

Cohort Data II

Survival Analysis



Questions About the Homework?



Resources for the Curious

- Paul D. Allison. *Survival Analysis Using SAS: A Practical Guide*
- David Machin, Yin Bun Cheung, Mahest Parmar. *Survival Analysis: A Practical Approach*
- John P. Klein and Melvin Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*
 - This is a very technical treatment



Why Survival Analysis?

- Time matters for many contexts
 - Slowing down the progression of a disease
 - Differing hospital lengths of stay
 - Time until a particular threshold is reached
- Understanding how risk changes over time gives a more nuanced view of the exposure-disease relationship than a snapshot at the arbitrary end of the study
- Cohorts have temporality built into them – we should exploit it!



Describing Time

- The Origin:
 - There is (at least one) natural time origin when a subject is *at risk of the event*
 - Death: Birth
 - Hospital Discharge: Hospital Admission
 - Initiation of Treatment: Death
 - Initiation of Treatment: Cure
- In trials, randomization is often taken as the origin, in observational studies we often choose



Immortal Person-Time

- The origin is the beginning of the time *at risk*
- A subject may spend time in the study not at risk
- It's inappropriate to include this when studying survival time
- Examples:
 - A workplace cohort where you enroll workers after 6 months of employment to study a disease outcome
 - By definition in those six months they could not have had the outcome
 - Infectious diseases: Any time when the subject had no exposure to the infectious agent



