Published in final edited form as:

Stat Med. 2011 August 30; 30(19): . doi:10.1002/sim.4295.

Some insight on censored cost estimators

H. Zhao^{a,*†}, Y. Cheng^b, and H. Bang^c

^aDepartment of Epidemiology and Biostatistics, School of Rural Public Health, Texas A&M Health Science Center, College Station, TX 77843, U.S.A

^bDepartment of Statistics, Texas A&M University, College Station, TX 77843, U.S.A

^cDivision of Biostatistics and Epidemiology, Department of Public Health, Weill Medical College of Cornell University, New York, NY 10065, U.S.A

Abstract

Censored survival data analysis has been studied for many years. Yet, the analysis of censored mark variables, such as medical cost, quality-adjusted lifetime, and repeated events, faces a unique challenge that makes standard survival analysis techniques invalid. Because of the ‘informative’ censorship imbedded in censored mark variables, the use of the Kaplan–Meier (*Journal of the American Statistical Association* 1958; 53:457–481) estimator, as an example, will produce biased estimates. Innovative estimators have been developed in the past decade in order to handle this issue. Even though consistent estimators have been proposed, the formulations and interpretations of some estimators are less intuitive to practitioners. On the other hand, more intuitive estimators have been proposed, but their mathematical properties have not been established. In this paper, we prove the analytic identity between some estimators (a statistically motivated estimator and an intuitive estimator) for censored cost data. Efron (1967) made similar investigation for censored survival data (between the Kaplan–Meier estimator and the redistribute-to-the-right algorithm). Therefore, we view our study as an extension of Efron’s work to informatively censored data so that our findings could be applied to other marked variables.

Keywords

inverse-probability weighting; marked process; medical cost; redistribute-to-the-right; replace-from-the-right; survival analysis

1. Introduction

In many applications in prospective studies, problems can be formulated in the marked point process framework as variables of interest are marks of the endpoint, which are accumulated over time but not observed when the survival time is censored [1]. Medical costs, quality-adjusted survival times, recurrent events, and customer lifetime value are some examples. A common feature of these data is that they are functions of a time process; therefore, they are subject to right censoring at a certain time, rendering impossible to observe the whole process for each subject. A methodological challenge in the analysis of these data is the so-called induced informative censoring, which was first formalized by Lin et al. [2]. Although it is often reasonable to assume that the censoring time and the survival time are

independent, which justifies the use of the Kaplan–Meier [3] estimator for survival analysis, the same ‘independent censoring’ assumption is not generally true for the corresponding marked variables—for example, a healthier subject may accumulate costs slowly, hence having less costs at the time of censoring as well as at the time of an event [4]. Because of this, the standard statistical techniques that have been developed for survival analysis, such as the Kaplan–Meier estimator, the log-rank test, and the proportional hazards model, are not well suited for the analysis of the censored mark variables.

Several estimators have been proposed over the past decade to handle this unique problem. For the estimation of mean medical cost, Lin et al. [2] proposed estimators that intended to reduce the bias of the naive estimators (e.g., sample average of all costs or uncensored costs) via survival probability-weighted estimators with partitioned time intervals. Bang and Tsiatis [5] proposed consistent estimators using the inverse probability weighting technique. Zhao and Tian [6] proposed a more efficient estimator, among others. Although these estimators look ostensibly different, being formulated differently, Zhao et al. [7] discovered some special situations in which different estimators become identical.

Although large sample properties of the aforementioned estimators have been established, it may be still difficult for applied researchers to understand why ordinary methods such as the Kaplan–Meier estimator do not work and why novel methods are needed. Actually, a similar struggle had occurred 40 years ago when researchers tried to understand the Kaplan–Meier estimator better. Particularly, Efron [8] provided an alternative way to construct the Kaplan–Meier estimator using the redistribute-to-the-right algorithm. This effort offered a more intuitive way to understand the underlying mechanism of the Kaplan–Meier estimator, one of the most popularly used statistical methods in history but not necessarily easy to understand [9, 10].

