

An obvious disadvantage of the PE model is the need to specify interval cutpoints. Zucker [25] has recently developed a method for estimating the restricted mean based on a stratified Cox model that avoids the need to divide the time axis into arbitrary intervals and that uses standard Cox model software (with some additional postprocessing routines). Extending the work of Anderson and Gill [26], Zucker showed that the resulting estimate is asymptotically normal and derived an appropriate variance estimator. He also developed a robust version of the technique that provides a valid group comparison when the underlying stratified Cox model is incorrect.

The author would like to thank Paul Meier for originally suggesting this problem and for his many valuable insights. Rick Chappell provided helpful comments on an earlier draft and two anonymous reviewers made several suggestions that led to considerable improvements in the manuscript.

REFERENCES

1. Irwin JO. The standard error of an estimate of expectational life. *J Hygiene.* 1949; 47:457–481.
2. Dickson ER, Fleming TR, Wiesner RH, Baldus WP, Fleming CR, Ludwig J, McCall JT. Trial of penicillamine in advanced primary biliary cirrhosis. *N Engl J Med.* 1985;312:1011–1015.
3. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: A statistical methodology. *NCI Mono No.* 6 1961;101–121.
4. Holford TR. Life tables with concomitant information. *Biometrics* 1976;32:587–597.
5. Friedman M. Piecewise exponential models for survival data with covariates. *Ann Stat* 1982;10:101–113.
6. Karrison T. Restricted mean life with adjustment for covariates. *J Am Stat Assoc.* 1987;82:1169–1176.
7. Cox DR. Regression models and life tables (with discussion). *J Royal Stat Soc Ser. B* 1972;34:187–220.
8. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966;50:163–170.
9. Peto R. Rank tests of maximal power against Lehmann-type alternatives. *Biometrika.* 1972;59:472–475.
10. Prentice RL. Liner rank tests with right censored data. *Biometrika.* 1978;65:153–158.
11. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics.* 1989;45:497–507.
12. Cutler SJ, Ederer F. Maximum utilization of the life table in analyzing survival. *J Chron Dis.* 1958;8:699–712.
13. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53:457–481.
14. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. *Statistical Inference Under Order Restrictions.* London: John Wiley; 1972.
15. Greenwood M. The natural duration of cancer. In: Reports on Public Health and Medical Subjects, No. 33. London, Her Majesty's Stationery Office. 1926;1–26.

16. 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. *Br J Cancer.* 1977;35:1–39.

17. Rao CR. *Linear Statistical Inference and Its Applications.* 2nd ed New York: John Wiley; 1973.

18. Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics.* 1982;38:613–621.

19. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis.* New York: John Wiley; 1991.

20. Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology.* 1989;10:1–7.

21. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics.* 1988;44:229–241.

22. Karrison T. Restricted mean life with adjustment for covariates. Unpublished PhD dissertation; University of Chicago, Department of Statistics; 1985.

23. SAS. SAS/STAT User's Guide, Version 6, 4th ed., Volume 2. Cary, NC: SAS Institute; 1990.

24. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* New York: John Wiley; 1980.

25. Zucker DM. Restricted mean life with covariates: Modification and extension of a useful survival analysis method. Technical Report, Department of Statistics, Hebrew University; 1996.

26. Anderson PK, Gill RD. Cox's regression model for counting processes: A large sample study. *Ann Stat.* 1982;10:1100–1120.

APPENDIX

Formulas are derived under the following situation: a study of 6 years duration with uniform enrollment over the first 2 years and true Weibull survival curves. For the calculation of restricted means the time axis is divided into 12 intervals of length 0.5 years, $(0,\tau_1]$, $(\tau_1,\tau_2]$, ..., $(\tau_{11},\tau_{12}]$.

Unadjusted (Homogeneous) Case

The Weibull survival and hazard rates are $S_g(t) = exp(-(\alpha_g t)^{\gamma_g})$ *and* $\lambda_g(t) = \alpha_g\gamma_g(\alpha_g t)^{\gamma_g - 1}$, $g = 1,2$, and the censoring distribution for both groups is

$$H(t) = \begin{cases} 0, & t \leq 4 \\ t/2 - 2, & 4 < t \leq 6 \\ 1, & t > 6. \end{cases} \tag{A1}$$

Since no censoring occurs prior to 4 years, the integral in Eq. (16) is just $n_g(S_g(\tau_{j-1}) - S_g(\tau_j))$ for $j \leq 8$. For $j > 8$, (16) is

$$n_g \int_{\tau_{j-1}}^{\tau_j} (3 - t/2)\alpha_g\gamma_g(\alpha_g t)^{\gamma_g - 1} \, exp(-(\alpha_g t)^{\gamma_g})dt. \tag{A2}$$

After separating terms and an integration by parts, Eq. (*A*2) can be written

$$3n_g(S_g(\tau_{j-1}) - S_g(\tau_j)) + \frac{n_g}{2}(\tau_j S_g(\tau_j) - \tau_{j-1}S_g(\tau_{j-1})) - \frac{n_g A}{2},$$

where $A = \int_{\tau_{j-1}}^{\tau_j} e^{-(\alpha_g t)^{\gamma_g}} \, dt$. The integral, A, was evaluated using Simpson's rule for numerical quadrature. The integral in Eq. (17) can be written

$$n_g \int_{\tau_{j-1}}^{\tau_j} (1 - H(t))S_g(t)dt = \begin{cases} n_g A, & \mathrm{j} = 1, 2, \ldots, 8 \\ 3n_g A - \frac{n_g}{2}B, & \mathrm{j} > 8 \end{cases} \tag{A3}$$

where $B = \int_{\tau_{j-1}}^{\tau_j} te^{-(\alpha_g t)^{\gamma_g}}\, dt$. B was obtained using numerical integration.

