Weighted log-rank test to compare two survival functions in the presence of dependent censoring

Philippe Flandre

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Weighted log-rank test to compare two survival functions in the presence of dependent censoring

Philippe Flandre

[1]Sorbonne Université, INSERM Institut Pierre Louis d'Epidémiologie et de Santé Publique (IPLESP), Paris, France

Correspondence
INSERM UMR-S1136, 56 boulevard Vincent Auriol, CS 81393, 75646 PARIS Cedex 13, France E-mail: philippe.flandre@iplesp.upmc.fr

Summary

Comparing survival functions with the log-rank test in the presence of dependent censoring can produce an invalid test result. We extend our previous work on the estimation of the survival function using prognostic variables to adjust for dependent censoring to the comparison of two survival distributions. In these estimators, the weights of a censored individual is redistributed among either a subset of patients in the risk set or all patients in the risk set but giving more weight to patients having smallest distances from the censored subject. The distance is based on two risk scores obtained from two working models, one for the failure time and one for the censoring time. Based on the estimators, we suggest a weighted log-rank test to compare two survival distributions. A simulation study compared performance of our method with the analysis of the observed data without using auxiliary variables and with a recent method based on multiple imputation (KMIB method). With appropriate parameters, the weighted log-rank approach provides sizes of the test comparable to the nominal value but higher powers than the two other methods. The method is illustrated with data from a breast cancer study.

KEYWORDS:
dependent censoring; weighted log-rank test; auxiliary variables; weighted Kaplan-Meier

1 | INTRODUCTION

Most standard statistical methods for the analysis of right-censored survival data fail to be consistent unless the censoring is independent of survival.[1] However, it is likely that the censoring mechanism is not independent of survival in some applications of survival analysis methods. Some mechanisms are very likely to yield dependent censoring, e.g., censoring due to subjects selectively dropping out of the study. Dependent censoring will lead to bias in survival estimate and then difficulties in performing nonparametric comparisons between two groups.[2,3] Then, log-rank test could be invalid. In this paper, we suggest a weighted log-rank test in the presence of dependent censoring. Our approach is based on a previous work introducing weighted Kaplan-Meier estimators for dependent censoring in the presence of continuous prognostic factors.[4]

In survival analysis, several approaches have been proposed to recover some of the loss of information due to censoring using prognostic covariates. A few of these methods use information from the prognostic covariates directly without modelling survival or censoring[5,6,2] or use working models to summarize prognostic covariates[2,3] to define homogeneous risk groups to improve estimation of the marginal survival distribution. Another approach was to incorporate the probability of censoring to improve the estimation of the marginal survival distribution.[7,8] Recently, a proportional hazards model using copulas and penalised likelihood has been introduced under dependent censoring.[9] Malani suggests a modification of the redistribution to

the right algorithm[10] as a new approach to recover information for censored individuals, the weighted Kaplan-Meier (WKM) approach.[5] The modification generalizes the Kaplan-Meier method in the presence of auxiliary information available from a disease marker. The method has been introduced when the disease marker is categorical. Recently, in the case of continuous markers, WKM estimators have been also introduced based on a modification of the redistribution to the right algorithm.[4] In this approach, the weight of a censored individual is redistributed among either a subset of patients in the risk set or all patients in the risk set but giving more weight to patients having smallest distances from the censored subject. The subset of patients correspond to the q neighbours having the smallest distances from the censored subject. The distance is based on two risk scores derived from two working proportional hazards (PH) models, one for the failure time and one for the censoring time.[4]

Hsu and colleagues introduced a multiple imputation procedure to impute event times for the censored observations, the KMIB (Kaplan-Meier Imputation bootstrap) approach.[2] They proposed using two risk scores, derived from two working proportional hazards (PH) models as introduced above, to define a neighborhood to impute event times for each censored observation. They showed that such a method can both reduce the bias due to dependent censoring and increase the efficiency compared with standard estimates.[2] They extended their work to propose a log-rank test from the multiple imputation method in the case of the comparison of two survival distributions.[11] Their work is described in details elsewhere[2,11] and, to facilitate its use, an R package, entitled InformativeCensoring, has been developed.

In the present study, based on our previous work[4], we suggest a weighted log-rank tests to compare two survival functions. This paper is organized as follows. In section 2, we briefly describe the WKM method and how to construct the weighted log-rank statistic. In section 3, we investigate properties of our weighted log-rank approach, compared with the statistics based on the KMIB procedure, in finite sample sizes though a simulation study. In section 4, we apply the methods to data from a breast cancer study. Some elements for discussion are given in section 5.

2 | METHODS

2.1 | Notation

Let $(T_i, \delta_i, X_i, \mathbf{Z}_i)$, $i = 1, \ldots, N$, denote an independent sample of right-censored survival data of two groups, where T_i is the possibly right-censored event time, δ_i is the censoring indicator with $\delta_i = 0$ if T_i is censored and $\delta_i = 1$ if T_i corresponds to an event; X_i is the group indicator for the two different groups ($X = 0$ and $X = 1$) and $\mathbf{Z}_i$ is the covariate vector. We have two treatment groups with n_0 and n_1 subjects in each group, with $N = n_0 + n_1$. Let $t_{(1)}, < \ldots < t_{(p)}$ denote the distinct ordered values of T_i and $\delta_{(i)}$ the corresponding values of the δ_i's. Also let $w_{ik}(t)$ denote the weight associated with the ith individual in group k after redistribution at time t and $w_{ik}(t^-)$ the weight associated with the ith individual in group k prior to redistribution at time t.

2.2 | Weighted Kaplan-Meier for data with dependent censoring

The redistribution algorithm starts by assigning a weight of $1/n_k$ to all individuals in group k at time 0, i.e., $w_{ik}(0) = 1/n_k$ for all i in group k. Without any censoring the weight in each group remains unchanged until the end of the study and the survival function in each group drops by $1/n_k$ at each failure time. In the case of censoring, this process is continued by moving through the ordered failure times until the time of the first censored observation. At this time, in the case of independent censoring the weight of the censored individual is re-allocated equally to all remaining individuals in the risk set belonging to the same group. This is done as follows $w_{ik}(t) = w_{ik}(t^-) + \frac{w_{ik}(t^-)}{n_k(t)}$, where $n_k(t) = \sum_i I(T_i > t) I(X_i = k)$. The weights are modified similarly at each subsequent censoring[10] and the KM estimator in group k is simply $\hat{S}_k(t) = \sum_i I(T_i > t) I(X_i = k) w_{ik}(t)$.

In the case of dependent censoring, a weighted Kaplan-Meier approach has been described which is briefly introduced here for the two sample data.[4] The new approach is based on the Malani estimator that has been introduced for a discrete disease marker.[5] Suppose that we have several continuous markers recorded at patient's entry. First we have to reduce these markers values into a single value. The procedure can be summarized into the following steps. Step 1: fit a working PH model to the observed failure time and the observed censoring time (observed event times are treated as censored observations), respectively. Step 2: compute the risk score for both working models (RS_f for the failure model and RS_c for the censoring model). Step 3: perform principal component analysis (PCA) on the two standardized risk scores to generate two orthogonal components. Step 4: the first component ($pca1$) is used to calculate the neighbours closest to the component value of the censored individual. Step 5: perform the weighted Kaplan-Meier using the two approaches described below.

