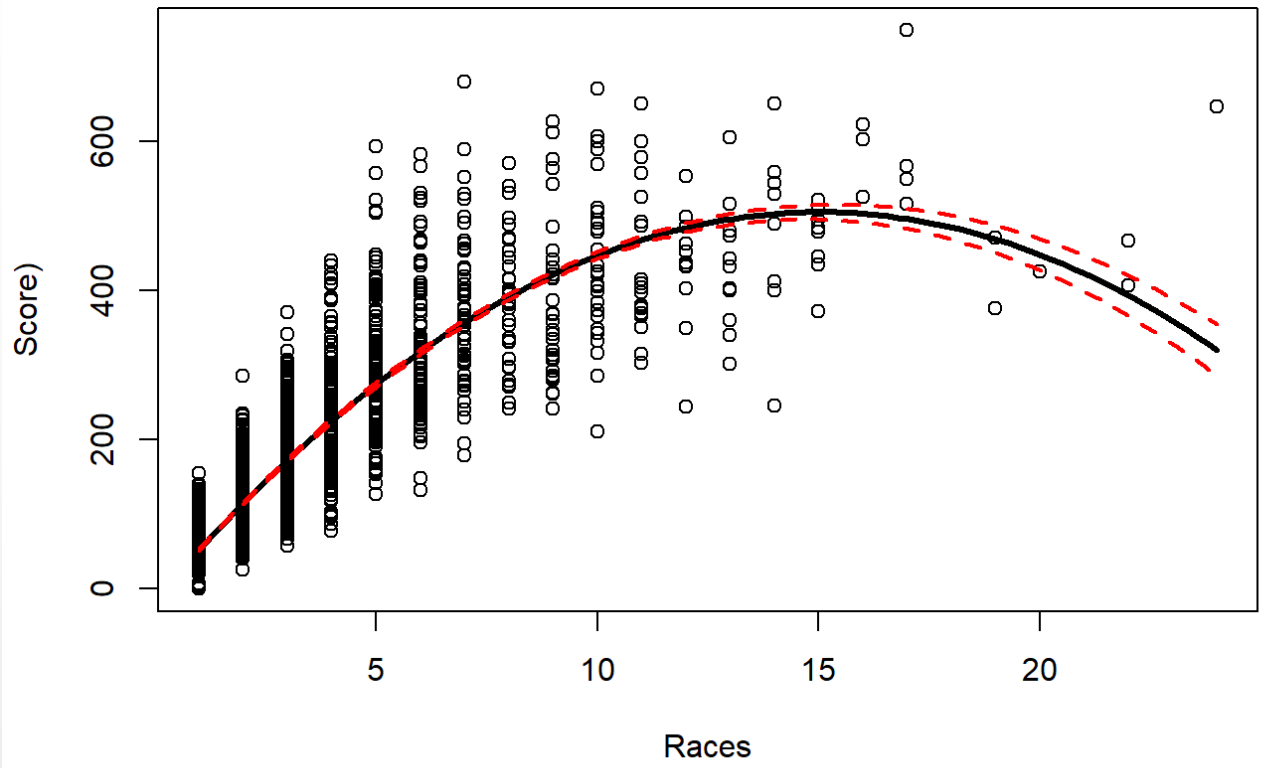
WASHINGTON STATE UNIVERSITY

Cohort Data II: Time to Event Data and Survival Analysis

Questions from Last Class?

PS 2

- A lot of you expressed that you weren't fully satisfied with your fits from the Capybara Racing League (real data, btw, about another type of competitive event
- This is a good thing – why?


Score)
0
200
400
600
5
10
15
20
Races

Remember Your Question!

