American Journal of Epidemiology

Vol. 189, No. 11
DOI: 10.1093/aje/kwaa086
Advance Access publication:
May 15, 2020

Practice of Epidemiology

Hidden Imputations and the Kaplan-Meier Estimator

Stephen R. Cole*, **Jessie K. Edwards, Ashley I. Naimi, and Alvaro Muñoz**

* Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: cole@unc.edu).

Initially submitted March 3, 2020; accepted for publication May 11, 2020.

The Kaplan-Meier (KM) estimator of the survival function imputes event times for right-censored and left-truncated observations, but these imputations are hidden and therefore sometimes unrecognized by applied health scientists. Using a simple example data set and the redistribution algorithm, we illustrate how imputations are made by the KM estimator. We also discuss the assumptions necessary for valid analyses of survival data. Illustrating imputations hidden by the KM estimator helps to clarify these assumptions and therefore may reduce inappropriate inferences.

censoring; imputation; loss to follow-up; survival; truncation

Abbreviations: AIDS, acquired immunodeficiency syndrome; KM, Kaplan-Meier.

Applied health scientists will sometimes express inconsistent views regarding survival analyses, such as "I don't want to impute, or make up, any data. Please just show me the Kaplan-Meier estimator!" However, statisticians have recognized for decades (1) that the Kaplan-Meier (KM) estimator (2) implicitly imputes event times for right-censored observations. Likewise, the KM estimator extended to account for late entry (3) implicitly imputes event times for left-truncated observations.

To review briefly, the survival function is the probability of remaining event-free. Formally, $S(t) = P(T > t)$ for some follow-up time t, where T is the time from the origin to the event (4). The KM estimator is the product, taken over the ordered set of distinct event times, of the complement of the number of events divided by the number at risk. Formally,

$$\hat{S}(t) = \prod_{k \in t_k \leq t} (1 - d_k/n_k),$$

where d_k is the number of events and n_k is the number at risk, both at time t_k, the kth distinct event time.

In the setting we discuss here, where we have both right-censored and left-truncated observations, the number at risk n_k excludes persons who have already experienced the event or been censored before time t_k, as well as those who enter follow-up at or after time t_k. The extended (to late entries) KM (5), or Lynden-Bell (3), estimator of the survival function is appropriate for data which include right-censoring and left-truncation. This extension to the original KM estimator simply defines n_k as

$$\sum_{i=1}^{n} I(W_i < t_k \leq T_i),$$

where $I(a)$ is the indicator function (i.e., returns 1 if a is true, 0 if false) and W_i is the time after the origin at which participant i entered the study.

The KM estimator of the survival function is widely used because it is a nonparametric maximum likelihood estimator (and is therefore asymptotically consistent and efficient) (6) and because it is algorithmically simple. But the simplicity of the KM estimator hides the fact that right-censored and left-truncated events are imputed. Here, we make 2 things explicit: 1) how right-censored observations are redistributed via proportions to future event times after the censoring times (i.e., following Efron's redistribution algorithm (1)); and 2) how late entries require imputing the unseen (due to truncation) individuals via odds of the probabilities (i.e., jumps in the KM) at events before the

late entries relative to the event probabilities past the late entries. We illustrate, using redistribution algorithms, how hidden imputations are made in the KM estimator, and we then discuss the critical assumptions necessary for valid analyses of survival data subject to right-censoring and late entries.

HIDDEN IMPUTATIONS

Say that we wish to study the time from diagnosis of acquired immunodeficiency syndrome (AIDS) to death (7), but a participant's data can be right-censored at time c years from the origin of AIDS diagnosis, due to dropout from the study. Additionally, a participant's data can be left-truncated at time w years from the origin, due to entry into the study at w years from the origin of interest (e.g., AIDS diagnosis), often referred as "late entries." In an alternate setting, we might wish to study the time from initiating attempts at pregnancy (i.e., the origin) to conception, where women's data can be right-censored due to dropout or left-truncated due to late entry.

