Research Article

Statistics in Medicine

Received 22 April 2008, Accepted 7 April 2010 Published online 25 May 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3969

A robust weighted Kaplan–Meier approach for data with dependent censoring using linear combinations of prognostic covariates

Chiu-Hsieh Hsu[a,b*†] **and Jeremy M. G. Taylor**[c]

The weighted Kaplan–Meier (WKM) estimator is often used to incorporate prognostic covariates into survival analysis to improve efficiency and correct for potential bias. In this paper, we generalize the WKM estimator to handle a situation with multiple prognostic covariates and potential-dependent censoring through the use of prognostic covariates. We propose to combine multiple prognostic covariates into two risk scores derived from two working proportional hazards models. One model is for the event times. The other model is for the censoring times. These two risk scores are then categorized to define the risk groups needed for the WKM estimator. A method of defining categories based on principal components is proposed. We show that the WKM estimator is robust to misspecification of either one of the two working models. In simulation studies, we show that the robust WKM approach can reduce bias due to dependent censoring and improve efficiency. We apply the robust WKM approach to a prostate cancer data set.

Keywords: dependent censoring; risk scores; prognostic covariates; weighted Kaplan–Meier estimator

1. Introduction

In survival analysis, there is a loss of information about the event times due to censoring and there is the potential for bias in survival estimates if the censoring times are dependent on the event times. The type of dependent censoring we consider in this paper is that censoring depends on survival only through the prognostic covariates, a natural generalization of missing at random in the language of missing data to censoring [1]. In addition to the event times, prognostic covariates for the event times are also often collected on all individuals. In this paper, we propose to deal with dependent censoring through the use of prognostic covariates. The prognostic covariates can be incorporated into survival estimates to recover some of the lost information. When the main interest is in estimating the marginal survival estimates or comparing two treatment-specific survival distributions several approaches [2–11] have been proposed to recover the lost information using prognostic covariates. Most of these approaches directly incorporate prognostic covariates into the model for estimating survival [4–6] or for the probability of censoring [2, 3, 9, 10] to improve estimation of the marginal survival distribution. A few of them use information from the prognostic covariates directly without modeling survival or censoring [7, 8, 11] or use working models to summarize prognostic covariates [11] to define homogeneous risk groups to improve estimation of the marginal survival distribution.

The weighted Kaplan–Meier (WKM) approach [7, 8] utilizes the prognostic covariates to try to improve efficiency by defining homogeneous risk groups based on these covariates. The WKM approach can be considered as a nonparametric method, which incorporates prognostic covariates without modeling assumptions, and has been shown to be robust to modeling assumptions in a situation with few prognostic covariates. The method does however require that the covariates be categorical or that continuous covariates be categorized. However, in many clinical studies there are multiple prognostic covariates, which are either categorical or continuous. As the number of prognostic covariates increases, it is increasingly

[a]*Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, 1295 N Martin, PO Box 245211, Tucson, AZ 85724-5211, U.S.A.*
[b]*Arizona Cancer Center, University of Arizona, 1295 N Martin, PO Box 245211, Tucson, AZ 85724-5211, U.S.A.*
[c]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[*]*Correspondence to: Chiu-Hsieh Hsu, Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health,University of Arizona, 1295 N Martin, PO Box 245211, Tucson, AZ 85724-5211, U.S.A.*
[†]*E-mail: phsu@azcc.arizona.edu*

difficult to define homogeneous risk groups. The WKM method also does not differentiate between covariates, which may be strongly associated with the event time, from those that are only weakly associated. If the relationship between the prognostic covariates and the event time is correctly used to define homogeneous risk groups that have different conditional survival distributions, then the WKM estimator is more efficient compared with the standard KM estimator, which does not make use of the prognostic covariates. However, the relationship between the prognostic covariates and the event time might be misspecified or censoring might be dependent on the event time. In a situation with dependent censoring, the WKM method, which only focuses on the prediction of the event time using the covariates, may give risk groups that are not homogeneous with respect to the censoring distribution and can produce biased survival estimates. In this paper, we extend the WKM approach to handle a situation with multiple prognostic covariates and dependent censoring, with the primary emphasis on using the information on the association between the prognostic covariates and the event times to estimate the marginal survival distribution, while at the same time trying to minimize the impact of possible dependent censoring.