(i) redistributing the weight among the q neighbours closest to the censored individual, or

(ii) redistributing the weight among all the individuals but according to a function giving more weight to individuals close to the censored individual.

For the individual censored at time $t_{(l)}$, the first step consists of determining the genuine number of neighbours available. This number is equal to $q_k^{(l)} = \min(q, n_k(t_{(l)}))$, where q is the desired number of neighbours. The criterion to select the $q^{(l)}$ neighbours of the individual censored at time $t_{(l)}$ can be summarized by $d_i^{(l)} = \left| pca1_{ik} - pca1_{(l)k} \right|,\ i \in \mathcal{R}_k(t_{(l)})$ where $pca1_{(l)k}$ denotes the pca_1 value of the censored individual at time $t_{(l)}$ in group k, and $\mathcal{R}_k(t_{(l)})$ denotes the set of the individuals at risk at time $t_{(l)}$ in group k.

Defining $\mathcal{Q}_k^{(l)} \subset \mathcal{R}_k(t_{(l)})$ the set of the $q_k^{(l)}$ individuals in group k with a minimum criterion $d^{(l)}$, the weight $w_{(l)k}(t_{(l)})$ of the lth censored individual in group k is redistributed according to

$$\begin{aligned} \forall i \in \mathcal{Q}_k^{(l)}, \quad & w_{ik}(t_{(l)}) = w_{ik}(t_{(l)}^-) + w_{(l)k}(t_{(l)}^-) \cdot f(pca1_{ik}, pca1_{(l)k}) \\ \forall i \notin \mathcal{Q}_k^{(l)}, \quad & w_{ik}(t_{(l)}) = w_{ik}(t_{(l)}^-) \end{aligned} \tag{2.1}$$

where $f(pca1_{ik}, pca1_{(l)k})$ is a weighted function. A previous work has shown that $f(pca1_{ik}, pca1_{(l)k}) = 1/q_k^{(l)}, \forall i \in \mathcal{Q}_k^{(l)}$ which means that the weight is equally distributed among these $q_k^{(l)}$ neighbours, provides reasonable estimates of the survival function.[4] For example, suppose we have a sample size of 100 in group k then $w_{ik}(0)$=1/100=0.01 for all i in group k. Suppose the desired number of neighbours is 4.The weight of the first individual censored is still 0.01 and the four observations having the closest pca1 value to the pca1 value of that censored individual will receive an additional weight of 0.0025 (0.01/4). Then the weight for this 4 observations is now 0.0125.

The second suggestion was to consider that the weight of a censored individual be redistributed among all individuals at risk but not equally, individuals having a pca1 value close to the value of the censored individual receiving more weight than the others. We explored two procedures for that redistribution, one based on the normal distribution the other one on the inverse of the distance $d_i^{(l)}$.

In the first procedure, the weight is then redistributed according to

$$\forall i \in \mathcal{R}_k(t_{(l)}), \quad w_{ik}(t_{(l)}) = w_{ik}(t_{(l)}^-) + w_{(l)k}(t_{(l)}^-) \cdot f(pca1_{ik}, pca1_{(l)k}, \sigma) \tag{2.2}$$

where $f(pca1_{ik}, pca1_{(l)k}, \sigma)$ is based on a normal distribution centered on $pca1_{ik}$

$$f(pca1_{ik}, pca1_{(l)k}, \sigma) = \frac{\exp\left\{ -\frac{(pca1_{ik} - pca1_{(l)k})^2}{2\sigma^2} \right\}}{\sum_{i \in \mathcal{R}_k(t_{(l)})} \exp\left\{ -\frac{(pca1_{i'k} - pca1_{(l)k})^2}{2\sigma^2} \right\}} \tag{2.3}$$

for the weight redistribution.

In the second procedure, the weight is then redistributed as above but with

$$f(pca1_{ik}, pca1_{(l)k}, p) = \frac{\left[1/d_i^{(l)})\right]^p}{\sum_{i \in \mathcal{R}_k(t_{(l)})} \left[1/d_i^{(l)})\right]^p} \tag{2.4}$$

where higher values of p redistribute more weights on observations having smaller distances.

All these estimators can be considered as weighted Kaplan-Meier estimates and estimating the survival function in group k at time t is still done by summing the corresponding weights of the individuals still at risk in group k at time t. Redistribution of the weights among the q neighbours, using a uniform distribution will be noted $\mathrm{WKM}_{U,x\%}$ with $x = q/N$ where q is the desired number of neighbours. The estimator based on the redistribution using a normal distribution is noted $\mathrm{WKM}_{\mathcal{N}(\sigma)}$ and the last estimator is noted $\mathrm{WKM}_{(1/d)^p}$.

2.3 | Log-rank test for data with dependent censoring

2.3.1 | Weighted log-rank test

The statistic of the test is derived in the similar way to the one proposed by Xie and Liu (2005).[12] As we have seen above, the weight of a censored individual is only redistributed among individuals at risk in the same group that the censored individual. At time t_j there are d_{jk} events out of Y_{jk} individuals at risk in group k. Then we can write $d_{jk} = \sum_{i:T_i=t_j} \delta_i I(X_i = k)$ and $Y_{jk} = \sum_{i:T_i \geq t_j} I(X_i = k)$. Denote $\bar{w}_k(t_j)$ the mean of weights among individuals still at risk at time t_j in group k. Then,

$\bar{w}_k(t_j) = \left(1/Y_{jk}\right) \sum_{i:T_i \geq t_j} w_{ik}(t_j)$. Denote $w^r_{ik}(t_j)$ the ratio between the individual weight and the mean weight for individuals in group k with $w^r_{ik}(t_j) = w_{ik}(t_j)/\bar{w}_k(t_j)$. Then an individual receiving weights from previous censored observations will have $w^r_{ik}(.) > 1$ whereas an individual who did not received weights will have $w^r_{ik}(.) < 1$. As in Xie and Liu (2005), the weighted number of events is

$$d^w_{jk} = \sum_{i:T_i=t_j} w^r_{ik}(t_j)\delta_i I(X_i = k)$$

But in contrast to Xie and Liu (2005), in our statistic the weighted number at risk in group k is equal to the number at risk in group k

$$Y^w_{jk} = \sum_{i:T_i \geq t_j} w^r_{ik}(t_j) I(X_i = k) = \sum_{i:T_i \geq t_j} \frac{w_{i0}(t_j)}{\left(1/Y_{jk}\right) \sum_{u:T_u \geq t_j} w_{u0}(t_j)} I(X_i = k) = Y_{jk}$$

Let $d_j = d_{j0} + d_{j1}$, $d^w_j = d^w_{j0} + d^w_{j1}$, and $Y_j = Y_{j0} + Y_{j1}$ denote the number of events, the weighted number of events and the number at risk in the combined sample at time t_j. The test of H_0 is based on the statistic

$$G^w = \sum_{j=1}^{p} \left(d^w_{j1} - Y_{j1}\left(\frac{d^w_j}{Y_j}\right)\right)$$

which is a weighted form of the standard log-rank test statistic. The variance of G^w based on similar arguments in Xie and Liu (2005)[12] is given by

$$\mathrm{Var}(G^w) = \sum_{j=1}^{p} \left\{ \frac{d_j(Y_j - d_j)}{Y_j(Y_j - 1)} \sum_{i=1}^{Y_j} \left[\left(\frac{Y_{j0}}{Y_j}\right)^2 (w^r_i)^2 X_i + \left(\frac{Y_{j1}}{Y_j}\right)^2 (w^r_i)^2 (1 - X_i) \right] \right\}$$

The weighted log-rank test statistic is proposed as $Z = G^w / \sqrt{\mathrm{Var}(G^w)}$. The test statistic has a standard normal distribution for large samples under the null hypothesis H_0.