Adjusted Case (Single Binary Covariate)

Here $S_g(t;z) = exp(-(\alpha_g t)^{\gamma_g} e^{\beta z})$, $\lambda_g(t;z) = \alpha_g \gamma_g (\alpha_g t)^{\gamma_g - 1} e^{\beta z}$, and $z = z_0$ or z_1. The observed information matrix can be partitioned as

$$I(\hat{\lambda},\hat{\beta}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix},$$

where I_{11} is the $2J \times 2J$ matrix with typical element

$$(I_{11})_{jg,jg} = \frac{d_{jg}}{(\hat{\lambda}_{jg})^2}, \tag{A4}$$

I_{12} is the $2J \times 1$ vector with typical element

$$(I_{12})_{jg} = \sum_{i \in R_{jg}} z_i e^{\hat{\beta} z_i} t_{ij}, \tag{A5}$$

$I_{21} = I_{12}'$, and I_{22} is the scalar

$$I_{22} = \sum_j \sum_g \sum_{i \in R_{jg}} z_i^2 \hat{\lambda}_{jg} e^{\hat{\beta} z_i} t_{ij}. \tag{A6}$$

Approximate expectations of Eqs. (A4), (A5), and (A6) were calculated as follows. For Eq. (A4) we substituted the expectations of d_{jg} and $\hat{\lambda}_{jg}$ into the numerator and denominator. The former is given by Eq. (20). For the latter, it can be shown that $\hat{\lambda}_{jg} = \dfrac{d_{jg}}{\sum_{i \in R_{jg}} e^{\hat{\beta} z_i} t_{ij}}$. Thus

$$E(\hat{\lambda}_{jg}) \doteq \frac{E(d_{jg})}{E(\sum_{i \in R_{jg}} e^{\beta z_i} t_{ij})} \tag{A7}$$

where the denominator is given by Eq. (21). For Eq. (A5),

$$E(\mathbf{I}_{12})_{jg} \doteq z_0 e^{\beta z_0} n_{g0} \int_{\tau_{j-1}}^{\tau_j} [1 - H(t)]S_g(t;z_0)dt + z_1 e^{\beta z_1} n_{g1} \int_{\tau_{j-1}}^{\tau_j} [1 - H(t)]S_g(t;z_1)dt\,, \tag{A8}$$

and finally, for Eq. (A6) we have

$$E(I_{22}) \doteq \sum_j \sum_g \left\{ z_0^2 \tilde{\lambda}_{jg} e^{\beta z_0} n_{g0} \int_{\tau_{j-1}}^{\tau_j} [1 - H(t)]S_g(t;z_0)dt \right. \\ \left. + z_1^2 \tilde{\lambda}_{jg} e^{\beta z_1} n_{g1} \int_{\tau_{j-1}}^{\tau_j} [1 - H(t)]S_g(t;z_1)dt \right\} \tag{A9}$$

where $\tilde{\lambda}_{jg}$ denotes the ratio of expectations in Eq. (A7). Note that the initial terms on the right-hand side of Eqs. (A8) and (A9) vanish if z is coded as $z_0 = 0$ and $z_1 = 1$.

To evaluate the above expressions two basic integrals are needed,

$$\int_{\tau_{j-1}}^{\tau_j} [1 - H(t)]\lambda_g(t;z)S_g(t;z)dt \tag{A10}$$

and

$$\int_{\tau_{j-1}}^{\tau_j} [1 - H(t)]S_g(t;z)dt. \tag{A11}$$

Analogous to the unadjusted case, Eq. (A10) is just $S_g(\tau_{j-1};z) - S_g(\tau_j;z)$ for $j \leqslant 8$. For $j > 8$, Eq. (A10) can be written

$$3\{S_g(\tau_{j-1};z) - S_g(\tau_j;z)\} + \frac{1}{2}[\tau_j S_g(\tau_j;z) - \tau_{j-1} S_g(\tau_{j-1};z)] - \frac{C}{2},$$

with $C = \int_{t_{j-1}}^{t_j} e^{-(\alpha_g t)^{\gamma_g} e^{\beta z}}\, dt$ evaluated using numerical integration. Similarly, Eq. (A11) is equal to C for $j \leqslant 8$ and $3C - \frac{D}{2}$ for $j > 8$ with $D = \int_{t_{j-1}}^{t_j} t e^{-(\alpha_g t)^{\gamma_g} e^{\beta z}}\, dt$ evaluated numerically.