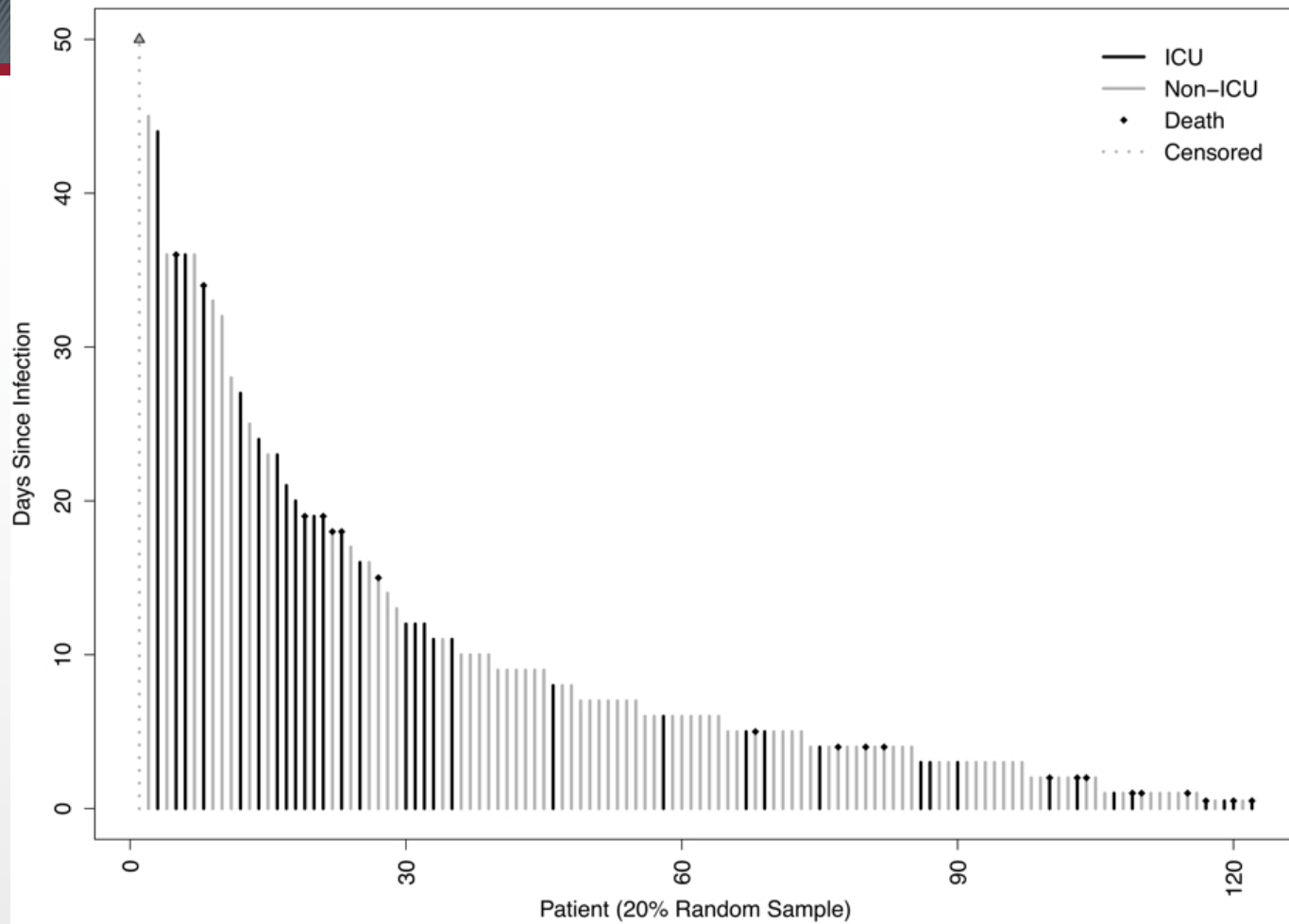
Event

- The event occurs at a time T
- This can be a single event (i.e. death) or a repeatable event (i.e. infection with a particular disease)



Censoring

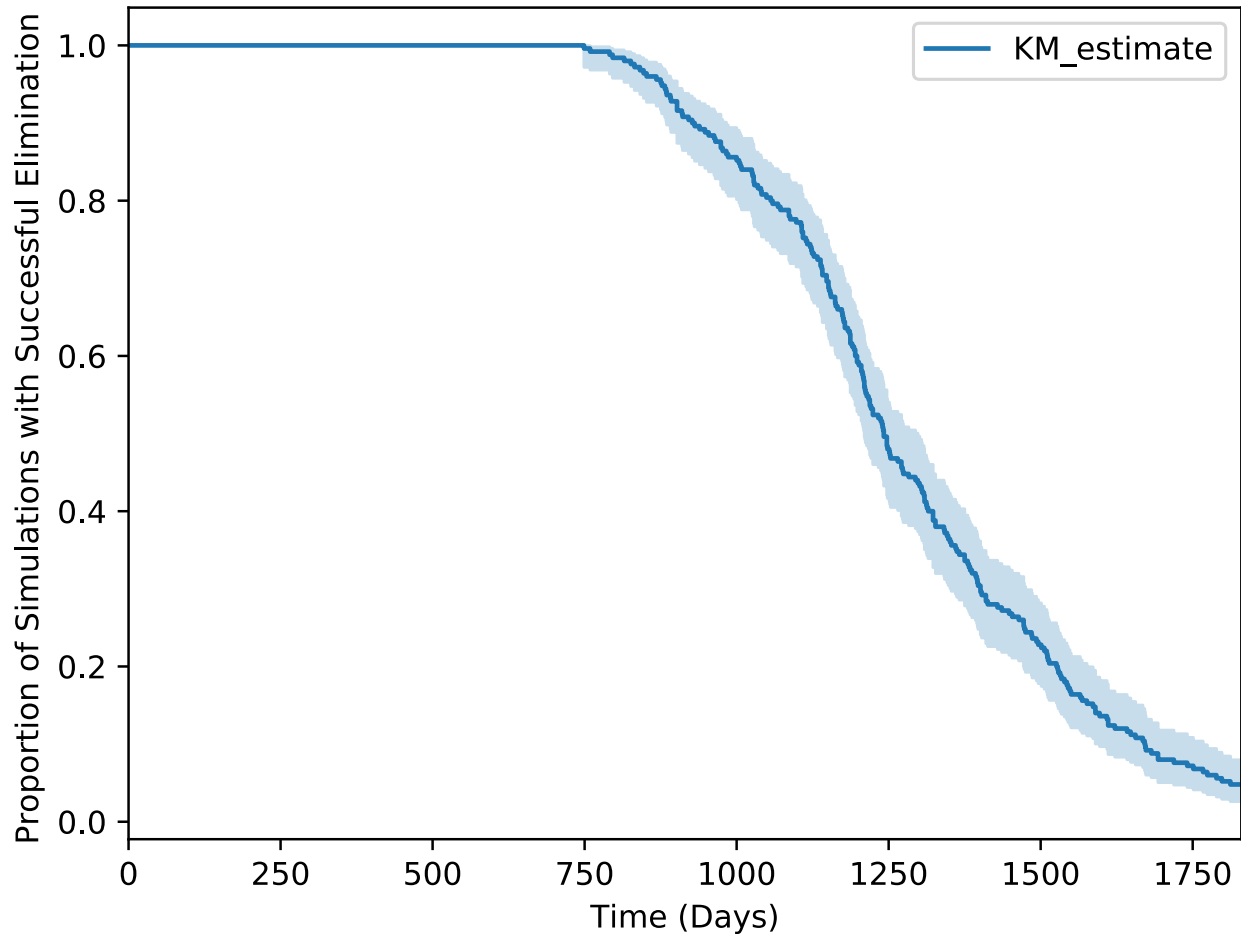
- This is one of the major problems in conducting survival analysis studies
- Sometimes we *don't know* T . This is known as censoring
- Left Censoring: We know T was before some value, but not when
- Right Censoring: We know T was after some value, but not when
- Interval Censoring: We know T was between two values