2.3.2 | Tests based on the Kaplan-Meier Imputation Bootstrap approach

The KMIB approach is based on M imputed data sets and also on two working PH models, one model for the event times and one for the censoring times. The two scale-free risk scores are used to select an imputing risk set for each censored observation by defining the distance between subjects. The distance, based on the original data, between subject j and k is defined as

$$d(j,k) = \sqrt{w_f [R\hat{S}_f(j) - R\hat{S}_f(k)]^2 + w_c [R\hat{S}_c(j) - R\hat{S}_c(k)]^2}$$

where w_f and w_c are nonnegative weights that sum to 1. The imputing risk set, $R(j+,\mathrm{NN})$, for the censored subject j consists of NN subjects who have longer survival time than the censoring time of subject j and the NN smallest distances from the censored subject j . Given M imputed data sets, we can perform time to event statistical analyses on each data set. The results can be combined to give a single p-value estimate in two distinct ways:

- meth1: Each data set produces a single point estimate for the null hypothesis ($\theta = \theta_0$) and these can be combined to obtain a single point estimate $\hat{\theta}$ with associated variance $V_1 = U_1 + (1 + M^{-1})B_1$ where B_1 is the sample variance of the M point estimates and U_1 is the average of the M variance estimates. The test statistic $D = (\hat{\theta} - \theta_0)' V_1^{-1} (\hat{\theta} - \theta_0)$ has a F_{1,v_1} distribution with v_1 degrees of freedom.
- meth2: Each data set produces a (normal) test statistic $Z_1, Z_2, \ldots, Z_m$ and these can be averaged to give an overall test statistic $\bar{Z}$ with variance $V_2 = 1 + (1 + M^{-1})B_2$ where B_2 is the sample variance of Z_i. A t-test statistic with v_2 degrees of freedom can be used with the statistic $s = \bar{Z}/\sqrt{V_2}$.

We refer the reader to[2,11] for further details. As recommended by the authors we used NN=5, M=10, $w_f = 0.8$ and $w_c = 0.2$.

3 | SIMULATION STUDY

3.1 | Methods

We perform a simulation study to investigate the properties of our weighted log-rank test approach. We consider a situation with multiple prognostic variables, binary and continuous. For each independent simulated data sets, five hypothetical auxiliary

TABLE 1 Monte Carlo results: size (per cent) of log-rank tests with dependent censoring. Misspecification implies using Z_1, Z_2 and Z_3 instead of Z_1, Z_2, Z_3, Z_4 and Z_5 in one of the working model. Results are based on 10,000 replications

	N=200		N=400		N=800		N=1600	
	α_0							
Method	-0.2	0.4	-0.2	0.4	-0.2	0.4	-0.2	0.4
FO	4.8	5.4	5.2	5.3	4.9	5.1	5.3	5.0
PO	4.8	5.1	5.0	5.2	4.8	5.1	5.2	5.1
Both working PH models correctly specified								
KMIB meth1	4.6	4.8	4.6	4.4	4.3	4.7	4.8	4.3
KMIB meth2	4.6	4.9	4.7	4.5	4.4	4.9	4.9	4.5
$WKM_{U,2\%}$	5.3	5.5	5.6	5.6	5.7	5.8	5.7	5.7
$WKM_{U,80\%}$	5.2	5.6	5.4	5.6	5.3	5.4	5.4	5.4
$KM_{\mathcal{N}(0.05)}$	5.1	5.6	5.2	5.0	5.4	5.2	5.6	4.8
$KM_{\mathcal{N}(8)}$	4.9	5.3	5.1	5.4	5.0	5.2	5.3	5.1
$WKM_{(1/d)^5}$	5.3	5.5	5.0	5.1	5.1	5.2	5.4	5.0
$WKM_{(1/d)^7}$	5.1	5.3	4.8	4.9	5.0	4.9	5.4	4.7
Only working failure PH model misspecified								
KMIB meth1	4.5	5.1	4.6	4.8	4.4	4.8	4.8	4.3
KMIB meth2	4.6	5.2	4.7	5.0	4.5	5.1	4.8	4.6
$WKM_{U,2\%}$	5.1	5.9	5.8	5.7	5.7	6.2	6.2	6.2
$WKM_{U,80\%}$	5.0	5.5	5.3	5.5	5.2	5.4	5.3	5.4
$KM_{\mathcal{N}(0.05)}$	5.1	5.7	5.4	5.1	5.2	5.0	5.9	4.7
$KM_{\mathcal{N}(8)}$	4.9	5.3	5.1	5.3	4.9	5.2	5.3	5.1
$WKM_{(1/d)^5}$	5.1	5.5	5.2	5.1	5.0	5.5	5.4	4.9
$WKM_{(1/d)^7}$	4.9	5.6	5.1	4.9	5.0	5.1	5.2	4.6
Only working censoring PH model misspecified								
KMIB meth1	4.6	5.3	4.6	4.5	4.4	4.7	4.7	4.5
KMIB meth2	4.7	5.5	4.7	4.6	4.5	5.0	4.8	4.7
$WKM_{U,2\%}$	5.0	5.5	5.5	5.5	5.7	6.0	5.9	5.7
$WKM_{U,80\%}$	5.1	5.5	5.3	5.6	5.2	5.4	5.3	5.4
$KM_{\mathcal{N}(0.05)}$	5.0	5.5	5.5	5.5	5.7	6.0	5.9	5.7
$KM_{\mathcal{N}(8)}$	5.1	5.5	5.3	5.6	5.2	5.4	5.3	5.4
$WKM_{(1/d)^5}$	5.1	5.6	5.1	5.3	4.8	5.1	5.1	4.3
$WKM_{(1/d)^7}$	4.9	5.4	5.0	5.0	4.7	4.8	5.0	4.0
Censoring rate	32%	45%	32%	45%	32%	45%	32%	45%

variables are generated, three (Z_1, Z_3, Z_5) from a *Bernoulli*(0.5) distribution and two (Z_2, Z_4) from a *Uniform*(0,1) distribution. The event time is generated from the PH model $\lambda_f(t) = t^4 \exp(\psi \text{Trt} - 2.0Z_1 + 0.5Z_2 - 2.0Z_3 + 2.0Z_4 + 2.0Z_5)$, where ψ is set equal to 0 for the study of size and -0.75 and 0.75 for the study of power and Trt is the treatment indicator generated from a *Bernoulli*(0.5) distribution. To induce dependent censoring, the censoring time is generated from the PH model $\lambda_c(t) = t^3 \exp(\alpha_0 + \alpha_1\psi\text{Trt} + \psi\text{Trt} - 3.0Z_1 + 0.5Z_2 - 2.0Z_3 + 1.5Z_4 + 2.0Z_5)$. The censoring model leads to several levels of censoring between groups 0 (Trt=0) and 1 (Trt=1). For example with $\alpha_0 = 0.4$ and $\alpha_1 = 0.15$, the censoring rate is 45% in group 0 and 39%, 45% and 52% in group 1 when $\psi = -0.75$, 0 and 0.75, respectively.