In [11], we used two working proportional hazards (PH) models to summarize the associations between prognostic covariates and event and censoring times into two risk scores. One is for the association between the covariates and the event times; the other is for the association between the covariates and the censoring times. The two risk scores are weighted to define the distance between subjects, which is then used to select an imputing risk set for each censored observation. The event times are imputed using a multiple imputation strategy including a Bootstrap stage. We called this estimate as Kaplan–Meier imputation with Bootstrap (KMIB). We showed that imputing event times for censored observation through the use of the two risk scores can induce a double robustness property in estimation of the survival distribution. Specifically, if one of the two working models is correctly specified, the survival estimate derived from the imputed data sets is consistent under defined conditions. Although the multiple imputation method has nice properties, the adequacy of imputation procedures will depend on the availability of possible donor observations, which diminishes in the tails of the survival distribution. This implies that the multiple imputation method could produce estimates with large variation in the tails of the survival distribution. In addition, it could involve intensively computational procedures to conduct multiple imputation, where each censored observation needs to be imputed multiple times. In this paper, we adapt the idea in [11] of fitting two working PH models to derive two risk scores. The two risk scores are then categorized into groups to define the homogeneous risk groups and the WKM method is applied. The WKM estimator derived from the risk groups does not have to deal with the limited availability of possible donor observations in the tails and does not need intensively computational procedures to recover information for censored observations. In addition, the WKM estimator is also expected to be robust to the misspecification of either one of the two working models; this will be explored in a simulation study.

In the prostate cancer application described in detail in Section 3, we consider a number of covariates (T-stage, prostate-specific antigen (PSA) and Gleason score), which are known to be prognostic for recurrence. Hence, the hope is that using the association between these covariates and recurrence time we could improve the efficiency of the estimated marginal time to recurrence distribution. In addition, the covariates may also be associated with the reasons that a participant is censored. Specifically, high values of these covariates would tend to be associated with the patient receiving additional treatment during the follow-up. Such additional treatments may impact when a patient drops out of the study or when they are considered censored. Hence, the data may be subject to dependent censoring, hence, we would like any estimation method not to be too greatly affected by dependent censoring.

In the application and simulation study, we will compare this WKM approach with a number of different alternatives. Robins and Finkelstein [9] proposed using the inverse probability of censoring weighted (IPCW) method, where the weight is derived from a PH model for censoring with auxiliary variables as the covariates. The IPCW method can be considered as a semi-parametric approach and can induce a property of double robustness locally after certain modifications [3]. Other methods we will be including are the KMIB multiple imputation method in [11] and the 'exact method' (denoted as ECS) where all of individual's survival estimates at each time point are derived from a PH model with the prognostic covariates as the covariates and then averaged to get the marginal (population) survival estimates [12].

This paper is organized as follows. In Section 2, we review the WKM and IPCW estimators, describe how to extend the WKM estimator to handle a situation with multiple prognostic covariates and dependent censoring and discuss the properties of the WKM estimator. In Section 3, we apply the WKM estimator to data from a prostate cancer study. In Section 4, we give results from simulation studies. A discussion follows in Section 5.

2. Method

2.1. WKM and IPCW estimators

Let T denote the event time, C denote the censoring time, $X=\min(T,C)$ and $\delta=I(T\leqslant C)$. For illustration, we assume that the prognostic factor, Z, is a time-independent categorical covariate and takes on values $1,\ldots,K$. The survival

Statistics in Medicine

function can be written as

$$S(t)=P(T>t)=\sum_{k=1}^{K}P(T>t|Z=k)P(Z=k)=\sum_{i=k}^{K}S_k(t)\theta_k,$$

where θ_k is the probability that a subject has covariate value $k,(k=1,\ldots,K)$ and $S_k(t)$ is the probability of survival conditional on having covariate value k. Based on the above expression, the WKM estimator is defined as $\mathrm{WKM}(t)\stackrel{\mathrm{def}}{=}\sum_{k=1}^{K}\hat{S}_k(t)n_k/n$, where $\hat{S}_k(t)$ is the KM estimator among those with covariate value k, n_k is the number of subjects in group k, and $n=\sum_{k=1}^{K}n_k$. For stability of the estimate it is desirable to avoid groups with small values of n_k. The asymptotic variance is equal to the sum of the weighted averages of within-group variation (variance of each $\hat{S}_k(t)$, where $k=1,\ldots,K$) and between-group variation (variation between $\hat{S}_k(t)$, $k=1,\ldots,K$) as follows:

$$\mathrm{var}(\sqrt{n}\mathrm{WKM}(t))=\sum_{k=1}^{K}\theta_k S_k(t)^2\int_0^t\frac{\lambda_k(u)\,\mathrm{d}u}{H_k(u)S_k(u)}+\sum_{k=1}^{K}\theta_k(S_k(t)-\bar{S}(t))^2,$$

where $\bar{S}(t)=(1/K)\sum_{k=1}^{K}S_k(t)$ and the $\lambda_k(.)$ and $H_k(.)$ are hazard and cumulative hazard functions, respectively. The first term of the variance can be easily estimated by calculating the weighted average of the variances derived from the Greenwood's formula for those K groups and the second term can be estimated by plugging in estimates of each component. We will be extending this method to the situation where the K groups are defined by two linear combinations of covariates.

