WEB MATERIAL

Hidden Imputations and the Kaplan-Meier Estimator

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

Correspondence to Dr. Stephen 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).

Table of Contents

Web Appendix 1: Calculations for 9 Participants Not Detailed in the Text

Web Appendix 2: Derivation of the Number of Truncated, or Unseen, Events

Web Appendix 3: Standard Errors and Confidence Intervals

Web Table 1. Redistribution of interval-censored (between t_1 and t_2) left-truncated observations (w) among 10 participants who were diagnosed with AIDS during (n = 4) or before (n = 6) study entry and were followed up to 16 years for death.

Reference

Web Appendix 1: Calculations for 9 Participants Not Detailed in the Text

Here we provide details of calculations for the 9 participants not discussed in the text. First, we cover the simplest case of an observed events that entered follow-up at diagnosis of acquired immunodeficiency syndrome (AIDS) (i.e., participants 1 and 5). When participant 1 entered follow-up (at their origin) the survival function was 1; hence the number of truncated events is 0. Because this participant died 3 years after AIDS diagnosis their entire unit mass is placed in that corresponding column. Likewise, for participant 5, but the unit mass is placed at year 9.

Next, we detail participants who entered study at AIDS diagnosis and were censored due to drop out before death (i.e., participants 4 and 8). Participant 4 entered follow-up at the origin, when the survival function was 1; hence the number of truncated events is 0. Participant 4 was censored at 8 years after AIDS diagnoses, and therefore their unit mass is distributed among the latter four event times occurring at: 9, 13, 15, and 16 years after AIDS diagnosis. This censored observation is distributed proportionally given the jumps in the extended Kaplan-Meier (KM) curve, which are: 0.133, 0.133, 0.2, and 0.2, respectively. Therefore, this censored observation is distributed (or imputed) as: 0.2, 0.2, 0.3, and 0.3, respectively. For example, the first mass is obtained as 0.2 = 0.133 / (0.133 + 0.133 + 0.2 + 0.2). In a similar fashion, participant 8 is redistributed with 0.5 of their mass at each of the event times 15 and 16 years after AIDS diagnosis.

Next, we detail participants who enter study after AIDS diagnosis and die during follow-up (i.e., participants 3, 7, 9, and 10). Participant 3 enters follow-up at 4 years after AIDS diagnosis. At that time, $S(4) = 0.833$. Therefore, the number of truncated events or ghosts that this late entry must carry is (1 – 0.833)/0.833 = 0.2. These 0.2 truncated events are distributed to the single observed event time between the origin and 4 years after AIDS diagnosis, which is at year 3. For a more interesting example, consider participant 10. This participant enters follow-up at 11 years after AIDS diagnosis. At that time, $S(11) = 0.533$. Therefore, the number of truncated events that this late entry must carry is (1 – 0.533)/0.533 = 0.875. These 0.875 truncated events are distributed to the three observed event times between the origin and 11 years after AIDS diagnosis, which are at years 3, 6, and 9. These 0.875 truncated events are distributed proportionally given the jumps in the extended KM curve, which are: 0.167, 0.167, and 0.133, respectively. Therefore, the 0.875 is distributed as: 0.3125, 0.3125, and 0.25, respectively. For example, the first mass is obtained as 0.3125 = 0.875 × [0.167/(0.167 + 0.167 + 0.133)]. In addition to the 0.875 truncated events, this participant died at 16 years after AIDS diagnosis. Consequently, their entire unit mass is placed in the column for 16 years. This same process is undertaken for participant 7. When this same process is undertaken for participant 9, there are no truncated events because although participant 9 enters follow-up late, at *w* = 2 years after AIDS diagnosis, the survival function remains at 1 at that time.

Last, we detail participant 2 who entered follow-up (at 1 years after AIDS diagnosis) the survival function remained 1; therefore, the number of truncated events is 0. Participant 2 was censored at 5 years after AIDS diagnosis and therefore their unit mass is distributed among the latter five event times occurring at: 6, 9, 13, 15 and 16 years after AIDS diagnosis. This censored observation is distributed proportionally given the jumps in the extended KM curve, which are: 0.167, 0.133, 0.133, 0.2, and 0.2, respectively. Therefore, this censored observation is distributed (or imputed) as: 0.2, 0.16, 0.16, 0.24, and 0.24, respectively. For example, the first mass is obtained as 0.2 = 0.167 / (0.167 + 0.133 + 0.133 + 0.2 + 0.2).