We also focus on model misspecification. As in Hsu and Taylor[11], we consider situations where either both of the two working PH models are correctly specified or one of them is misspecified. Specifically, for each treatment group the working failure time model is either correctly specified or misspecified as $\lambda_{wf}(t) = \lambda_{0f}(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$, and the censoring time model is either correctly specified or misspecified as $\lambda_{wc}(t) = \lambda_{0c}(t)\exp(\gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3)$. For investigation of the size, the sample sizes are 200/400/800/1600 subjects, with half of patients in each group ($n_k = 100$, 200, 400 or 800) and 10,000 replications. To

TABLE 2 Monte Carlo results: power (per cent) analysis with dependent censoring. Misspecification implies using Z_1, Z_2 and Z_3 instead of Z_1, Z_2, Z_3, Z_4 and Z_5 in one of the working model (N=200). Results are based on 1,000 replications

	$\alpha_1 = 0.15$				$\alpha_1 = 0.75$			
	$\alpha_0 = -0.2$		$\alpha_0 = 0.4$		$\alpha_0 = -0.2$		$\alpha_0 = 0.4$	
	Treatment effect (ψ)							
Method	-0.75	0.75	-0.75	0.75	-0.75	0.75	-0.75	0.75
FO	63.5	61.5	59.8	59.7	62.6	64.3	63.6	60.4
PO	42.1	36.4	32.1	29.7	28.4	16.4	17.7	10.2
Both working PH models correctly specified								
KMIB meth1	53.9	46.2	38.7	33.0	47.8	28.2	32.0	15.4
KMIB meth2	54.3	47.1	39.3	33.8	48.3	29.6	32.4	15.6
$WKM_{U,2\%}$	60.5	56.0	52.0	48.2	59.6	52.8	53.0	38.1
$WKM_{\mathcal{N}(0.05)}$	59.9	55.7	49.4	46.6	59.8	51.4	51.4	36.9
$WKM_{(1/d)^5}$	59.6	55.0	51.0	47.9	59.0	50.7	51.8	37.5
$WKM_{(1/d)^7}$	59.5	54.4	50.7	47.2	58.7	51.0	51.6	37.2
Only working failure PH model misspecified								
KMIB meth1	51.0	43.1	35.6	31.9	44.3	24.2	27.7	12.3
KMIB meth2	51.2	43.3	36.6	32.9	44.6	24.8	28.6	13.0
$WKM_{U,2\%}$	58.2	54.3	48.1	43.1	56.4	41.4	45.6	26.9
$WKM_{\mathcal{N}(0.05)}$	56.4	53.9	47.4	42.4	55.8	40.8	45.1	26.7
$WKM_{(1/d)^5}$	56.4	52.4	46.8	42.5	55.8	39.2	42.9	26.5
$WKM_{(1/d)^7}$	56.2	52.1	46.1	42.5	55.9	39.4	42.8	26.7
Only working censoring PH model misspecified								
KMIB meth1	54.0	45.4	37.8	35.4	47.9	28.2	30.5	14.1
KMIB meth2	54.4	45.8	38.4	36.3	48.2	29.1	31.1	15.5
$WKM_{U,2\%}$	58.2	52.8	46.9	42.3	56.0	42.6	45.2	27.3
$WKM_{\mathcal{N}(0.05)}$	58.2	52.8	46.9	42.3	56.0	42.6	45.2	27.3
$WKM_{(1/d)^5}$	57.7	52.0	45.7	41.8	53.6	41.4	43.0	26.3
$WKM_{(1/d)^7}$	57.2	51.6	44.8	40.7	53.4	41.3	42.7	26.2
Censoring rate								
Overall	29%	35%	42%	49%	26%	41%	37%	54%
group 0	32%	32%	45%	45%	32%	32%	45%	45%
group 1	26%	39%	39%	52%	19%	49%	29%	63%
case	1	2	3	4	5	6	7	8

investigate the power of the tests the sample size $N = 1600$ was not used and results are based on 1,000 replications. For each of the independent data sets, we compute the log-rank tests for the 'Fully Observed' (FO) analysis, treated as the gold standard, (KM estimates are derived for each simulated data set before any censoring is applied), for the partially observed (PO) analysis (with censoring), for the two approaches based on the KMIB method[11] and for the weighted log-rank test approach. The R code for the simulation study available upon request.

3.2 | Results

Table 1 provides the sizes of the log-rank test in a situation with dependent censoring for few parameters of the WKM approach (complete results are given in Table S1 supplemental material). The results indicate that for the PO analysis the sizes are comparable to the nominal level (5 per cent). The KMIB method, in a situation with both of the working models correctly specified, produces also sizes comparable to the 5 per cent level. For the results of the WKM approach displayed in Table 1, sizes are also comparable to the nominal level. However, globally sizes are slightly higher for $WKM_{U,2\%}$, $WKM_{U,80\%}$, $WKM_{\mathcal{N}(0.05)}$ and