The above asymptotic variance of the WKM estimator are provided by Malani [7] and Murray and Tsiatis [8], who show that the WKM estimator is consistent under an assumption that censoring is independent conditional on Z. The WKM estimator is consistent only up to the minimum event time of the K latest event times [8], which consists of the latest event time from each of the K groups of Z. Furthermore, it is only defined up to the minimum event time of the K latest event or censoring times. This indicates that the range of times for which the WKM estimate is defined decreases with K. In a situation that the latest time ($t_{\max}$) for a risk group k is less than the time of interest (t) and the sample size is small, we assume $\hat{S}_k(t)=\hat{S}_k(t_{\max})$ because this assumption has better performance than other methods in a small sample size situation [13].

For the IPCW method, we use the expression of point estimator [9, Equation 10]

$$\hat{S}(t)=\prod_{\{i;X_i<t\}}\left(1-\frac{\delta_i\hat{W}_i(X_i)}{\sum_{j=1}^{n}Y_j(X_i)\hat{W}_j(X_i)}\right)$$

where $Y(u)=I(X\geqslant u)$ is the at-risk indicator and $\hat{W}_i(X_i)=\hat{K}_i^0(X_i)/\hat{K}_i^Z(X_i)$ is the subject-specific weight at time X_i for subject i. $\hat{K}_i^0(X_i)$ is the usual Kaplan–Meier estimator of the probability being uncensored by time X_i and $\hat{K}_i^{Z_i}(X_i)$ is the conditional probability of being uncensored by time X_i given Z_i derived from a PH model for censoring time using the prognostic variables Z as the covariates. The expressions of standard errors for the IPCW method involve complicated formulas and can be found in the appendix of [9].

2.2. *WKM with multiple covariates*

It is problematic to extend the WKM method to the situation of p prognostic covariates, especially if p is large and some of the covariates are continuous. One potential solution is to dichotomize each prognostic covariate into two groups and then derive the WKM estimator based on the resulting 2^p categorized groups. With this strategy the number of groups increases with the number of the covariates, which could be problematic for both consistency and variation of the estimator, especially in a situation with a small sample size.

In this paper we adapt the ideas in [11] to incorporate multiple covariates into the WKM method. Let $\mathbf{Z}=\{\mathbf{z_1},\ldots,\mathbf{z_p}\}$ denote the p prognostic covariates. They could be either categorical or continuous. We first propose to fit a working PH model to the observed failure time to reduce the covariates to a single risk score, which provides a summary measure of an individual's risk of failure. The risk score effectively weights each covariate in the linear combination, with more weight given to the more important covariates. In a situation with dependent censoring, the WKM method with risk groups that are not homogeneous with respect to the censoring distribution could produce biased survival estimates. Therefore, we will also investigate a second PH model that calculates risk scores by summarizing the association between the covariates and the censoring time. The risk scores are defined as $\widehat{\mathrm{RS}}_f=\mathbf{Z}\hat{\boldsymbol{\beta}}$ for the failure time model and $\widehat{\mathrm{RS}}_c=\mathbf{Z}\hat{\boldsymbol{\gamma}}$ for the censoring time model, where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ denote the estimates of the regression coefficients for the two working PH models. These two risk scores can be categorized separately and the two categorized risk scores can then be jointly used to define the risk groups. However, in a dependent censoring situation, these two risk scores could be highly

correlated and separate categorization of the two risk scores could lead to some groups with very small number of observations. To alleviate this sparseness problem and create groups with more homogeneous number of observations, we propose using principal component analysis on the two standardized risk scores (centered and scaled) to derive two orthogonal components (linear combinations of two risk scores) and then categorize these two components separately based on their percentiles into $I*J$ groups, where I is the number of categories for the first component and J is the number of categories for the second component. The WKM estimator can then be derived based on the $I*J$ categorized groups (denoted as $\mathrm{WKM}_{I,J}$). We would expect I to be greater than J and J may even be equal to 1. The above procedures can be summarized into the following six steps. Step 1: fit a PH model to the observed failure time and the observed censoring time, respectively. Step 2: calculate the risk score for both PH models. Step 3: standardize the two risk scores, respectively, by subtracting the mean and dividing by the standard deviation. Step 4: perform principal component analysis on the two standardized risk scores to generate two orthogonal components. Step 5: categorize the two components to define risk groups. Step 6: perform the WKM on the risk groups derived from Step 5.