Recently, Pfeifer and Bang [11] proposed the replace-from-the-right (RR) algorithm, a modified version of Efron’s original algorithm, for estimating the mean cost with censored data and reported numerical equivalency between this intuitive, user-friendly estimator and an inverse probability-weighted estimator proposed by Bang and Tsiatis [5], the first consistent estimator in the literature. However, they did not provide mathematical justifications for their observation, and we are not yet sure when this equivalency is true or if it is always true.

In this paper, we will look into these two estimators more closely and provide a mathematical proof that these two estimators are indeed identical. First, we will revisit the link between the Kaplan–Meier estimator and the redistribute-to-the-right algorithm for survival data that was already documented in the literature. Then, we will review the definitions of the two medical cost estimators—one from the inverse probability weighting and the other from the RR algorithm—and provide a proof of the equivalency of these two estimators. Finally, we will discuss our findings and their implications on related problems.

2. The Kaplan–Meier estimator and the redistribute-to-the-right algorithm

We assume a standard survival data setting in which the survival time to an event (death, for example) is the variable of interest and in which it is subject to right censoring. The i th subject has survival time T_i and censoring time C_i ($i = 1, \dots, n$). What we normally observe for each subject in reality is that whichever time occurs first, $X_i = \min(T_i, C_i)$, along with the event indicator, $\Delta_i = I(T_i < C_i)$. Here, we call X observation or follow-up time.

To estimate the survival function of T , $S(t) = \Pr(T > t)$, we can construct the well-known Kaplan–Meier estimator as follows. Allowing for possible ties, we assume that the observed events occur at D distinct times $\tau_1 < \tau_2 < \dots < \tau_D$. Denote d_i as the total number of deaths

observed at time τ_i , and R_i as the total number of subjects still at risk at time τ_i ; that is,
 $R_i = \sum_{j=1}^n I(X_j \geq \tau_i)$. The Kaplan–Meier estimator $\hat{S}(t)$ can be defined by

$$\hat{S}(t) = \prod_{i: \tau_i \leq t} \left(1 - \frac{d_i}{R_i}\right). \quad (1)$$

Efron [8] proposed a redistribution-to-the-right construction, which illustrates how the censored observations contribute to the Kaplan–Meier estimator. First, arrange the observational times X_i 's ($i=1, \dots, n$) from the smallest to the largest and from left to right, such that $X_1 \leq X_2 \leq \dots \leq X_n$. If two or more X_i values are equal, place those with $\Delta = 1$ before those with $\Delta = 0$. Each observation has the same weight $1/n$. Starting from the smallest time and moving to the right, whenever we reach a censored time (which can be tied), we distribute its mass (or their masses) uniformly to all times to the right. We repeat this process until we reach the largest censoring time. We obtain the survival probability estimate at time t by adding up all the weights at times greater than t . Here is an example with observed data $[X = \{1, 2, 3, 4, 5\}, \Delta = \{1, 0, 1, 0, 1\}]$.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
Step 1:	$\frac{1}{5}$		$\frac{4}{15} (= \frac{1}{5} + \frac{1}{3} \cdot \frac{1}{5})$	$\frac{4}{15}$	$\frac{4}{15}$
Step 2:	$\frac{1}{5}$		$\frac{4}{15}$		$\frac{8}{15} (= \frac{4}{15} + \frac{4}{15})$
$\hat{S}(t) =$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{8}{15}$	$\frac{8}{15}$	0

In Step 0, all the observation times (X_1 to X_5) get the same weight $1/5$. In Step 1, we find the smallest censoring time at $X_i = 2$ and distribute its weight of $1/5$ to all three observations to its right, so that each of them gets an additional weight of $1/15$, making the total weight at these times become $4/15$. Moving to the next largest censoring time, $X_i = 4$, we distribute its weight of $4/15$ to one observation to its right, making the mass at time $X_i = 5$ become $8/15$. We can then obtain the survival function estimate at any time t by summing up all the masses at times greater than t . This is exactly the same as the Kaplan–Meier estimate for survival probability at a given time point.