Consider the data for 10 individuals shown in Figure 1A, with the case identifier as the y-axis and time from AIDS diagnosis to death as the x-axis. The data include the number of years between the origin (i.e., AIDS diagnosis) and study entry, denoted as W, and the number of years between this same origin and study exit, denoted as T^*, where T^* is the minimum of T (i.e., years between AIDS diagnosis and death) and C (i.e., years between AIDS diagnosis and censoring). Solid black lines denote time under follow-up, from W_i to T_i^*. Death before censoring is denoted by $\delta = 1$, and in Figure 1A by lines ending with solid dots. Participants 1, 4, 5, and 8 enter the study at AIDS diagnosis (i.e., $w = 0$), while participants 2, 3, 6, 7, 9, and 10 enter the study late, at a known time after AIDS diagnosis (i.e., $w > 0$).

Turnbull extended Efron's redistribution algorithm to allow for left-truncation (8), and this extended redistribution algorithm is equivalent to the extended KM estimator detailed above. In Figure 1B, we show the extended KM estimator for the data in part A. In the example, the extended KM estimate of the survival function at $t = 10$ years is

$$S(10) = \prod_{k \in t_k \leq 10} \left(1 - \frac{d_k}{n_k}\right)$$
$$= \left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{5}\right)\left(1 - \frac{1}{5}\right) = 0.533.$$

In Figure 2 we illustrate the redistribution of all 10 participants. The upper portion (i.e., top 5 rows) of Figure 2 details the 6 distinct event times t_k, the number of events d_k, the size of the risk sets n_k, the extended KM estimator, and the jumps (i.e., step sizes) in the estimator, say p_k. The kernel of Figure 2 consists of the central shaded 10×6 rectangle. Each row in this rectangle represents a participant, and each column represents an event time. The third and fourth columns in Figure 2 (labeled $S(w)$ and "No. Truncated," respectively) provide information on the number of

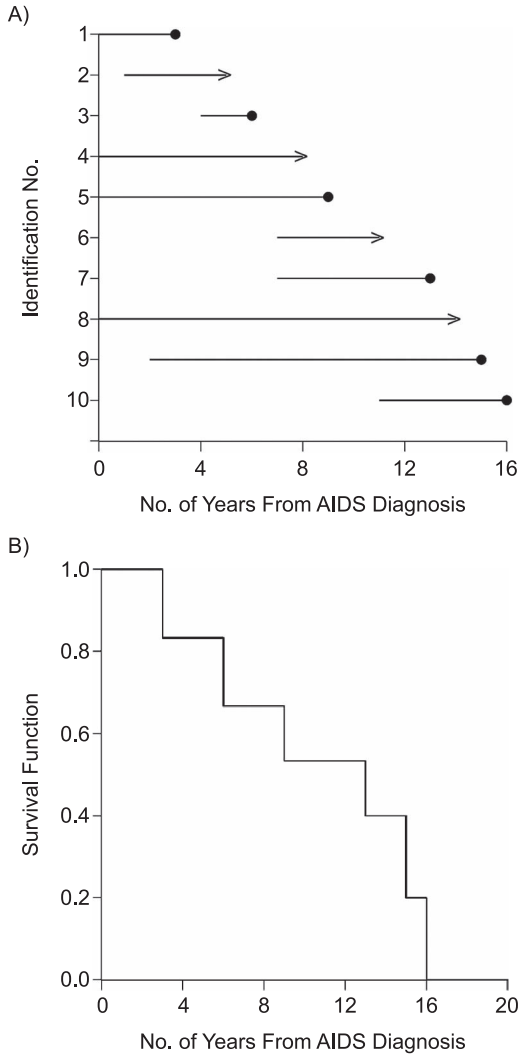

Figure 1. A) Data from 10 hypothetical study participants who were diagnosed with acquired immunodeficiency syndrome (AIDS) during ($n = 4$) or before ($n = 6$) study entry and were followed up to 16 years for death. B) Extended Kaplan-Meier estimator of the survivor function.

truncated events. Here we detail only the most illustrative case, the case of participant 6, who enters the study late and is censored. In Web Appendix 1 (available at https://doi.org/10.1093/aje/kwaa086), we provide details for the remaining 9 participants.