Survival Function

- Probability of an event taking place greater than some specified time t
- $S(t) = P(T > t)$
- $S(0) = 1$, $S(\text{infinity}) = 0$ (in most cases)





Hazard

- A function of time
- $h(t) = \lim_{\Delta t \rightarrow 0} P(t \leq T < t + \Delta t)$
- Super clear, right?
- Relating the two:
 - The hazard is the slope of the survival function at t , divided by $S(t)$
- A constant hazard results in an exponentially distributed survival function



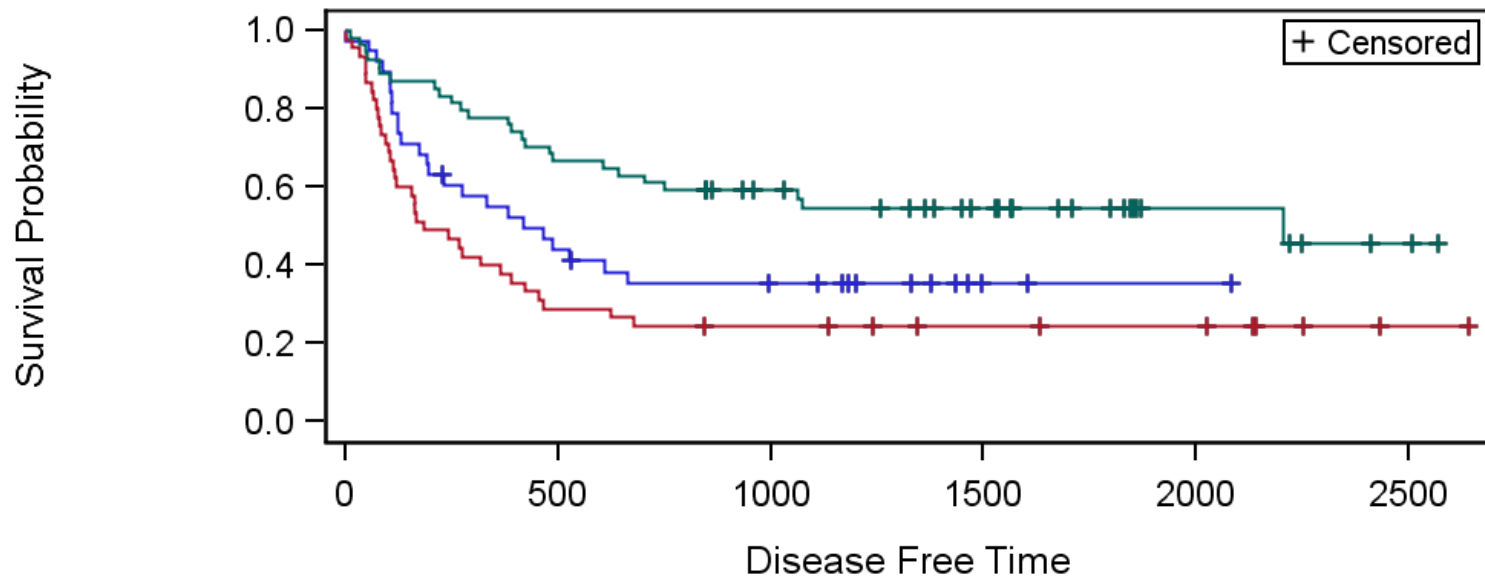
Kaplan-Meier Methods

- Non-parametric way to calculate a survival function
- Fairly approachable – one can in principle calculate these by hand
- One of the preferred methods of analysis in epidemiology
- Often can only compare stratified groups
 - There are ways of controlling for many variables using inverse probability weights