Later, Dinse [12] modified the redistribute-to-the-right algorithm so that the computation could be a little easier. Now starting from the largest censoring time and moving to the left, we re-distribute the mass at each censoring time to all the event times larger than itself, in proportion to the mass at each event time, as shown in the following graph.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
Step 1:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$		$\frac{2}{5} (= \frac{1}{5} + \frac{1}{5})$
Step 2:	$\frac{1}{5}$		$\frac{4}{15} (= \frac{1}{5} + \frac{1}{3} \times \frac{1}{5})$		$\frac{8}{15} (= \frac{2}{5} + \frac{2}{3} \times \frac{1}{5})$
$\hat{S}(t) =$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{8}{15}$	$\frac{8}{15}$	0

In Step 0, all the observation times (X_1 to X_5) get the same weight $1/5$. In Step 1, we locate the largest censoring time, X_4 in this example, and redistribute its weight to the right, resulting in a weight of $2/5$ for observation 5. In Step 2, we move to the next largest censoring time, X_2 in this case, and redistribute its weight of $1/5$ to the uncensored times to its right (X_3 and X_5), in proportion to their current weights ($1/5$ and $2/5$). Hence, X_3 gets an extra weight of $1/3 \times 1/5 = 1/15$, making its final weight become $4/15$. Similarly, X_5 gets an additional weight of $2/3 \times 1/5 = 2/15$, making its final weight become $2/5 + 2/15 = 8/15$. The survival estimator at each time is equal to the total weights to its right, which is the same as the Kaplan–Meier survival estimator.

The Efron's redistribute-to-the-right algorithm clearly demonstrates that the Kaplan–Meier estimator for the survival function is equivalent to redistributing the weights from the censoring times equally among all the times that are larger than themselves. Dinse's algorithm demonstrates further that the Kaplan–Meier estimator is also equivalent to redistributing the weights at censoring times to all the larger uncensored times in a mass-weighted manner. The Kaplan–Meier estimator is shown to be the most efficient and consistent non-parametric estimator when censoring and survival times are independent [13].

3. Medical cost estimators

If we have complete cost data for each person (i.e., no censoring), it is straightforward to estimate the mean cost. However, when the cost data for some patients are censored, it is challenging to do the same task. Censoring occurs typically in a clinical trial where cost information may be collected only when subjects stay in the study. Because of the presence of censoring, the marginal distribution of lifetime cost may be nowhere identifiable non-parametrically, so it is not possible to estimate the mean lifetime costs, say, without making some parametric assumptions about the cost data [14]. Therefore, we consider the mean costs accumulated over a time limit L , where L is usually chosen to be a time at which a reasonable number of subjects are still being observed. As a consequence, people who are still alive at time L are considered to have complete information on survival and costs (as if they had a failure event at time L).

A naive use of the Kaplan–Meier estimator in this situation could be to treat the censored cost and uncensored cost as the equivalent of censored time and uncensored time. We may apply the Kaplan–Meier method to obtain a survival function for costs, and then compute the area under the curve. However, this estimator has been shown to be biased due to induced informative censoring because the censored cost is usually positively correlated with the uncensored cost even when the corresponding times are independent [2]. Therefore, the direct application of the redistribute-to-the-right algorithm to cost data will not be appropriate.

3.1. The Bang and Tsiatis estimator

A consistent estimator for the mean cost was proposed by Bang and Tsiatis [5] based on the inverse probability weighting technique [15]:

$$\mu^{BT} = \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i M_i}{\widehat{K}(T_i)}, \quad (2)$$

where M_i is the total cost observed for subject i ($i = 1, \dots, n$) and $\widehat{K}(T_i)$ is the Kaplan–Meier estimator for the survival function for the censoring time, $K(t) = \Pr(C_i > t)$, evaluated at time T_i . The underlying idea is that each complete cost data represents a total of $1/\widehat{K}(T_i)$ observations that might have been observed. Note that when $\Delta_i = 1$, $X_i = T_i$; that is, the observed cost M_i is equal to the true cost.