Web Appendix 2: Derivation of the Number of Truncated, or Unseen, Events

One may ask why there are $[1 - S(W_i)]/S(W_i)$ truncated events for participant i. First, recognize that each participant observed from their origin to w years either survived to, or died by, w years. Specifically, $1 = S(w) + [1 - S(w)]$. Next, ask what size study do we need to see one participant survive w years? Dividing the above formula by $S(w)$ yields a study of size $1/S(w)$. Specifically, $\frac{1}{S(w)} = \frac{S(w)+[1-S(w)]}{S(w)} = 1 + \frac{[1-S(w)]}{S(w)}$. Where the first addend on the rightmost side is the single participant surviving to time w, and the second addend is the number of truncated events. Consequently, when we observe participant i to enter study alive at time W_i, the number of truncated events is $[1 - S(W_i)]/S(W_i)$.

For example, say at time w years the survival function is at 1/4, then we need a sample of size 4 to see 1 participant survive to time w with 3 truncated events by time w, or

$$\frac{1}{S(w)} = 1 + \frac{[1 - S(w)]}{S(w)}$$

$$\frac{1}{\frac{1}{4}} = 1 + \frac{\left[1 - \frac{1}{4}\right]}{\frac{1}{4}}$$

$$4 = 1 + 3.$$

Web Appendix 3: Standard Errors and Confidence Intervals

After both imputations of events due to censoring and late entries are completed, one will have M_k events at k^{th} time t_k when events were observed, which is the sum of the number of observed events at that time (typically one in the common case of no ties) plus the imputed events due to the censored observations before t_k and the late entries after t_k (shown in the second to last row in Table). It follows that the hazard function at t_k is estimated as $h_k = M_k/\sum_{j\geq k} M_j$ and thus, by following the same usage of the delta method in Greenwood's formula (1), the standard error of the cumulative hazard function (= -logarithm of the survival function) at t_k is $SE_k = \sqrt{\sum_{j\leq k} h_j^2/[(1-h_j)d_j]}$. Hence, the logarithm of the cumulative hazard will have standard error $SE_k/(-\log(S_k))$ and correspondingly, a pointwise 95% confidence interval for the survival function S_k $(= \sum_{j>k} M_j/\sum_{all\ j} M_j)$ at t_k is given by $S_k^{\exp\,[\pm SE_k/(-\log(S_k))]}$.

To illustrate the application of the imputation approach to interval-censored data and the calculation of 95% confidence intervals, the Supplemental Material (Table 1) below shows all calculations for data where the only change from that in the manuscript is that the participant 2 is not right-censored at time 5 (i.e., event occurs at any time after 5 years) but in an interval between 5 and 10 years (i.e., event to be redistributed to only at observed event times 6 and 9).

Web Table 1. Redistribution of interval-censored (between t_1 and t_2) left-truncated observations (w) among 10 participants who were diagnosed with AIDS during (*n* = 4) or before (*n* = 6) study entry and were followed up to 16 years for death

ID		**3**	**6**	**9**	**13**	**15**	**16**	**Total**
	Data: w, t_1, t_2							
1	0, 3, 3	1	0	0	0	0	0	1
2	1, 5, 10	0	0.576	0.424	0	0	0	1
3	4, 6, 6	*0.200*	1	0	0	0	0	1.2
4	0, 8, ∞	0	0	0.263	0.184	0.277	0.277	1
5	0, 9, 9	0	0	1	0	0	0	1
6	7, 11, ∞	*0.271*	*0.356*	0	0.250	0.375	0.375	1.627
7	7, 13, 13	*0.271*	*0.356*	0	1	0	0	1.627
8	0, 14, ∞	0	0	0	0	0.500	0.500	1
9	2, 15, 15	0	0	0	0	1	0	1
10	11, 16, 16	*0.368*	*0.483*	*0.356*	0	0	1	2.207
	Total events, M_k	2.110	2.771	2.043	1.434	2.152	2.152	12.662
	Survival, S_k	0.833	0.615	0.453	0.340	0.170	0	
	95% CI for S_k	0.27-0.97	0.13-0.89	0.08-0.78	0.05-0.67	0.01-0.52	N/A	

Abbreviations: CI, confidence interval; ID, identification; N/A, not applicable.

Reference

1. Miller R. *Survival Analysis*. New York, NY: John Wiley & Sons, Inc.; 1981.