Hidden imputations for censoring

What is hidden in the KM algorithm is that right-censored observations are allocated as (partial) events in proportion to all events that occur *after* the observation's censoring time. Specifically, if $t_{k(c)}$ is the first event time *after* an individual is censored at time c, the KM estimator imputes $p_i / \sum_{k \geq k(c)} p_k$ events at the ith event time *after* c, which corresponds to the probability that the individual censored

Event times (t_k)	3	6	9	13	15	16
No. of events (d_k)	1	1	1	1	1	1
Risk set (n_k)	6	5	5	4	2	1
Survival (S_k)	0.833	0.667	0.533	0.400	0.200	0.000
Jumps (p_k)	0.167	0.167	0.133	0.133	0.200	0.200

ID	Data: $(w, t; \delta)$	$S(w)$	No. Truncated						Total
1	0, 3; 1	1	0	1	0	0	0	0	0	1
2	1, 5; 0	1	0	0	0.200	0.160	0.160	0.240	0.240	1
3	4, 6; 1	0.833	*0.2*	*0.200*	1	0	0	0	0	1.2
4	0, 8; 0	1	0	0	0	0.200	0.200	0.300	0.300	1
5	0, 9; 1	1	0	0	0	1	0	0	0	1
6	7, 11; 0	0.667	*0.5*	*0.250*	*0.250*	0	0.250	0.375	0.375	1.5
7	7, 13; 1	0.667	*0.5*	*0.250*	*0.250*	0	1	0	0	1.5
8	0, 14; 0	1	0	0	0	0	0	0.500	0.500	1
9	2, 15; 1	1	0	0	0	0	0	1	0	1
10	11, 16; 1	0.533	*0.875*	*0.3125*	*0.3125*	*0.250*	0	0	1	1.875
			Total no. of events (M_k)	2.0125	2.0125	1.610	1.610	2.415	2.415	12.075
			Jumps (p_k)	0.167	0.167	0.133	0.133	0.200	0.200	1.0

Figure 2. Redistribution of right-censored and left-truncated observations among 10 hypothetical study participants who were diagnosed with acquired immunodeficiency syndrome (AIDS) during ($n = 4$) or before ($n = 6$) study entry and were followed up to 16 years for death. ID, identification.

at c will develop the event at the ith event time *after* c. For example, participant 6 was censored at 11 years after AIDS diagnosis and therefore has their unit mass distributed among the 3 remaining events occurring at 13, 15, and 16 years after AIDS diagnosis. This censored observation is distributed proportionally given the jumps in the extended KM estimator, which are 0.133 (i.e., $0.533-0.4 = 0.133$), 0.2 (i.e., $0.4-0.2 = 0.2$), and 0.2 (i.e., $0.2-0 = 0.2$), respectively. Therefore, this censored observation is distributed (or imputed) as 1/4, 3/8, and 3/8, respectively. For example, the first allocation is obtained as $1/4 = 0.133/(0.133+0.200+0.200)$.

Hidden imputations for left-truncation

Also hidden in the extended KM algorithm is that left-truncated events are allocated as odds of the event probabilities at events before the late entries relative to the sum of the probabilities past the late entries. For a late entry at time w, we first determine the number of unseen truncated events, which is calculated as $[1-S(w)]/S(w)$ (see Web Appendix 2 for derivation). Then, we allocate these "ghosts" (8) proportional to all events that occurred *before* time w. This 2-step algorithm is equivalent to imputing $p_j/\sum_{k\geq k(w)}p_k$ events at the jth event time *before* w (i.e., $j < k(w)$), where $t_{k(w)}$ is the first event time *after* an individual enters the study at time w. This number of imputed events corresponds to the number of unseen individuals who are peers of the recruited individuals but were truncated (not enrolled) because they developed the event before the recruited individual at w years since origin. For example, using the 2-step algorithm, participant 6 entered follow-up at 7 years after AIDS diagnosis, when the survival function is $S(7) = 2/3$. Consequently, participant 6 contributes $[1 - S(7)]/S(7)$, or 1/2 of a truncated event. This 1/2 truncated event is distributed to the 2 observed event times between the origin and 7 years after AIDS diagnosis, which occur at years 3 and 6 for participants 1 and 3, respectively. In effect, the step sizes in the risk function at times 3 and 6 will account for the unobserved risk contributed by the truncated "ghosts" of participant 6 prior to their late entry. This 1/2 truncated event is distributed proportionally given the jumps in the extended KM curve at times 3 and 6 years, which are both 0.167. Because the jump sizes are the same at times 3 years and 6 years, the truncated event is divided equally as 0.250 and 0.250. In this way, the extended KM estimator is able to "impute" events due to truncated observations by redistributing them in proportion to the size of the jumps in the survival function.