2.3. *Properties of the WKM estimator*

Malani [7] and Murray and Tsiatis [8] established theoretical properties of the WKM estimator. In particular they showed that the WKM is consistent when censoring is independent. Hsu *et al*. [11] showed that when one of the two working models is correctly specified, censoring is independent conditional on the two risk scores $(\mathrm{RS}_f, \mathrm{RS}_c)$. The two orthogonal components derived from principal component analysis are one to one transformation of the two risk scores because they are simple linear combinations of $(\mathrm{RS}_f, \mathrm{RS}_c)$. As a result, censoring is also independent conditional on the two orthogonal components. For an intuitive explanation of why conditionally independent censoring is sufficient, it is simplest to consider the case of categorical risk scores. Within a group that is defined using the two risk scores, the event times are independent of the censoring times. Thus, the Kaplan–Meier estimator is consistent within each group, as the number of observations in the group increases. Therefore, the WKM is also consistent. In a situation that the two risk scores are continuous and the sample size is finite, the WKM estimator based on the categorical groups may have a small bias because censoring will be close to, but not exactly, independent within each group. The magnitude of this bias will be explored in simulation studies. We note that the estimator also has the double robustness property in that only one of the models defining the risk scores needs to be correct.

It has been shown [14, 15] that to first order the estimates of the regression coefficients when fitting a PH model are consistent up to a constant factor, i.e. they are proportional to the regression coefficient of an accelerated failure time (AFT) model, when the true model is from an AFT family. Note that in the construction of the WKM both RS_f and RS_c are rescaled. Thus, the definition of the risk groups in WKM is not dependent on the absolute magnitude of β and γ, but is dependent on the relative magnitude of the individual coefficients within β and γ. It is exactly the relative magnitude of the coefficients which is robust to the link misspecification. Thus, the WKM estimator that uses two risk scores from two working PH models to define risk groups can produce reasonable survival estimates when the true failure and censoring time models are AFT models. The effect of misspecification of link functions of the two working PH models on the WKM estimator and the other estimators will be investigated in simulations.

3. Illustration of the method on prostate cancer data set

We demonstrate the WKM approach on a prostate cancer data set. The data consist of 503 patients with localized prostate cancer, who underwent external-beam radiation therapy at the University of Michigan and affiliated institutions between July 1987 and February 2000. Post-treatment each patient was followed for clinical recurrence, including local, regional, or distant metastases. This data set has been previously used to develop individualized prediction models of disease progression using serial PSA [16–18]. Patients were excluded from this analysis if their total radiation dose was not between 50 Gy and 100 Gy. For each patient several baseline characteristics (e.g. Age, T-Stage, PSA, and Gleason score) were measured. Of the 503 patients, the mean age was 69.0 years old, the mean baseline PSA was 14.8 and the mean total radiation dose was 71.9 Gy. The percentage with Gleason scores of $\leqslant$5, 6, 7, and $\geqslant$8 were 19, 36, 37 and 8 per cent, respectively. The median follow-up time was 5.72 years.

It is well known that PSA, Gleason score and T-stage are prognostic covariates associated with time to recurrence, with high values of each associated with higher risk. To demonstrate the WKM approach, baseline PSA value, Age, Gleason score, Total Radiation Dose and T-Stage are treated as prognostic time-independent covariates in the two working PH models. The results for estimation of those two working models are provided in Table I. Gleason score, T-Stage and Total Radiation Dose are significantly associated with failure time. Log-transformed baseline PSA, T-Stage and Total Radiation Dose are significantly associated with censoring time. The presence of the same significant variables in both models does indicate the potential for dependent censoring for these data. The risk scores derived from the two working

Statistics in Medicine

Table I. Data analysis: estimation of two working Cox PH models.

	Failure time model			Censoring time model		
Covariates	Estimate	SE	p-value	Estimate	SE	p-value
Age	−0.024	0.0170	0.15	0.031	0.0078	0.71
Log(PSA)	0.173	0.1267	0.17	−0.115	0.0536	0.03
Gleason	0.405	0.1037	<0.01	0.092	0.0494	0.06
T-Stage	1.355	0.2184	<0.01	−0.679	0.0849	<0.01
Total Dose	−0.111	0.0300	<0.01	0.176	0.0141	<0.01

Table II. Data analysis: estimation of recurrence-free probability at 5 years and 10 years.

	$t=5$ years		$t=10$ years	
Method	$\hat{S}(t)$	SE*	$\hat{S}(t)$	SE*
PO	0.852	0.0175	0.742	0.0285
ECS	0.866	0.0136	0.791	0.0188
IPCW	0.868	0.0157	0.770	0.0279
KMIB	0.864	0.0164	0.757	0.0317
$WKM_{2,2}$	0.865	0.0160	0.788	0.0227
$WKM_{4,1}$	0.869	0.0156	0.799	0.0208
$WKM_{4,2}$	0.868	0.0157	0.797	0.0212
$WKM_{4,4}$	0.866	0.0158	0.791	0.0218
$WKM_{8,1}$	0.870	0.0155	0.799	0.0208
$WKM_{8,2}$	0.868	0.0156	0.798	0.0211
$WKM_{16,1}$	0.869	0.0156	0.796	0.0213

*Estimated standard error.

models are used to define risk groups to obtain the WKM estimator. The two risk scores are highly correlated with Spearman's correlation coefficient of −0.77. This suggests that dependent censoring exists in the data. Based on principal component analysis, about 89 per cent of variation of the two risk scores is explained by the first principal component.