- The MERS data set has an outcome of *fatal* MERS infections
- That does not imply (necessarily) that being a healthcare worker is protective against *any* MERS infection

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
  - Death: Initiation of Treatment
  - Cure: Initiation of Treatment
- In trials, randomization is often taken as the origin, in observational studies we often choose

Immortal Person-Time

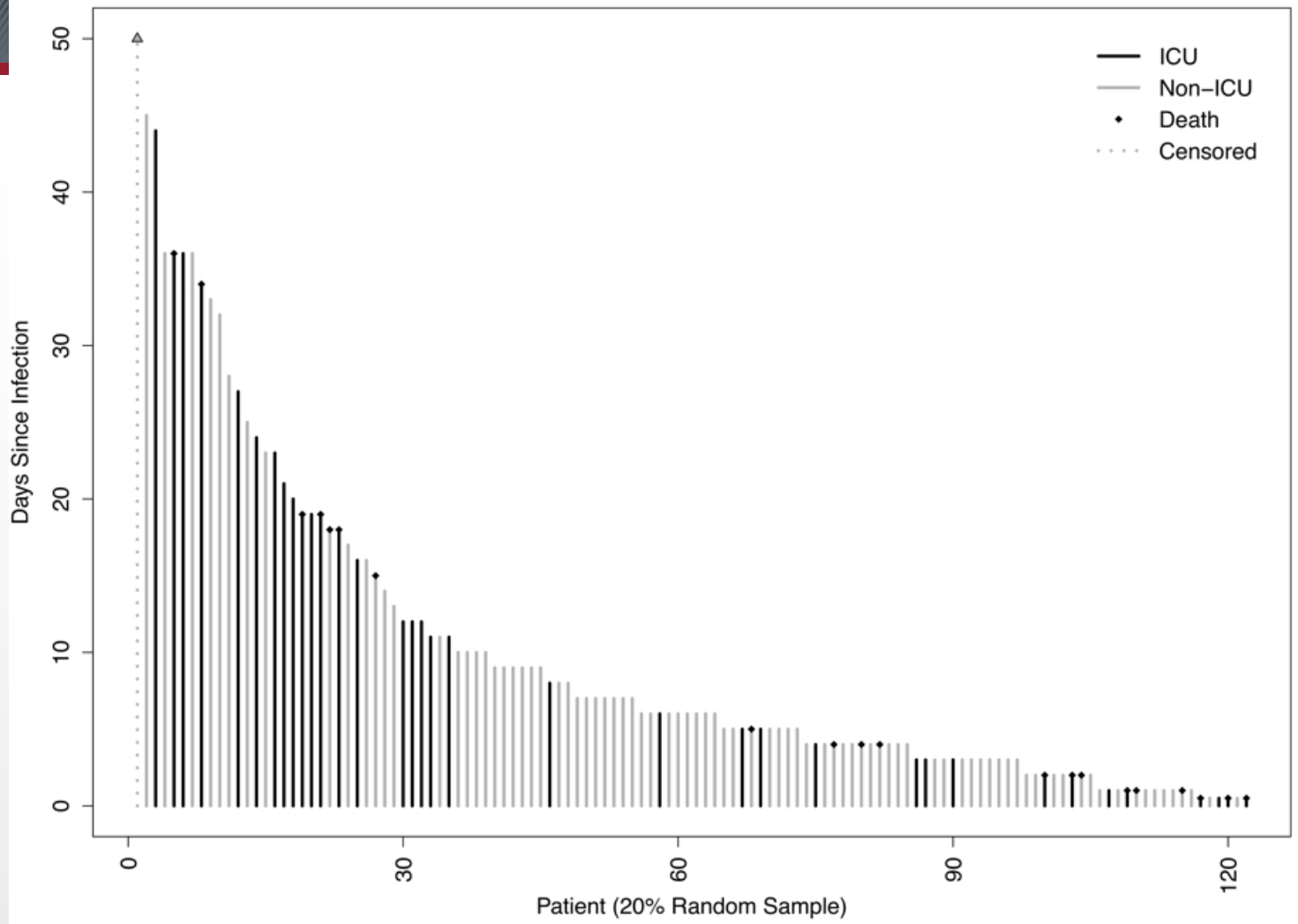
- The origin is the beginning of the time *at risk*
- A subject may spend time in the study not at risk
- It's inappropriate to include this when studying survival time
- Examples:
  - A workplace cohort where you enroll workers after 6 months of employment to study a disease outcome
    - By definition 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)

