Let V be the time at which T is reported, so in the context above we have $V = A_k$. Note that at time V the computer contains the full data structure $X \equiv \bar{X}(V)$, which corresponds with observing T, $\bar{V}_1(T)$, and $\bar{W}(T)$. We assume that the reporting delay is finite; that is, if a subject dies, then it will eventually always be reported, which is a reasonable assumption for these central registries, although the reporting delay can be large. Let C represent the time at which the data analyst stops receiving information on the subject. For example, C might be the time at data analysis, but it could also be the time at which the subject switches treatment or leaves the study so that his/her true survival time can no longer be recovered. Living in the time scale of the data analyst, one observes the process X up to the minimum of C and V, and one knows whether this minimum is the censoring time or is the time V at which T is reported. If $C \geq V$, then the full data $\bar{X}(V)$ are observed, and if $C < V$, then the right censored data $\bar{X}(C)$ is observed. Thus, the observed data structure can be represented as

$$Y = (\tilde{T} \equiv C \wedge V(T), \Delta \equiv I(V(T) \leq C), \bar{X}(\tilde{T})), \tag{3.4}$$

which corresponds with the data structure studied in this chapter. We observe n independent and identically distributed observations $Y_1, \ldots, Y_n$ of Y. Note that CAR allows the probability of being censored at time c, given one is not censored yet, to depend on the reporting delay history and the observed covariate history. As in the previous examples, possible parameters of interest are the marginal distribution of T and regression parameters in Cox proportional hazard models or linear regression models with outcome $\log(T)$.

One should note that reporting delay causes bias in the naive Kaplan–Meier estimator of the distribution of T. For simplicity, let us assume that the U_j's are reported immediately (i.e., $A_j = U_j$) but that T is reported at a possibly delayed time A_k. Let C be the time at analysis, which is assumed to be independent of T. If death is reported before the censoring time C (i.e., $A_k < C$), then the censoring variable is simply C. Suppose now that at time C death has yet to be reported (i.e., $A_k > C$) and C is between $U_{j-1} = A_{j-1}$ and $U_j = A_j$. Then we cannot be sure that T did not happen between U_{j-1} and C since all we know is that $T > U_{j-1}$. It is common practice to set $C = U_{j-1}$ and thus let T be right-censored at U_{j-1}. The censoring variable is now a function of A_k and thus of T, which implies that censoring is no longer independent of T. This can lead to serious bias in the Kaplan–Meier estimator, as nicely illustrated in a simulation in Hu and Tsiatis (1996) and an analysis of the California AIDS Registry in Hubbard, van der Laan, Enaroria, and Colford (2000) earlier analyzed in Colford et al. (1997).

We already illustrated the applications of this data structure in studies in which data on a subject are reported with delay to the data analyst. The reporting delay data structure also appears naturally in the following interval-censored data type of application. Consider a T defined by an event

that can only be detected at a monitoring time; for example, T might be the time of onset of a tumor or the time at which the CD4 count of an AIDS patient drops below a particular value. At the monitoring time, one can find out whether T happened, and if it happened one might be able to determine the precise value of T (or use an extrapolated approximation) between the last and current monitoring times. In this case, we have no reporting delay for the U_j's (i.e., $A_j = U_j$), $j = 1, \ldots, k-1$, but T is reported at the monitoring time A_k following T. If the precise value of T between two subsequent monitoring times can only be guessed, then in these kinds of applications it is also common practice to apply the Kaplan–Meier estimator to the guessed T's and be satisfied with estimation of their distribution. In both situations, the Kaplan–Meier estimator can be expected to be biased due to the "reporting delay" of T, while the data structure above will acknowledge this reporting delay phenomenon.

In van der Laan and Hubbard (1998), locally efficient estimators of the survival function have been developed and implemented analogously to the methods presented in this chapter.

3.2.4 *Univariately right-censored multivariate failure time data*

Consider a longitudinal study in which various time variables $\vec{T} = (T_1, \ldots, T_k)$ on the subject are of interest. For example, consider a study in which HIV-infected subjects have been randomized to treatment groups. In such a study, one might be concerned with comparing the multivariate treatment-specific survival functions of time from the beginning of the study until AIDS diagnosis, death, and time until particular AIDS-related events (e.g., types of opportunistic infections). As a second example, one might be interested in the bivariate survival function of time until recurrence of cancer (measured from extraction of tumor) and time until death. In this setting, the researcher could also have interest in the estimation of functions of the joint distribution, such as the distribution of the gap time $T_2 - T_1$ from recurrence to death. We will not require that time variables $T_1, \ldots, T_k$ be ordered. Let $L(t)$ represent a time-dependent covariate process that one measures on the subject over time. This process includes the baseline covariates $L(0)$. The full data on a subject is defined as $X = (\vec{T}, \bar{L}(T))$, where $T \equiv \max(T_1, \ldots, T_k)$.

Let C be the common right-censoring time, which could be the minimum of time until end of study or time until dropout of the subject. Each subject is observed until $\tilde{T} = \min(T, C)$. Let $\tilde{T}_j \equiv \min(T_j, C)$, $j = 1, \ldots, k$. Thus, the researcher observes, for each subject, the following data structure:

$$Y = (\tilde{T}_1, \ldots, \tilde{T}_k, \Delta_1 = (T_1 < C), \ldots, \Delta_k = (T_k \leq C), \bar{L}(\tilde{T})). \tag{3.5}$$