3.2. The replace-from-the-right estimator

Although the Bang and Tsiatis (BT) estimator is not difficult to calculate, it may not be so intuitive to practitioners why this method produces a consistent estimator and how censored data contribute to estimation. Pfeifer and Bang [11] proposed an estimator using the RR algorithm, a modified version of Efron's redistribute-to-the-right algorithm. This method shows how uncensored cost and censored cost data are used in the estimation process.

The RR method can be summarized as follows. First, arrange all the cost data from the smallest time to the largest time—note that data (M_i, X_i, Δ_i) are sorted by X_i , not by M_i . We assume that the data with the largest observation time is uncensored, which can be satisfied by considering the restricted mean cost, as described earlier. Starting from the largest observational time, we move to the left to find the subject with the largest censoring time. Because we do not know this subject's true total cost, we could use a sensible guess, that is, the average of the costs from subjects who have larger survival times (i.e., all subjects on the right side). We do this one by one—use the observed cost for an uncensored subject, and use the replaced cost (i.e., the downstream mean) for a censored subject. We continue doing this until we reach the smallest censoring time. Here, let us denote the replaced cost at censoring times by M^{RR} , then the RR estimator can be written as follows:

$$\mu^{RR} = \frac{1}{n} \sum_{i=1}^n \{ \Delta_i M_i + (1 - \Delta_i) M_i^{RR} \}. \quad (3)$$

The succeeding graph is a simple example to illustrate how this algorithm works. Suppose we have the following data, which are the same as the previous survival data but with cost data added: $[X = \{1, 2, 3, 4, 5\}, \Delta = \{1, 0, 1, 0, 1\}, M = \{10, 50, 100, 60, 40\}]$. There are two censored data, the second and fourth subjects. In Step 1, we deal with the fourth subject ($i = 4$) and replace its observed cost by the cost for the fifth subject, who is the sole subject who survived longer than the fourth subject. In Step 2, we replace the cost for the second subject by the average of 100 (true), 40 (replaced), and 40 (true), which is equal to 60. Therefore, the mean cost from the RR method gives an estimate of 50, as shown in the following graph.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
$M_i =$	10	50	100	60	40
Step 1:				40 (from cost at X_5)	
Step 2:		60 (average of costs at X_3, X_4 and X_5)			
$\mu^{RR} =$	$(10 + 60 + 100 + 40 + 40)/5 = 50.$				

If we apply the BT estimator to the same data, we get

$$\mu^{BT} = \frac{1}{5} \sum_{i=1}^5 \frac{\Delta_i M_i}{\widehat{K}(T_i)} = \frac{1}{5} \left(\frac{10}{1} + \frac{100}{3/4} + \frac{40}{3/8} \right) = \frac{1}{5} (10 + 400/3 + 320/3) = 50,$$

where the Kaplan–Meier estimates for $K(t) = \Pr(C_i > t)$ are computed as $\widehat{K}(\hat{X}_i) = (1, 3/4, 3/4, 3/8, 3/8)$ at $X_i = \{1, 2, 3, 4, 5\}$, respectively.

Although Pfeifer and Bang [11] noticed the numerical equivalency of the RR estimator and the BT estimator, they did not provide a mathematical proof of this equivalency or condition(s) under which they are equivalent. In the next section, we prove that these two estimators are mathematically identical.

3.3. Equivalency of the Bang and Tsiatis estimator and the replace-from-the-right estimator

Suppose we have data (X_i, Δ_i, M_i) , $i = 1, \dots, n$ for cost estimation. Arrange the data from the smallest X_i to the largest X_i from left to right. Denote the ordered distinctive censoring times as t_j ($j = 1, \dots, J$). Let Y_j represent the number of people who have observation times greater than t_j ; that is, $Y_j = \sum_{i=1}^n I(X_i > t_j)$. Let n_j represent the number of people who are censored

at time t_j . Here, if an event time T_i occurs at a censoring time t_j , we assume that T_i happened shortly before t_j .

