Kaplan-Meier estimator with delayed entry

Since this matter seemed confusing on our first day, or at least raised interest, I did some searching for literature on the topic. There is quite a bit of this.

Suppose data t_i represent in the usual manner right-censored survival data, but that the entry time for each individual is some known value e_i. That is, t_i, known to be from the sampling scheme at least e_i, is the smaller of a failure time t_i^* and a censoring time c_i^*, where the censoring is "uninformative". I think you may see that formulating a precise probability model for all of this is at least daunting --- should the censoring times and the entry times be considered as observations from some distributions? One may need to worry less about precise assumptions on the censoring times (this is pretty standard) but it may become critical for our concerns to think about a probability model for the entry times.

I suspect one of the best papers on this is that by Keiding & Gill (Ann. Stat. 1990, p. 582), but it is pretty difficult. They refer to quite a few other papers. The first one I had encountered in my search was the much simpler one by Tsai, Jewell & Wang (Biometrika 1987, p. 883), but I found that probably misleading considered alone.

Note first that without delayed entry the K-M estimator is estimating the common cdf of the T_i^*, that is the distribution of failure times had there been no censoring. With delayed entry I think it is impossible to do any such estimation without pretty strong assumptions about the entry times. For example, suppose first that all the entry times are a given number e^1. Then I think it is clear that the usual K-M estimator is estimating $pr(T_i^* > t \mid T_i^* > e^1)$ (for all values of *t*). Now suppose that there were two entry times e^1, e^2 and that it were know which entry time corresponds to which individual. Clearly there are two K-M estimators that can be computed, one for each part of the sample according to entry time, and these are estimating respectively $pr(T_i^* > t \mid T_i^* > e^2)$ and $pr(T_i^* > t \mid T_i^* > e^2)$. Now there might be some kind of (modified) K-M estimator to be used for all the data together, but the issue is in this setting *what is it estimating*? A possible answer is that it estimates some weighted average of those two distributions.

I think the resolution of this by Keiding & Gill involves assuming that the entry times are a sample from some distribution. Then what I think they show is that it is possible, although maybe not straightforward, to estimate *both* the unconditional distribution $pr(T_i^* > t)$ and the distribution of entry times. Note that by first principles we have that $pr(T_i^* > t) = E\{pr(T_i^* > t \mid T_i^* > e)$, where the expectation here is over the distribution of the conditioning event (a generalization of the weighted average I suggested above).

So it seems a bit less hopeless than I said to use the K-M estimation with delayed entry, but it is complicated and the result may or may not be what one really wishes to estimate. Further, even to carry out the estimation just discussed, some modification of the basic K-M estimator is perhaps required, and I think this is what the Tsai et al paper is about.

I am confident that usually the best resolution of all this is simply to estimate the *hazard function* of the failure times, rather than in essence the *cumulative hazard* (equivalent to the survival function). They delayed entry is simply no problem for estimating the hazard function, and all the difficulties arise when one tries to estimate the integral of this. Why make the problem so difficult, when there is a simple solution?