Product-Limit Survival Estimates

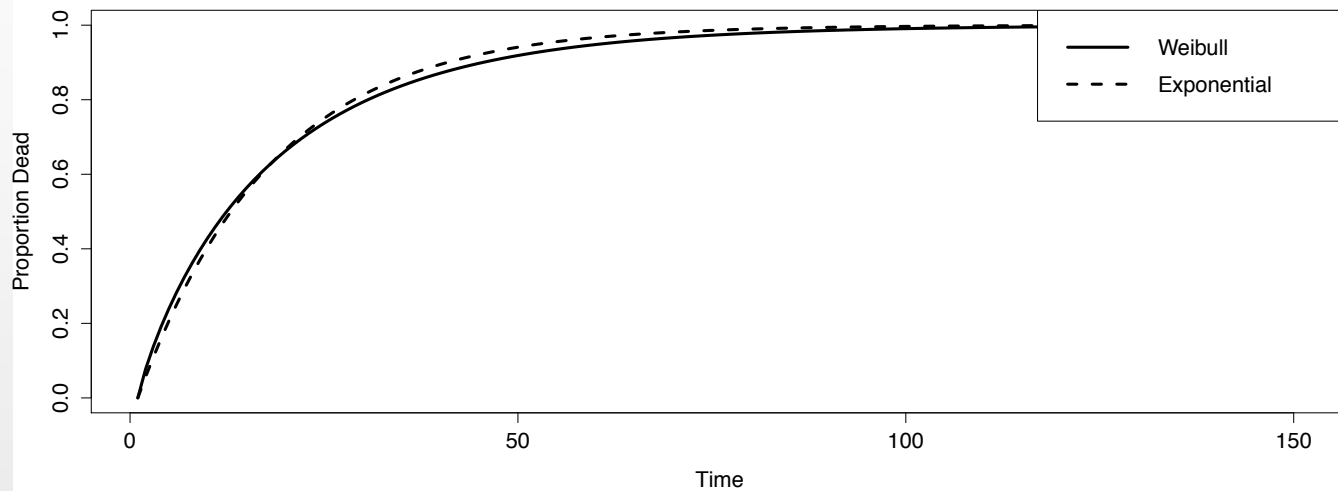
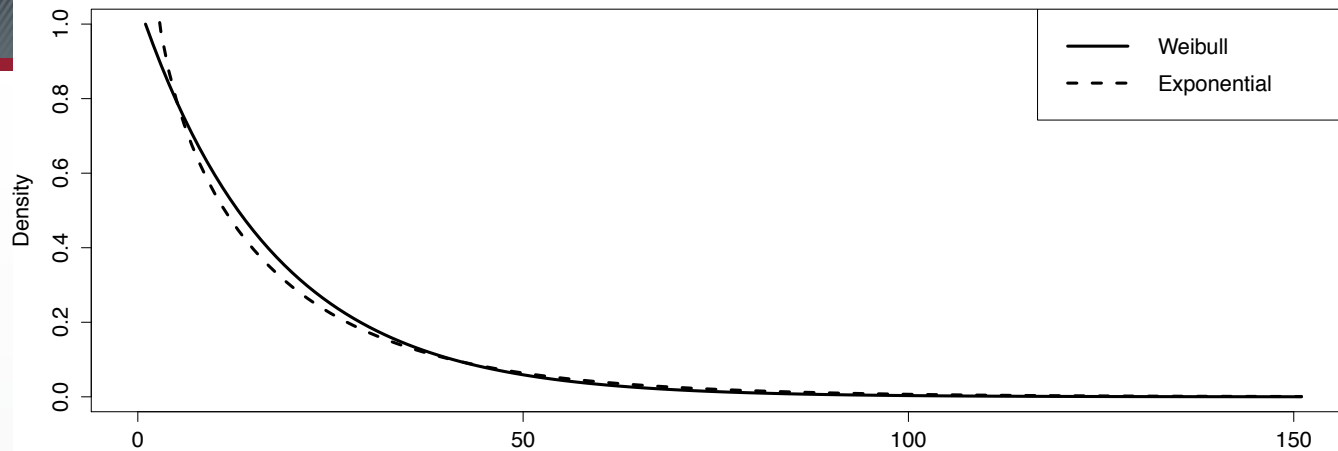
With Number of AML Subjects at Risk