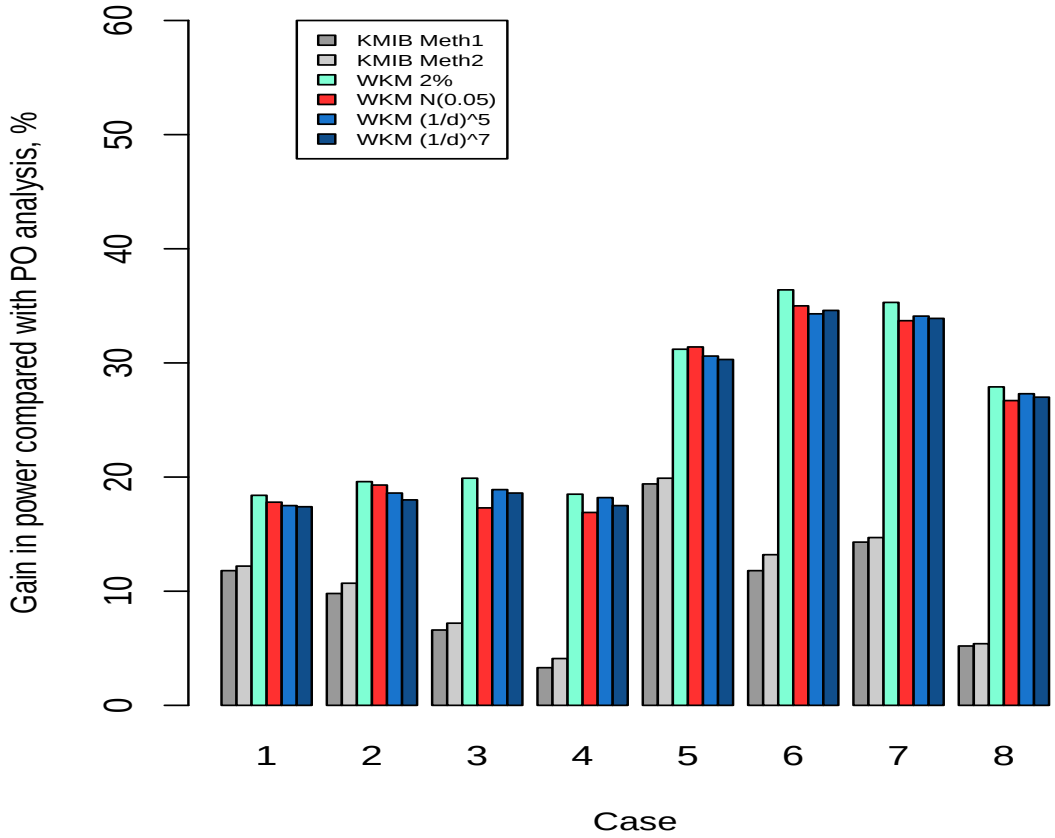

FIGURE 1 From Table 2 (N=200), gain in power when both working PH models are correctly specified for both KMIB and WKM methods compared with the PO analysis. The cases 1-8 correspond to the cases introduced in bottom of Table 2.

$\text{WKM}_{\mathcal{N}(8)}$ than that for $\text{WKM}_{(1/d)^5}$ and $\text{WKM}_{(1/d)^7}$. In a situation with only the working failure time model misspecified, the sizes of the KMIB method in general are slightly greater than when both models are correctly specified. The sizes for our WKM approach in general are almost similar than when both models are correctly specified with no clear trend. In a situation with only the working censoring time model misspecified, as previously the sizes of the KMIB method in general are greater than when both models are correctly specified. For $\text{WKM}_{\mathcal{N}(\sigma)}$, in general the sizes are greater than when both models are correctly specified and for $\text{WKM}_{U,x\%}$ sizes are mostly similar. For $\text{WKM}_{(1/d)^p}$ the sizes are slightly either lower or greater than the sizes when both models are correctly specified with no clear trend.

As expected, similar sizes to the conventional log rank test were found when $x = 100\%$ and $\sigma = 200$ (Table S1). For $\text{WKM}_{U,x\%}$, the sizes increase slightly from x=2 to 20% and then decrease from 20 to 100%, especially when the percent of censoring is 45% (Figure S1 supplemental material). Similar findings were found for $\text{WKM}_{\mathcal{N}(\sigma)}$ with increasing sizes from $\sigma = 0.05$ to 0.5 and then decreasing sizes from $\sigma = 0.5$ to 200 (Figure S1). The sizes for $\text{WKM}_{(1/d)^p}$ with p=1 and 3 are higher than for the FO analysis especially for p=1 with few values larger than 7% (Table S1).

Table 2 provides the powers of the log-rank tests in a situation with dependent censoring when N=200 (complete results in Table S2 supplemental material). The results indicate clearly that the power based on the PO analysis is much lower than that of the FO analysis. The loss in power is dependent of both the global rate of censoring and the difference in rates of censoring between the two groups. For example, with a difference in rates of censoring around 6 per cent between the two groups ($\alpha_1 = 0.15$), on the average the PO analysis provides a power about 26 per cent lower than the FO analysis (from 21% for an overall censoring rate of 29% to 30% for an overall censoring rate of 49%). With a higher difference in rates of censoring between the two groups, a difference of 16 per cent (α_1=0.75), on the average the loss in power is 45 per cent (from 34 per cent for an overall censoring rate of 26% to 50 per cent for an overall censoring rate of 54%).

TABLE 3 Monte Carlo results: power (per cent) analysis with dependent censoring. Misspecification implies using Z_1, Z_2 and Z_3 instead of Z_1, Z_2, Z_3, Z_4 and Z_5 in one of the working model (N=400). Results are based on 1,000 replications

	$\alpha_1 = 0.15$				$\alpha_1 = 0.75$			
	$\alpha_0 = -0.2$		$\alpha_0 = 0.4$		$\alpha_0 = -0.2$		$\alpha_0 = 0.4$	
	Treatment effect (ψ)							
Method	-0.75	0.75	-0.75	0.75	-0.75	0.75	-0.75	0.75
FO	90.9	91.5	90.9	90.3	91.5	89.3	89.8	89.5
PO	68.2	62.0	60.1	50.6	50.9	27.4	31.4	15.7
Both working PH models correctly specified								
KMIB meth1	79.3	76.5	70.1	56.6	77.9	47.2	57.2	26.0
KMIB meth2	79.6	77.1	70.8	58.1	78.3	47.7	58.2	27.1
$WKM_{U,2\%}$	89.3	88.8	87.8	81.9	91.1	81.4	83.3	69.1
$WKM_{\mathcal{N}(0.05)}$	88.6	86.8	86.2	78.4	90.5	78.1	80.6	65.3
$WKM_{(1/d)^5}$	87.9	86.8	85.3	78.2	89.7	78.8	80.3	66.3
$WKM_{(1/d)^7}$	87.7	86.4	84.9	77.7	89.8	77.9	79.6	65.2
Only working failure PH model misspecified								
KMIB meth1	78.3	73.5	68.0	54.5	74.4	42.6	52.6	22.5
KMIB meth2	78.3	74.3	68.9	55.9	74.6	43.4	53.7	23.9
$WKM_{U,2\%}$	87.7	85.9	84.1	75.8	88.9	73.3	77.1	54.4
$WKM_{\mathcal{N}(0.05)}$	86.7	85.1	80.6	71.3	88.3	68.4	76.1	49.5
$WKM_{(1/d)^5}$	86.3	83.6	81.1	72.0	87.2	67.8	74.2	50.9
$WKM_{(1/d)^7}$	85.9	83.1	80.8	71.4	87.2	67.2	74.1	50.8
Only working censoring PH model misspecified								
KMIB meth1	80.3	75.8	71.4	57.3	78.2	47.0	57.1	26.3
KMIB meth2	80.7	76.1	71.9	58.0	78.6	48.0	57.7	27.4
$WKM_{U,2\%}$	87.1	86.1	83.9	75.1	87.5	71.8	77.0	55.3
$WKM_{\mathcal{N}(0.05)}$	87.1	86.1	83.9	75.1	87.5	71.8	77.0	55.3
$WKM_{(1/d)^5}$	85.2	82.5	80.2	71.4	85.9	67.7	73.4	50.1
$WKM_{(1/d)^7}$	84.9	81.8	79.4	69.7	86.1	67.4	73.0	49.0
Censoring rate								
Overall	29%	35%	42%	49%	26%	41%	37%	54%
group 0	32%	32%	45%	45%	32%	32%	45%	45%
group 1	26%	39%	39%	52%	19%	49%	29%	63%
case	1	2	3	4	5	6	7	8