To complete Figure 2, we sum across and down the central 10×6 rectangle. Taking the column totals (2.0125, 2.0125, 1.610, 1.610, 2.415, and 2.415) and dividing by the sum of the row totals (12.075) yields the bottom row in Figure 2, which is identical to the jumps from the extended KM estimator given in the top portion of Figure 2. The completed Figure 2 provides an explicit picture of how the extended KM estimator of the survival function imputes right-censored and left-truncated events. Discussion of the standard errors and confidence intervals using the total number of events after the imputations are completed is provided in Web Appendix 3.

ASSUMPTIONS

The extended KM estimator is simple and succinct, but it does not make apparent that right-censored and left-truncated events are imputed in the manner illustrated above by redistribution. The validity of the extended KM estimator rests on the dual conditions that right-censoring and left-truncation are both random, conditional on any variables used to stratify the curves. The explicit imputation shown in Figure 2 helps to make clear the assumptions necessary for valid estimation of the survival curve. For example, we allocate a censored observation *proportionally* to *all* events after the censoring time as a direct consequence of the assumption that the individual was censored at random. If the individual was instead censored at random conditional on some set of measured covariates, that person's unit mass ought to be redistributed proportionally, but only to those events with the same covariate set. If the individual was instead censored conditional on some set of unmeasured covariates, then we do not know how to redistribute them, and the survival curve is not identifiable or computable given the observed data. Likewise, we allocate any unseen events *proportionally* to *all* events before the late entry time as a direct consequence of the assumption that the individual entered the study at random. If the individual instead entered at random conditional on some set of measured covariates, then their unseen events should be redistributed proportionally, but only to those events with the same covariate set. If the individual instead entered conditional on some set of unmeasured covariates, then we do not know how to redistribute them, and the survival curve is again not identifiable given the observed data.

The characterization of the hidden imputation used by the extended KM estimator presented here for the case of uninformative censoring and truncation can also be used to arrive at unbiased estimates of the survival function for cases of interval-censored data and for data with informative censoring where the informative mechanism is known. The key is to redistribute not to all event times after censoring, or to all event times before late entries, but to only the pertinent subsets (e.g., if an individual's event occurs in the interval c_1, c_2, it should only be redistributed to event times observed in that interval).

In summary, methods with which to appropriately account for right-censoring and left-truncation in epidemiologic studies impute right-censored and left-truncated events, to the chagrin of those who wish to avoid imputations. These or any method used to recover right-censored or left-truncated events must operate under assumptions. While epidemiologists are generally aware that assumptions are required for valid estimation of the survival function, illustrating hidden imputations in survival analyses helps to clarify these assumptions and may reduce unnecessary violations.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Stephen R. Cole, Jessie K. Edwards); Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania (Ashley I. Naimi); and Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (Alvaro Muñoz).

This work was supported in part by National Institutes of Health grants K01AI125087 (J.K.E.) and R01 HD093602 (A.I.N.).

This work is dedicated to Camilo Alvaro Steil (the senior author's grandson), who was born on the day the paper was provisionally accepted for publication.

Conflict of interest: none declared.

REFERENCES

1. Efron B. The two-sample problem with censored data. In: Le Cam LM, Neyman J, eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 4: Biology and Problems of Health*. Berkeley, CA: University of California Press; 1967:831–852.
2. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282): 457–481.
3. Lynden-Bell D. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon Not R Astron Soc*. 1971;155(1):95–118.
4. Cole SR, Hudgens MG. Survival analysis in infectious disease research: describing events in time. *AIDS*. 2010;24(16): 2423–2431.
5. Lamarca R, Alonso J, Gómez G, et al. Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population. *J Gerontol A Biol Sci Med Sci*. 1998; 53(5):M337–M343.
6. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am J Epidemiol*. 2014;179(2):252–260.
7. Cox C, Chu H, Schneider MF, et al. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat Med*. 2007;26(23):4352–4374.
8. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc Series B Stat Methodol*. 1976;38:290–295.