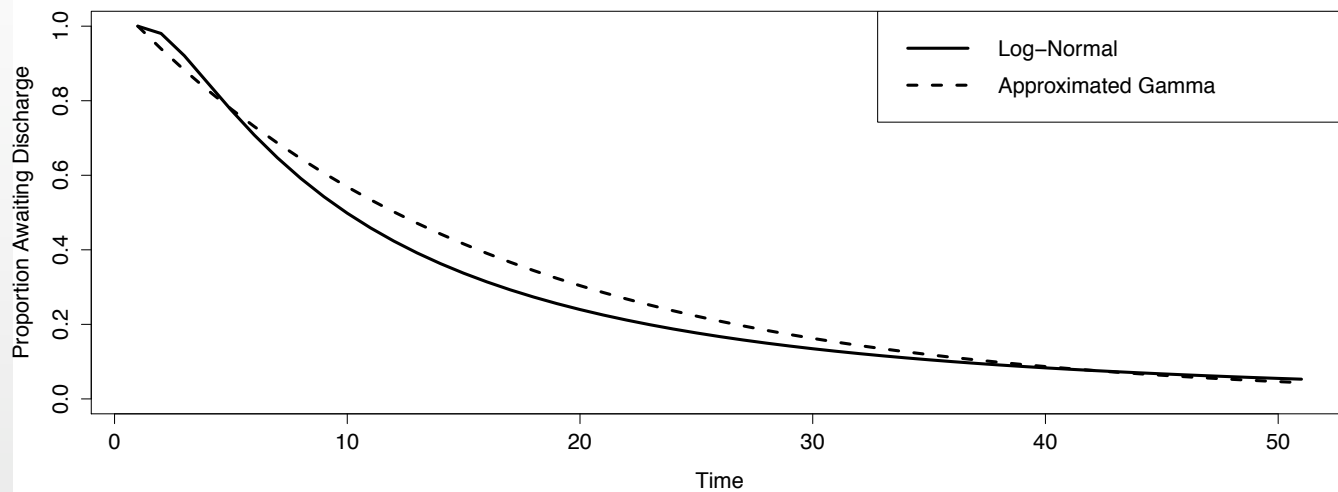
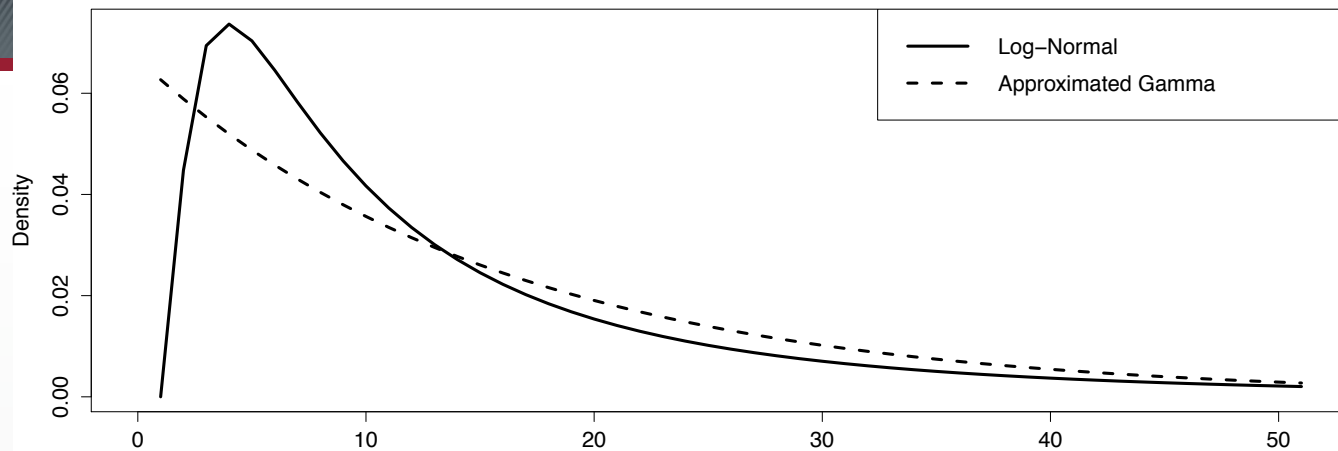




Parametric Survival Models

- Estimating the survival function directly using a parametric model
- Useful for projecting survival beyond the data, or when you need to use a known distribution to generate survival times for another purpose
 - Mathematical modeling, etc.
- May be more precise
- May be more robust to model misspecification







Problems...

- These estimate relative *time* not relative *hazard*
- Not comparable to a RR
- Exponential and Weibull distributions have transformations, more complex distributions do not



Hazard Ratios

- $HR = h_1(t) / h_0(t)$
- $\exp(\beta) = h_1(t) / h_0(t)$
- $h_1(t) = h_0(t) * \exp(\beta)$



Cox Proportional Hazards Model

- “Semi-parametric”
- Uses a partial likelihood method that factors out $h_0(t)$ so it doesn't need to be estimated
- Because of this it is semi-parameteric
 - You have a parameter for the *ratio* of hazards, but not for the underlying hazard itself
- As the name suggests, this assumes hazards are proportional through time



Check Your Assumptions

- $\log\text{-}\log S(t)$ over time should be parallel
- Fit a variable that is a function of time, make sure it's ~ 0 .