When both working PH models are correctly specified, Figure 1 displays the gain in power for both KMIB and WKM methods compared with the PO analysis. The KMIB method produces powers between 4 and 20 per cent higher than the PO analysis. In all situations the WKM method outperforms the KMIB approach. In particular, on the average $WKM_{(1/d)^p}$, with p=5 and 7, produces a power about 18 per cent higher than the PO analysis with a 6 per cent difference in censoring rates between the two groups ($\alpha_1 = 0.15$) and about 32 per cent with a 16 per cent difference rates ($\alpha_1 = 0.75$). On the average $WKM_{U,2\%}$ and $WKM_{\mathcal{N}(0.05)}$ produces a power at least 18 per cent higher than the PO analysis. For both the KMIB and WKM methods, the gain in power is lower when one of the working PH model is misspecified (Table S2 and Figures S3-S4). Both methods, however, outperform the PO analysis and the WKM approach still outperforms the KMIB approach. As previously, the power for $x = 100\%$ and $\sigma = 200$ is similar to the power of the conventional log rank test (Table S2). As expected, the gain in power for $WKM_{U,80\%}$ and $WKM_{\mathcal{N}(8)}$ compared with the PO analysis is close to zero. Overall, the power decreases markedly from 20 to 100% and from $\sigma = 0.5$ to 200 (figures S2-S10).

Table 3 provides results when N=400 (complete results in Table S3 supplemental material). On the average the PO analysis produces powers about 31 and 59 per cent lower than the FO analysis when the difference in rates of censoring between the

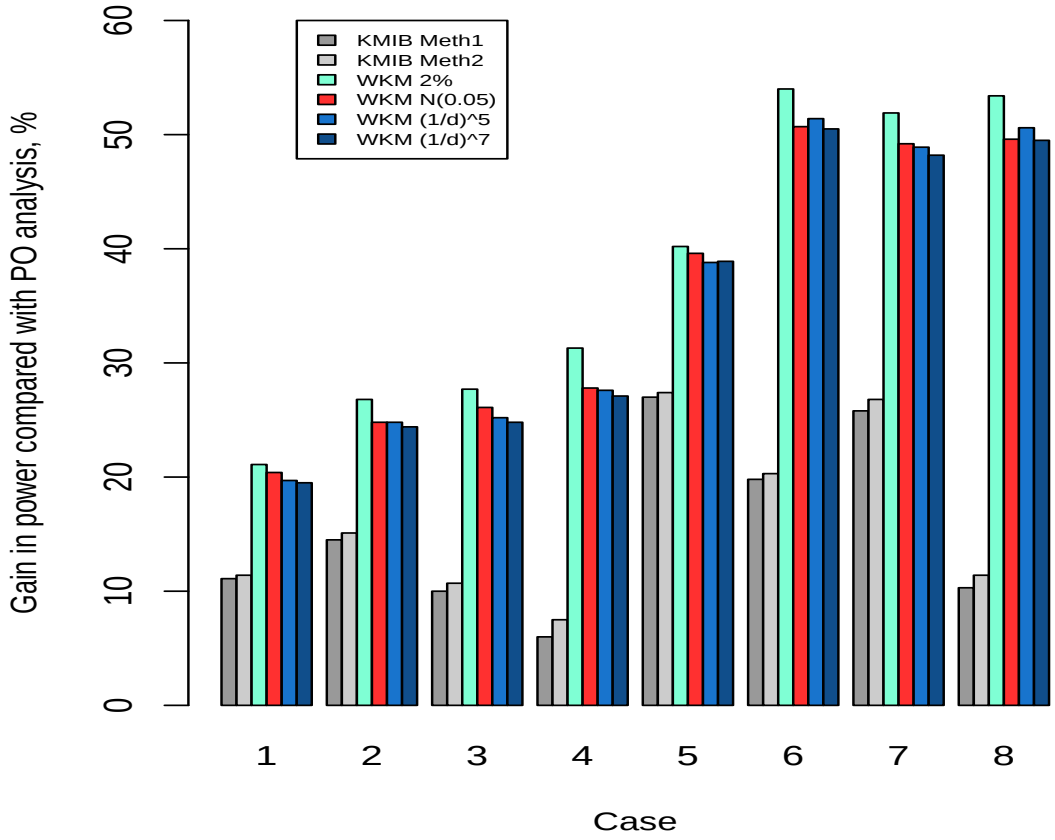

FIGURE 2 From Table 3 (N=400), gain in power when both working PH models are correctly specified for both KMIB and WKM methods compared with the PO analysis. The cases 1-8 correspond to the cases introduced in bottom of Table 3.

two groups is about 6 (α_1=0.15) and 16 (α_1=0.75) per cent, respectively. When both models are correctly specified, the gain in power for both KMIB and WKM methods compared with the PO analysis is greater than with N=200 (Figure 2). With a difference in rates of censoring between the two groups around 16 per cent, on the average the gain in power of the WKM approach is 47 per cent. Again, in all situations the WKM method outperforms the KMIB approach. For both KMIB and WKM methods, as previously, the gain in power is lower when one of the working PH model is misspecified (Table S3). Similar trends are observed for N=800 (Table S4) though gain in power are greatly reduced when the difference in rates of censoring between the two groups is about 6 per cent. Of note, in many situations the WKM method provides a power similar to the FO analysis

In summary, the PO analysis and the KMIB method produce sizes comparable to the nominal level even, for the KMIB approach, when one of the working PH model is misspecified. For the WKM approach, $\text{WKM}_{(1/d)^p}$ with $p = 5$ and 7 produces sizes comparable to the nominal level in all sampe sizes investigated providing, for our method, the best finite sample property. The WKM method provides acceptable sizes for $\text{WKM}_{U,2\%}$ and $\sigma = 0.05$ and 0.15 for $\text{WKM}_{\mathcal{N}(\sigma)}$ whereas slightly higher than with $\text{WKM}_{(1/d)^p}$, $p = 5$ and 7. As expected, the PO analysis provides a much lower power than the FO analysis. The KMIB procedure consistently produces a power higher than the PO analysis but smaller than the FO analysis. $\text{WKM}_{(1/d)^p}$ with $p = 5$ and 7, provides a power much higher than the PO analysis and higher than the KMIB method. With a large sample size (N=800), our method provides a power almost similar to that of the FO analysis. When one of the working PH model is misspecified, the WKM method still outperforms both the PO analysis and KMIB approach.