Starting from the largest censoring time t_J , there are Y_J subjects who have complete costs and whose survival times are greater than t_J . Hence, the RR cost at t_J for each censored observation is a simple average of all the costs from Y_J subjects on the right side; that is,

$$M_J^{RR} = \frac{1}{Y_J} \sum_{i: X_i > t_J} \Delta_i M_i.$$

Moving to the second largest censoring time, t_{J-1} , where the number of subjects who survived longer than that is Y_{J-1} , the RR cost for each observation censored at this time is

$$\begin{aligned} M_{J-1}^{RR} &= \frac{1}{Y_{J-1}} \sum_{i: X_i > t_{J-1}} \{(1-\Delta_i)M_i^{RR} + \Delta_i M_i\} \\ &= \frac{1}{Y_{J-1}} \left(n_J M_J^{RR} + \sum_{i: X_i > t_{J-1}} \Delta_i M_i \right) \\ &= \frac{1}{Y_{J-1}} \left(\frac{n_J}{Y_J} \sum_{i: X_i > t_J} \Delta_i M_i + \sum_{i: X_i > t_{J-1}} \Delta_i M_i \right) \\ &= \frac{1}{Y_{J-1}} \sum_{i: t_{J-1} < X_i \leq t_J} \Delta_i M_i + \frac{1}{Y_{J-1}} \left(1 + \frac{n_J}{Y_J} \right) \sum_{i: X_i > t_J} \Delta_i M_i. \end{aligned}$$

Similarly, we also have

$$\begin{aligned} M_{J-2}^{RR} &= \frac{1}{Y_{J-2}} \sum_{i: X_i > t_{J-2}} \{(1-\Delta_i)M_i^{RR} + \Delta_i M_i\} \\ &= \frac{1}{Y_{J-2}} \left(n_{J-1} M_{J-1}^{RR} + n_J M_J^{RR} + \sum_{i: X_i > t_{J-2}} \Delta_i M_i \right) \\ &= \frac{1}{Y_{J-2}} \sum_{i: t_{J-2} < X_i \leq t_{J-1}} \Delta_i M_i + \frac{1}{Y_{J-2}} \left(1 + \frac{n_{J-1}}{Y_{J-1}} \right) \sum_{i: t_{J-1} < X_i \leq t_J} \Delta_i M_i + \frac{1}{Y_{J-2}} \left(1 + \frac{n_{J-1}}{Y_{J-1}} \right) \left(1 + \frac{n_J}{Y_J} \right) \sum_{i: X_i > t_J} \Delta_i M_i. \end{aligned}$$

It is clear that for a complete observation, for example, one with $T_i > t_J$, the contribution to all the RR costs at the first censored time smaller than it is

$$\frac{n_J}{Y_J};$$

its contribution to the second censored time to its left is

$$\frac{n_{J-1}}{Y_{J-1}} \left(1 + \frac{n_J}{Y_J} \right);$$

its contribution to the third censored time to its left is

$$\frac{n_{J-2}}{Y_{J-2}} \left(1 + \frac{n_{J-1}}{Y_{J-1}}\right) \left(1 + \frac{n_J}{Y_J}\right);$$

and so on.

To generalize this, consider any complete observation M_i with death time T_i . Denote the ordered distinctive censoring times smaller than this observation as $t_1, \dots, t_{J(i)}$. Then we can show that the following are true:

the contribution of M_i to $t_{J(i)}$ is $\frac{n_{J(i)}}{Y_{J(i)}}$;

the contribution of M_i to $t_{J(i)-1}$ is $\left(1 + \frac{n_{J(i)}}{Y_{J(i)}}\right) \cdot \frac{n_{J(i)-1}}{Y_{J(i)-1}}$;

...

the contribution of M_i to t_1 is $\left(1 + \frac{n_{J(i)}}{Y_{J(i)}}\right) \cdots \left(1 + \frac{n_2}{Y_2}\right) \cdot \frac{n_1}{Y_1}$.