The results for estimating the recurrence-free probability are provided in Table II and Figure 1. Table II displays selected estimates from the partially observed (PO) analysis, that is the standard analysis of the observed censoring event time data, and from the WKM methods. We have also included the ECS, KMIB and IPCW methods for comparison. For the KMIB method, a weight of 0.8 and 0.2 is used for failure and censoring risk scores, respectively, when defining the distance between subjects, and the size of the imputing risk set for each censored observation is set at 5. All of the $WKM_{I,J}$ methods, as well as the ECS, IPCW and KMIB methods, produce slightly higher estimated survival compared with the PO analysis. The ECS, IPCW and WKM methods also produce lower associated estimated standard errors compared with the PO analysis. The KMIB method produces higher associated estimated standard errors in the tail compared with the PO analysis. The IPCW method produces similar estimated survival and associated estimated standard error early on but lower survival and greater associated estimated standard error at longer follow-up times compared with the $WKM_{I,J}$ method. The ECS method produces similar estimated survival but lower associated estimated standard errors compared with the $WKM_{I,J}$ method. Figure 1 displays the estimated survival curves for the PO analysis, the ECS and IPCW methods and the $WKM_{8,1}$ method. The ECS, IPCW, KMIB and $WKM_{8,1}$ methods consistently produce slightly higher estimated survival compared with the PO analysis, especially the ECS and $WKM_{8,1}$ methods. This indicates that the ECS, IPCW, KMIB and WKM methods all have potential to reduce bias due to dependent censoring through using information from the prognostic covariates.

4. Simulation study

We performed several simulation studies to investigate the properties of the WKM methods derived from the two risk scores. We consider a situation with multiple time-independent prognostic covariates. We investigate the effects of censoring mechanism, sample size, number of groups for categorizing the two components, misspecification of one of the two working models and misspecification of the two link functions on survival estimates.

For each of 1000 independent simulated data sets, there are five hypothetical prognostic covariates $(Z_1, \ldots, Z_5)$ independently generated from a $U(0,1)$ distribution. In a situation that the link functions are correctly specified for the two working PH models, i.e. the true failure and censoring time models are from a PH family, the event time

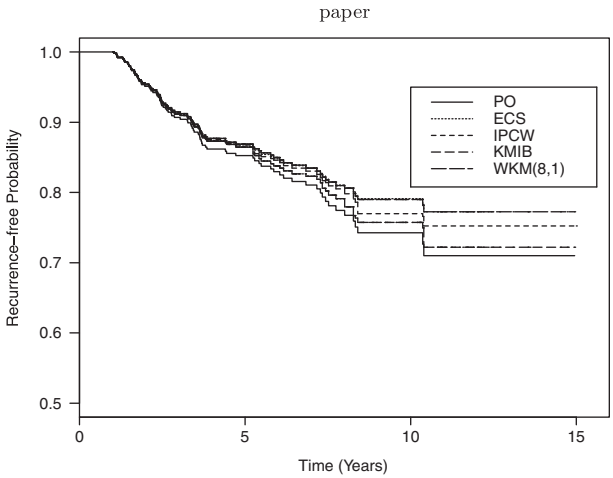

Figure 1. Prostate Cancer Study: Recurrence-free curves derived from the methods considered in this paper.

is generated from a hypothetical PH model conditional on prognostic covariates, where the hazard function is $\lambda(t)=4*t^3*\exp(-2.0Z_1+0.5Z_2-2.0Z_3+2.0Z_4+2.0Z_5)$, the censoring time is generated from Exponential(0.6) under an independent censoring scenario and the censoring time is generated from a hazard function $\lambda_c(t)=3*t^2*\exp(-3.0Z_1+0.5Z_2-2.0Z_3+1.5Z_4+2.0Z_5)$ under a dependent censoring scenario. In a situation that the link functions are incorrectly specified for the two working PH models, i.e. the true failure and censoring time models are not from a PH family, the event time is generated from lognormal$(0.1-2Z_1+0.5Z_2-2Z_3+2Z_4+2Z_5, 1)$ and the censoring time is generated from lognormal$(0.08-2.5Z_1+0.5Z_2-2Z_3+2.5Z_4+2Z_5, 1)$ and only dependent censoring is considered. We note that the coefficients in the linear combinations are quite similar, this will lead to a strong dependent censoring and provide challenging situations in which to evaluate the properties of the methods. We use Spearman's correlation coefficient between the event time and the censoring time to summarize the strength of dependency.