TABLE 4 . Data analysis of the GBSG data set (N=686) using grade (tumor grade) nodes (number of positive lymph nodes) and pgr (progesterone receptors (fmol/l)) as covariates in both failure and censoring models

	Failure time model			Censoring time model		
Covariates	Estimate	SE	p-value	Estimate	SE	p-value
grade	0.387	0.182	0.03	0.273	0.172	0.09
nodes	0.032	0.012	0.01	0.037	0.016	0.02
pgr	-0.002	0.001	0.02	0.001	3.10^{-4}	0.07

TABLE 5 Data analysis of the GBSG data set (N=686) using grade (tumor grade) nodes (number of positive lymph nodes) and pgr (progesterone receptors (fmol/l)) as covariates in both working PH models for both KMIB and WKM methods.

Method	*p*-value	Method	*p*-value	Method	*p*-value
PO analysis	0.091	KMIB meth1	0.104	$WKM_{\mathcal{N}(0.05)}$	0.139
		KMIB meth2	0.102	$WKM_{\mathcal{N}(0.10)}$	0.026
		$WKM_{U,2\%}$	0.040	$WKM_{(1/d)^5}$	0.041
		$WKM_{U,5\%}$	0.042	$WKM_{(1/d)^7}$	0.040

4 | ILLUSTRATION OF THE METHODS ON BREAST CANCER PATIENTS

Our illustrative example is based on German Breast Study Group (GBSG) data set. The GBSG data set contains patient records from a 1984-1989 trial conducted by the German Breast Cancer Study Group (GBSG) of 720 patients with node positive breast cancer; it retains the 686 patients with complete data for the prognostic variables. The study has previously been used in methodological investigation [13,14] and can be found in R survival package (data(gbsg, package="survival")). Overall, 246 (36%) patients had received hormonal therapy. The outcome of interest is recurrence or death and was observed in 299 (44%) patients; 94 in the hormonal therapy group and 205 in the no treatment group. Considering the 686 patients, patients in the hormonal therapy group shows a lower risk of recurrence or death (Hazard Ratio = 0.695; 95%CI 0.54 to 0.89) with a *p*-value for the log-rank test of 0.0036.

We randomly select 191 patients from the GBSG data set. In this sample of 191 patients, the hazard ratio is similar than previously but the log-rank test is not statistically significant (p=0.093). We then analyze this sample with the statistical methods described in this work. Among the 7 potential prognostic factors we used the variables grade (tumor grade) nodes (number of positive lymph nodes) and pgr (progesterone receptors (fmol/l)) as covariates in the two working PH models. It is well known that the independence assumption between censoring and survival cannot be tested with right-censored data unless additional data are collected or assumptions are made about an underlying statistical model. [15] However, Hsu and Taylor suggest that the presence of the same significant variables in two working models, one for the failure time and one for the censoring time, does indicate the potential for dependent censoring. [3] The results for estimation of those two working models are provided in Table 4. For both models the three variables are either significant or close to significance.

Table 5 displays the *p*-values of the different statistics of the methods used in this work. The WKM approach provided smaller *p*-values than the classical log-rang test based on the PO analysis except for $WKM_{\mathcal{N}(0.05)}$. The difference in the *p*-values between $\sigma = 0.05$ and $\sigma = 0.10$ illustrates the large variability we have observed in few examples of using $WKM_{\mathcal{N}(\sigma)}$ with $\sigma < 0.10$. For this reason, we recommend using $WKM_{(1/d)^p}$ with p=5 or 7. The KMIB approach provided slightly higher *p*-values then the conventional log-rank test. Of note, the KMIB method uses a random number generator to impute data sets and then slightly different p-values may be found. Survival estimates for the methods are given in Figure 3.

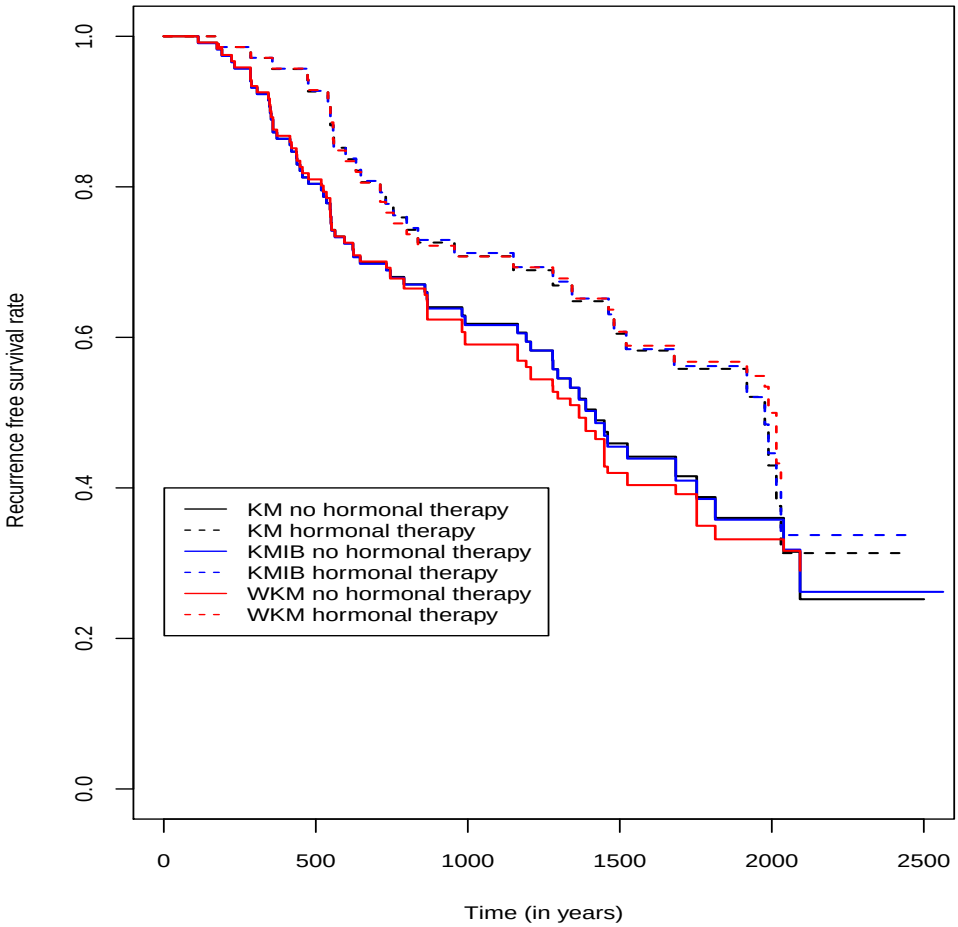

FIGURE 3 Survival estimates of two hormonal therapy groups in GBSG data set, WKM corresponds to $\text{WKM}_{(1/d)^5}$.