Hence, the total contribution of M_i to the RR estimator is a summation of its own contribution of 1, plus its contribution to all the censoring times before itself; that is,

$$\begin{aligned} 1 + \left(\frac{n_{J(i)}}{Y_{J(i)}}\right) + \left(1 + \frac{n_{J(i)}}{Y_{J(i)}}\right) \cdot \frac{n_{J(i)-1}}{Y_{J(i)-1}} + \cdots + \left(1 + \frac{n_{J(i)}}{Y_{J(i)}}\right) \cdots \left(1 + \frac{n_2}{Y_2}\right) \cdot \frac{n_1}{Y_1} &= \left(1 + \frac{n_{J(i)}}{Y_{J(i)}}\right) \cdot \left\{1 + \frac{n_{J(i)-1}}{Y_{J(i)-1}} + \cdots + \left(1 + \frac{n_{J(i)-1}}{Y_{J(i)-1}}\right) \cdots \left(1 + \frac{n_2}{Y_2}\right)\right\} \\ &= \left(1 + \frac{n_{J(i)}}{Y_{J(i)}}\right) \cdot \left(1 + \frac{n_{J(i)-1}}{Y_{J(i)-1}}\right) \cdots \left(1 + \frac{n_2}{Y_2}\right) \cdot \left(1 + \frac{n_1}{Y_1}\right) \\ &= \prod_{j=1}^{J(i)} \left(1 + \frac{n_j}{Y_j}\right). \end{aligned}$$

Using the formula for the Kaplan–Meier estimator (1) and noticing that we need the number of observations at risk at each censoring time for the survival function of the censoring time $K(t)$, here $R_j = Y_j + n_j$, we have

$$\widehat{K}(T_i) = \prod_{j: t_j \leq T_i} \left(1 - \frac{n_j}{Y_j + n_j}\right) = \prod_{j=1}^{J(i)} \frac{Y_j}{Y_j + n_j}.$$

Then the contribution of a complete observation M_i to the BT estimator (2) becomes

$$\frac{1}{\widehat{K}(T_i)} = \frac{1}{\prod_{j=1}^{J(i)} \frac{Y_j}{Y_j + n_j}} \prod_{j=1}^{J(i)} \left(1 + \frac{n_j}{Y_j}\right).$$

This is the same as the contribution of M_i to the RR estimator. Therefore, we have shown that the two formulas (2) and (3) are identical.

Note that if we replace the total observed cost M_i at each time by the observation time X_i , and perform the replacement-from-the-right algorithm, we will get the mean survival time, which is mathematically the same as the area under the Kaplan–Meier survival curve, assuming that the largest observation is not censored [16].

4. Discussion

In this paper, we provide a proof that links two mean cost estimators, the BT estimator which is based on the inverse probability weighting technique and the RR estimator which is based on the RR algorithm. Although these two estimators look substantially different in that one is far more intuitive and the other is motivated purely by statistical theory, we learned that these two estimators are mathematically equivalent. Hence, we provide a theoretical justification for the use of the intuitive estimator and also provide some insight on a more sophisticated inverse probability-weighted estimator. Efron [8] and Dinse [12] reported a similar relationship for survival data, which led us to understand the mechanism of the Kaplan–Meier estimator in a more intuitive way. In that regard, our work can be considered as an extension of their works for censored survival data to censored cost data.

Bang and Tsiatis [5] showed that the estimator for the survival distribution of medical cost, $S_M(x) = \Pr(M_i > x)$, proposed by Huang and Louis [1], is identical to the simple-weighted survival function estimator for cost:

$$S_M^{SW}(x) = \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i I(M_i > x)}{\widehat{K}(T_i)}.$$

Because an integration of the simple-weighted survival estimator will give us back the BT cost estimator, we can conclude that an integration of the Huang and Louis (HL) estimator is also equivalent to the RR estimator. We therefore established the equivalencies among the RR, HL, and BT estimators. Although we used the special case of medical cost estimator here, the implication of this equivalency could be generalized to other mark variables [1].