For the 'Fully Observed' (FO) analysis, treated as the gold standard, we derive KM estimates for each simulated data set before any censoring is applied. For the 'PO' analysis, we derive KM estimates with censoring. The estimate of standard error for both FO and PO analysis is based on Greenwood's formula. For the ECS method, either all five prognostic covariates (Z_1–Z_5) or only the first three prognostic covariates (Z_1–Z_3) are included in the PH model for the event time to derive the expected marginal survival (denoted as ECS_5 and ECS_3, respectively). For the IPCW method, either all five prognostic covariates (Z_1–Z_5) or only the first three prognostic covariates (Z_1–Z_3) are included in the PH model for the censoring time to derive the weight (denoted as IPCW_5 and IPCW_3, respectively). For the KMIB method, we define an imputing risk set for each censored observation using the risk scores derived from the two working PH models with prognostic covariates as covariates. Based on the results in [11], the number of imputation is set at 10, the size of each imputing risk set is set at 5 and the weights on risk scores derived from working failure time and censoring time models are set at 0.8 and 0.2, respectively. For the WKM method, we derive WKM estimates based on the risk groups defined using the risk scores derived from the two working PH models with prognostic covariates as covariates. For both KMIB and WKM, when the working failure time model or working censoring model is misspecified only the terms for Z_1, Z_2 and Z_3 are included in fitting the working model. The standard errors are calculated using the formula in Section 2.1. The performance of these methods for estimating survival at two time points, the median survival time and the 65th percentile survival time (i.e. t such that $S(t)=0.35$), are investigated. The simulation results for independent censoring are shown in Table I‡ of supplementary material and the findings discussed. The results show that all methods except ECS reasonable coverage rate and that the WKM method recovers between 15 and 24 per cent of information lost due to censoring while estimating the marginal survival and the IPCW method recovers about 14 per cent of the lost information.

In a situation with dependent censoring and the link functions correctly specified (Table III), the PO analysis as expected produces biased survival estimates in all situations and a lower coverage rate, which decreases as sample size increases (supplementary Table II). The performance of the ECS and IPCW methods depends on the number of covariates included in fitting the PH model for event/censoring time. When all five covariates are included in the model, the point survival estimates are comparable to the true values (0.5 and 0.35) for both the ECS_5 and IPCW_5 methods and

‡*Supporting information may be found in the online version of this article.*

Table III. Monte Carlo results for five time-independent covariates with dependent censoring (censoring rate=35 per cent) and the link functions correctly specified: the effects of number of risk groups. $n=200$. Spearman's $\rho=0.48$.

	True value: 0.5					True value: 0.35				
Method	Est*	SD†	MSE‡	SE§	CR¶	Est	SD	MSE	SE	CR
FO	0.501	0.0352	0.0012	0.0353	94.6	0.350	0.0343	0.0012	0.0336	94.0
PO	0.568	0.0383	0.0061	0.0394	58.6	0.426	0.0404	0.0074	0.0411	53.7
ECS_3	0.534	0.0413	0.0029	0.0380	82.5	0.385	0.0401	0.0028	0.0370	84.8
ECS_5	0.503	0.0398	0.0016	0.0320	88.3	0.353	0.0370	0.0014	0.0300	89.9
$IPCW_3$	0.538	0.0434	0.0033	0.0407	83.6	0.387	0.0461	0.0035	0.0408	82.5
$IPCW_5$	0.503	0.0432	0.0019	0.0402	92.8	0.351	0.0442	0.0020	0.0383	90.5
Both working PH models correctly specified										
KMIB	0.513	0.0407	0.0018	0.0404	93.5	0.364	0.0405	0.0018	0.0394	92.5
$WKM_{4,1}$	0.508	0.0411	0.0018	0.0406	94.0	0.357	0.0393	0.0016	0.0388	94.4
$WKM_{4,2}$	0.513	0.0418	0.0019	0.0402	92.8	0.365	0.0408	0.0019	0.0389	92.6
$WKM_{4,4}$	0.518	0.0418	0.0021	0.0392	90.9	0.376	0.0414	0.0024	0.0384	88.9
$WKM_{8,1}$	0.506	0.0409	0.0017	0.0397	93.7	0.361	0.0411	0.0018	0.0385	93.1
$WKM_{8,2}$	0.516	0.0414	0.0020	0.0392	91.6	0.374	0.0413	0.0023	0.0383	89.7
$WKM_{16,1}$	0.512	0.0409	0.0018	0.0391	93.0	0.371	0.0411	0.0021	0.0381	90.7
Only working failure PH model mis-specified										
KMIB	0.521	0.0408	0.0021	0.0406	90.3	0.372	0.0410	0.0022	0.0400	91.8
$WKM_{4,1}$	0.513	0.0417	0.0019	0.0411	92.9	0.363	0.0402	0.0018	0.0396	94.1
$WKM_{4,2}$	0.514	0.0414	0.0019	0.0404	93.3	0.367	0.0406	0.0019	0.0391	93.5
$WKM_{4,4}$	0.519	0.0417	0.0021	0.0394	90.7	0.377	0.0409	0.0024	0.0385	88.3
$WKM_{8,1}$	0.511	0.0420	0.0019	0.0401	92.6	0.366	0.0419	0.0020	0.0391	93.1
$WKM_{8,2}$	0.517	0.0416	0.0020	0.0394	91.3	0.376	0.0415	0.0024	0.0385	89.1
$WKM_{16,1}$	0.518	0.0417	0.0021	0.0393	91.8	0.377	0.0412	0.0024	0.0385	88.0
Only working censoring PH model mis-specified										
KMIB	0.514	0.0407	0.0019	0.0402	92.7	0.365	0.0404	0.0019	0.0393	93.0
$WKM_{4,1}$	0.512	0.0411	0.0018	0.0407	93.4	0.361	0.0396	0.0017	0.0392	94.4
$WKM_{4,2}$	0.513	0.0409	0.0018	0.0401	92.9	0.365	0.0407	0.0019	0.0388	93.1
$WKM_{4,4}$	0.518	0.0407	0.0020	0.0393	91.9	0.375	0.0408	0.0023	0.0384	89.5
$WKM_{8,1}$	0.510	0.0410	0.0018	0.0399	93.4	0.363	0.0402	0.0018	0.0388	93.5
$WKM_{8,2}$	0.515	0.0412	0.0019	0.0392	91.8	0.373	0.0408	0.0022	0.0383	90.3
$WKM_{16,1}$	0.515	0.0411	0.0019	0.0392	92.2	0.373	0.0403	0.0022	0.0382	90.6