5 | DISCUSSION

This work introduces a weighted log-rank statistic to test for the difference between two survival curves in the presence of dependent censoring. The non-parametric test is based on WKM estimators that have been recently proposed.[4] The WKM approach used information from prognostic variables via two working PH models. These models are used to identify a neighborhood of similar observations that would received weights of censored individuals. The simulation study shows that, with appropriate parameters, our approach leads to a valid log-rank test. In particular, $\text{WKM}_{(1/d)^p}$ with $p = 5$ and 7 provides sizes comparable to the nominal value even if one of the working model is misspecified. In addition, it produces a power much higher than the power produced by analyzing the observed data without using the auxiliary variables (PO analysis) and also higher than the KMIB approach. We then recommend to use $\text{WKM}_{(1/d)^p}$ with $p = 5$ and 7 though $\text{WKM}_{U,\%}$ $\text{WKM}_{\mathcal{N}(\sigma)}$ with appropriate value of x and σ may produced promising results as well. As expected, our numerical results show that $\text{WKM}_{U,100\%}$ provides similar sizes and powers than the conventional log rank test (similar findings are found with $\sigma = 200$ for $\text{WKM}_{\mathcal{N}(\sigma)}$). Our simulation results show that the weight of censored observations should be redistributed among 'similar' individuals as measured by the distance defined in our approach. Increasing the number of neighbours, as well as increasing σ or decreasing p, markedly decrease the power the WKM test in redistributing the weight among observations having larger distances than censored observations.

In our previous work, we have shown that to estimate the survival function, the KMIB method outperformed both the PO analysis and the Inverse of Probability Censoring Weighted (IPCW) method.[4] In particular the IPCW method performed poorly in terms of relative bias and coverage probability. In addition, Hsu and Taylor showed that the size of the log-rank test based on

the IPCW method is well above the nominal level.[11] For these reasons we did not include the IPCW method in our simulation study.

In the simulation study we considered a situation with time-independent variables, which were known at baseline and are directly incorporated into the working failure PH model. As for the KMIB procedure, time-dependent variables can be considered for both working PH models.[11] Then the two working models needs to be fitted at every censored observation to the data of those at risk at the censoring time using the currently available auxiliary variables as fixed covariates. An attractive feature of our method is that the reliance on some specific parametric models is weak because the working failure PH model is only used to identify the neighborhood of similar observations at the time of censored observations. Then, the performance of our approach is not highly dependent on the assumptions of both working models. Such an attractive aspect is also share by the KMIB method where the neighborhood is used to develop an imputation method.[11]

For both KMIB and WKM methods, other misspecification could be investigated. In particular misspecification of the link function could be studied via a simulation study. Such a misspecification implies to modify the InformativeCensoring package as no option allowed to modify the link function. It is difficult to provide empirical result of such misspecification but we suspect that both methods would be disturbed in the same magnitude. This point, however, needs further study. For both methods, one has to choose the variables included in models used to derive the estimators. Both methods, however, have shown good performance even when one working model is misspecified. In addition, for the KMIB method we have to choose weights for failure and censoring risks as well as the number of imputed datasets and the size of the imputing risk set. Here we used the parameters recommended by their authors, NN=5, M=10, $w_f = 0.8$, and $w_c = 0.2$.[3,11] The choice of the number of neighbours, the value of σ and the value of p is a limitation of our method. Simulation results of our previous work show that this choice is important to provide reasonable estimates of $S(t)$ in the presence of dependent censoring.[4] In this work, we propose a new procedure for the redistribution of the weights $\text{WKM}_{(1/d)^p}$ given more weights on observations having smaller distances with larger value of p. Our simulation study and the example show that the log-rank test based on $\text{WKM}_{(1/d)^p}$ with $p = 5$ and 7 provide the most promising results and should be used in the situation of dependent censoring.

In our previous work, we have shown that using the distance based on the risk score for the failure time model provides, in general, similar results than using a PCA step.[4] Using the distance $d(j,k)$ defined in section 2.3.2 provides similar results of size and power than using a PCA step in the WKM approach (data not shown). Then the difference of power between the KMIB and WKM approaches did not resulted from the addition of a PCA step in our method. In the KMIB approach, imputation of the j^{th} censored observation is done in computing the KM estimator for the risk set $R(j^+, \text{NN})$ as described in section 2.3.2 (see[2] and[11] for more details). They then sample $U \sim [0,1]$ and take the time at which the KM estimator equals u as the imputed event time. Thus, the procedure imputes observed failure times unless the longest time in the imputing risk set is censored, in which case some imputed times may include this censored time. With NN=5 and a large rate of censoring, we suspect that many imputed times will be censored times. In the WKM approach, the weight of a censored observation at time t is likely redistributed among censored and uncensored observations at risk at t. But the weights of these censored observations receiving an extra weight is redistributed beyond time t among again censored and uncensored observations and so on. We suspect that the reason why the WKM approach outperforms the KMIB approach in term of power. Of note, a risk set imputation (RSI) procedure was introduced that will frequently impute censored time but was not investigated further to compare two survival functions.[2,11]

6 | ACKNOWLEDGMENTS

The author is very grateful to the referees and the editors for insightful comments on the article.

7 | DATA AVAILABILITY STATEMENT

The data used in the illustration can be uploaded in R survival package (data(gbsg, package="survival")). We used a sample (data358) of the original dataset that can be obtained as follow

```
N.total <- nrow(gbsg)
set.seed(358)
gbsguni < −runif(N.total, min = 0, max = 1)
data358 <- subset(gbsg.ana, uni<0.3)
```

References

1. Kalbfleisch J, Prentice R. *The Statistical analysis of Failure Time Data*. Wiley; New York . 2002.
2. Hsu CH, Taylor JM, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine* 2006; 25(20): 3503-17.
3. Hsu CH, Taylor JM. A robust weighted Kaplan-Meier approach for data with dependent censoring using linear combinations of prognostic covariates. *Statistics in Medicine* 2010; 29(21): 2215-23.
4. Ahmed I, Flandre P. Weighted Kaplan-Meier estimators motivating to estimate HIV-1 RNA reduction censored by a limit of detection. *Statistics in Medicine* 2020; 39(2): 968-83.
5. Malani H. A modification of the redistribution to the right algorithm using diseases markers. *Biometrika* 1995; 90: 577-584.
6. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; 52(1): 137-51.
7. 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(3): 779-88.
8. Robins JM, Rotnitzky A. *Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers*: 24-33; Boston: Birkhauser. . 1992.
9. Jing Xu MHC, Brodaty H. Propotional hazard model estimation under dependent censoring using copulas and penalized likelohood.. *Statistics in Medicine* 2017(37): 2238-2251.
10. Efron B. The two sample problem with censored data.. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 5 of Berkeley, pages 831-853. ; 1967.
11. Hsu CH, Taylor JM. Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Statistics in Medicine* 2009; 28(21): 462-75.
12. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 2005; 24: 3089-3110.
13. Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies.. *Statistics in Medicine* 1996; 15: 263-71.
14. Royston P, Altman D. External validation of a Cox prognostic model: principles and methods.. *BMC Medical Research Methodology* 2013: 13-33.
15. Tsiatis AA. A nonidentifiability aspect of the problem of competing risks.. *Proceedings of the National Academy of Sciences, U.S.A.* 1975; 72: 20-22.