As one referee suggested, we want to acknowledge that the relationships among various medical cost estimators with censored data have been studied [7, 17–20]. In addition, the inverse probability weighting technique permits covariate information in the cost estimation [21–24]. We did not cover the situation with covariates but focused on the situation without covariates in this paper. Future research may be directed to cases with covariates, other types of censoring [25], or finding user-friendly ways to characterize more sophisticated estimators in this field.

Acknowledgments

We would like to thank Dr. Pfeifer who provided a motivating idea to this work. The authors were partly supported by the National Heart, Lung, and Blood Institute (grant R01 HL096575). We thank reviewers for providing constructive comments.

References

1. Huang Y, Louis T. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*. 1998; 85:785–798.
2. Lin D, Feuer E, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics*. 1997; 53:419–434. [PubMed: 9192444]
3. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.

4. Lin D. Regression analysis of incomplete medical cost data. *Statistics in Medicine*. 2003; 22:1181–1200. [PubMed: 12652561]
5. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika*. 2000; 87:329–343.
6. Zhao H, Tian L. On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics*. 2001; 57:1002–1008. [PubMed: 11764238]
7. Zhao H, Bang H, Wang H, Pfeifer P. On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine*. 2007; 26:4520–4530. [PubMed: 17380543]
8. Efron, B. *Proceedings of the 5th Berkeley Symposium*. Vol. 4. Berkeley: University of California Press; 1967. The two sample problem with censored data; p. 831–853.
9. Miller JRG. What price Kaplan–Meier? *Biometrics*. 1983; 39:1077–1081. [PubMed: 6671119]
10. Efron B. Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association*. 1988; 83:414–425.
11. Pfeifer PE, Bang H. Non-parametric estimation of mean customer lifetime value. *Journal of Interactive Marketing*. 2005; 19:48–66.
12. Dinse GE. An alternative to efron’s redistribution-of-mass construction of the Kaplan–Meier estimator. *The American Statistician*. 1985; 39:299–300.
13. Kalbfleisch, J.; Prentice, R. *The Statistical Analysis of Failure Time Data*. Wiley; New Jersey: 2002.
14. Huang Y. Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association*. 2002; 97:318–327.
15. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
16. Satten GA, Datta S. The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *American Statistician*. 2001; 55:207–210.
17. O’Hagan A, Stevens J. On estimators of medical costs with censored data. *Journal of Health Economics*. 2004; 23:615–625. [PubMed: 15120473]
18. Raikou M, McGuire A. Estimating medical care costs under conditions of censoring. *Journal of Health Economics*. 2004; 23:443–470. [PubMed: 15120465]
19. Hallstrom A, Sullivan S. On estimating costs for economic evaluation in failure time studies. *Medical Care*. 1998; 36:433–436. [PubMed: 9520967]
20. Strawderman RL. Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of American Statistical Association*. 2000; 95:1192–1208.
21. Lin D. Linear regression analysis of censored medical costs. *Biostatistics*. 2000; 1:35–47. [PubMed: 12933524]
22. Willan A, Lin D, Manca A. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine*. 2005; 24:131–145. [PubMed: 15515137]
23. Pullenayegum E, Willan A. Semi-parametric regression models for cost-effectiveness analysis: improving the efficiency of estimation from censored data. *Statistics in Medicine*. 2007; 26:3274–3299. [PubMed: 17309112]
24. B ser O, Gardiner JC, Bradley CJ, Given CW. Estimation from censored medical cost data. *Biometrical Journal*. 2004; 46:351–363.
25. Betensky R. Redistribution algorithms for censored data. *Statistics and Probability Letters*. 2000; 46:385–389.