*Average of 1000 point estimates.
†Empirical standard deviation.
‡Mean square error: $\text{bias}^2+\text{SD}^2$.
§Average estimated standard error.
¶Coverage rate of 1000 95 per cent confidence intervals.

the coverage rates are comparable to the nominal level (95) for the $IPCW_5$ method but slightly lower than the nominal level for the ECS_5 method due to lower estimates of standard errors compared with the SD. The mean square error (MSE) for the ECS_5 and $IPCW_5$ methods is similar to the MSE of the KMIB and WKM methods. When only the first three covariates are included in the model, for both the ECS_3 and $IPCW_3$ methods the point estimates are higher than the true values and the coverage rates are lower than the nominal level. The MSE for the ECS_3 and $IPCW_3$ methods is higher compared with the KMIB and WKM methods. The performance of the $WKM_{I,J}$, where the risk groups are derived from the principal component analysis, depends on the number of groups and sample size. In general, as the number of the first component (I) increases, the bias decreases. However, when the number is larger than some number (e.g. 8), this bias starts to increase again. When I is fixed, as the number of the second component (J) increases, the bias increases. This indicates that the first component is more important in defining the risk groups. As for the magnitude of bias, it remains similar across all three situations of misspecification of the two working PH models. For the $WKM_{I,J}$ method, the magnitude of bias decreases with sample size (supplementary Table II). However, the rate is slow. The KMIB method produces slightly higher point estimates than the WKM method, especially when the working failure time model is misspecified. The KMIB and WKM methods produce similar MSE in all three scenarios of working model misspecification.

In a situation with dependent censoring and the link functions incorrectly specified, the results are presented in the supplementary material (Table III). We find that all methods have some bias, which decreases with sample size for the

WKM and KMIB methods. The IPCW method has poor efficiency with underestimated SE leading to lower coverage rates.

In summary, the ECS method in which the marginal survival estimator is derived from a PH model for failure time can produce reasonable survival estimates when the true failure time model is from a PH family and all prognostic covariates are included in the model. The IPCW method in which the weight is derived from a PH model for censoring time can produce reasonable survival estimates when the true censoring time model is from a PH family and all prognostic covariates are included in the model. The WKM method in which the risk groups are derived from two risk scores based on two working PH models can provide reasonable survival estimates and is robust to misspecification of either one of the two working models and is robust to misspecification of the link functions of failure time and censoring time models. The KMIB method in which the imputing risk sets are derived from two risk scores based on two working PH models can provide reasonable survival estimates and is robust to misspecification of either one of the two working models but the bias is consistently higher than for the WKM method.

5. Discussion

The research in this paper generalizes the WKM estimator, which is often used to incorporate prognostic covariates into survival analysis, to handle a situation with multiple prognostic covariates and potential dependent censoring through the use of prognostic covariates. This approach is expected to have weak reliance on a statistical model, because the model is only used to identify risk groups. Once the risk groups are defined, the WKM estimator is conducted on the risk groups. The simulation study shows that the use of this WKM method can lead to improved performance of estimators. In general, the WKM point estimates are less variable and closer to the truth than the estimates produced by analyzing the observed data without using the prognostic covariates.

The form of the IPCW method in [9] does not have the property of double robustness. As a result, when the censoring time model is misspecified, the point estimate is associated with bias as seen in simulation results. As pointed out in [3], the double robustness property can be induced after modifying the survival estimator when the true model is from a PH family. In a situation with the true model is from an AFT family, we suspect that the inverse probability of censoring weighted method is biased because the misspecified link function is utilized in construction of the estimator. The estimated standard errors can also be expected to be biased.

The major reason for the remaining bias in the WKM method in the case of dependent censoring is the sample size. In particular, a risk group contains some individuals whose risks are not close enough to the majority of the group, so some remnants of dependent censoring remain within the risk group. This is likely to be more of a problem with high-dimensional covariates compared with cases with less than say 5 prognostic covariates. An additional complication with high-dimensional covariates is that it will be hard to obtain good estimates of the coefficients in the working models with many covariates, making it even harder to define a risk group that truly has similar risk for each individual within the group.

Theoretical arguments for large samples indicate that defining the risk groups allows for good estimation from the WKM method, even with dependent censoring. Numerical results indicate that when the working model for the event time is misspecified, the bias is greater than when it is correctly specified. In addition, we also observed more gains in efficiency when the failure time model is correctly specified. Thus, although double robustness is a very useful property, it should not be used as a replacement for trying to find reasonable fitting working models for both the failure time and the censoring time, rather it should be used in addition to seeking good models for the observed data.

The two risk scores from the event time and the censoring times models are given equal emphasis in our approach, because they are both standardized prior to the principal component analysis. Extensions of the method which give more or less emphasis to the censoring time risk score depending on the perceived impact of the dependent censoring would be interesting to investigate.

In this paper, we demonstrate that the bias of $\mathrm{WKM}_{I,J}$ can increase with I or J. This is due to both the 'nearness' of the risk group and the tail problem for each risk group, where the Kaplan–Meier estimator is only defined up to the latest event time if the largest observed time is censored. The 'nearness' of the risk group mainly depends on how many risk groups are defined by categorizing the two risk scores. The tail problem mainly depends on the sample size. As I or J increases, a more homogeneous risk group can be found, but it could increase the chance of having the tail problem for all risk groups due to the small sample size for each risk group. Hence, the total number of risk groups depends on the sample size. The results in simulation indicate a minimum sample size of 30 for each risk group performs better overall. As for how to choose the ratio between I and J, it can be decided by calculating the percentage of variation for the two standardized risk scores explained by each component. For example, in a situation that the first component explains 80 per cent of the variation, a ratio of 4 to 1 can be selected for I and J.

References

1. Heitjan DF. Ignorability in general incompleted-data models. *Biometrika* 1994; **81**:701–710.
2. Robins JM, Rotnitzky A. Recovery of informationand adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*: *Methodological Issues*, Jewell N, Dietz K, Farewell V (eds). Birkhauser: Boston, 1992; 297–331.
3. Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the Biopharmaceutical Section*, American Statistician Association, 1993; 24–33.
4. Finkelstein DM, Schoenfeld DA. Analysing survival in the presence of an auxiliary variable. *Statistics in Medicine* 1994; **13**:1747–1754.
5. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* 1994; **13**:955–968.
6. Gray RJ. A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* 1994; **81**:527–539.
7. Malani HM. A modification of the redistribution to the right algorithm using disease markers. *Biometrika* 1995; **82**:515–526.
8. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; **52**:137–151.
9. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics* 2000; **56**:779–788.
10. Satten GA, Datta S, Robins JM. An estimator for the survival function when data are subject to dependent censoring. *Statistics and Probability Letters* 2001; **54**:397–403.
11. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine* 2006; **25**:3503–3517.
12. Ederer F, Axtell L, Cutler S. The relative survival rate: a statistical methodology. *National Cancer Institute Monograph* 1961; **6**:101–121.
13. Klein JP. Small-sample moments of some estimators of the variance of the Kaplan–Meier and Nelson–Aalen estimators. *Scandinavian Journal of Statistics* 1991; **18**:333–340.
14. Solomon PJ. Effect of misspecification of regression models in the analysis of survival data. *Biometrika* 1984; **71**:291–298; Amendment (1986) **73**:245.
15. Struthers CA, Kalbfleisch JD. Misspecified proportional hazards models. *Biometrika* 1986; **73**:363–369.
16. Law NJ, Taylor JMG, Sandler HM. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* 2002; **3**:547–563.
17. Yu M, Law NJ, Taylor JMG, Sandler HM. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* 2004; **14**:835–862.
18. Taylor JMG, Yu M, Sandler HM. Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* 2005; **23**:816–825.