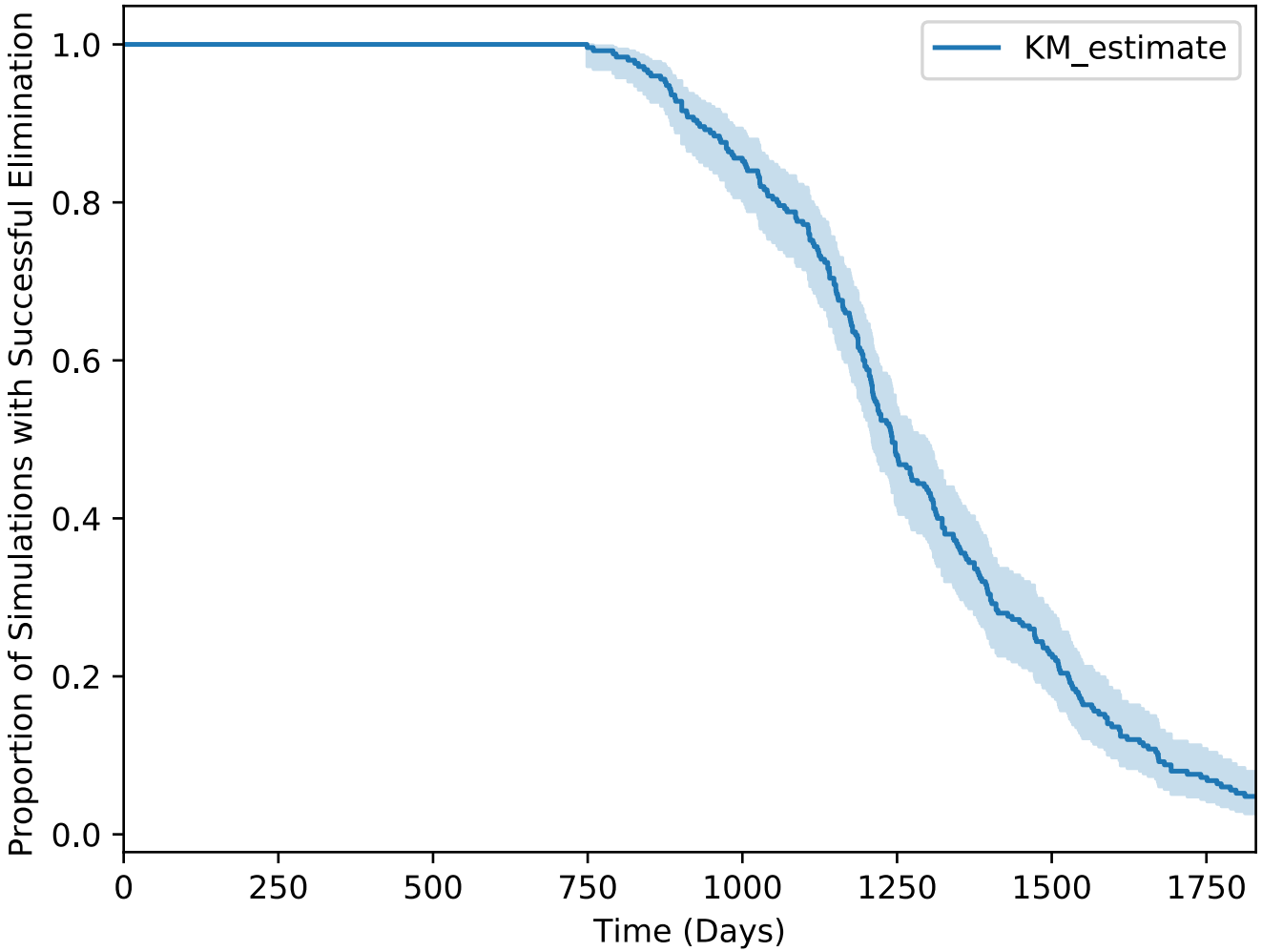
Censoring

- 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


ICU
Non–ICU
Death
Censored
Days Since Infection
0
10
20
30
40
50
0
30
60
90
120
Patient (20% Random Sample)

Survival Function

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


KM_estimate
Proportion of Simulations with Successful Elimination
1.0
0.8
0.6
0.4
0.2
0.0
0
250
500
750
1000
1250
1500
1750
Time (Days)

Hazard

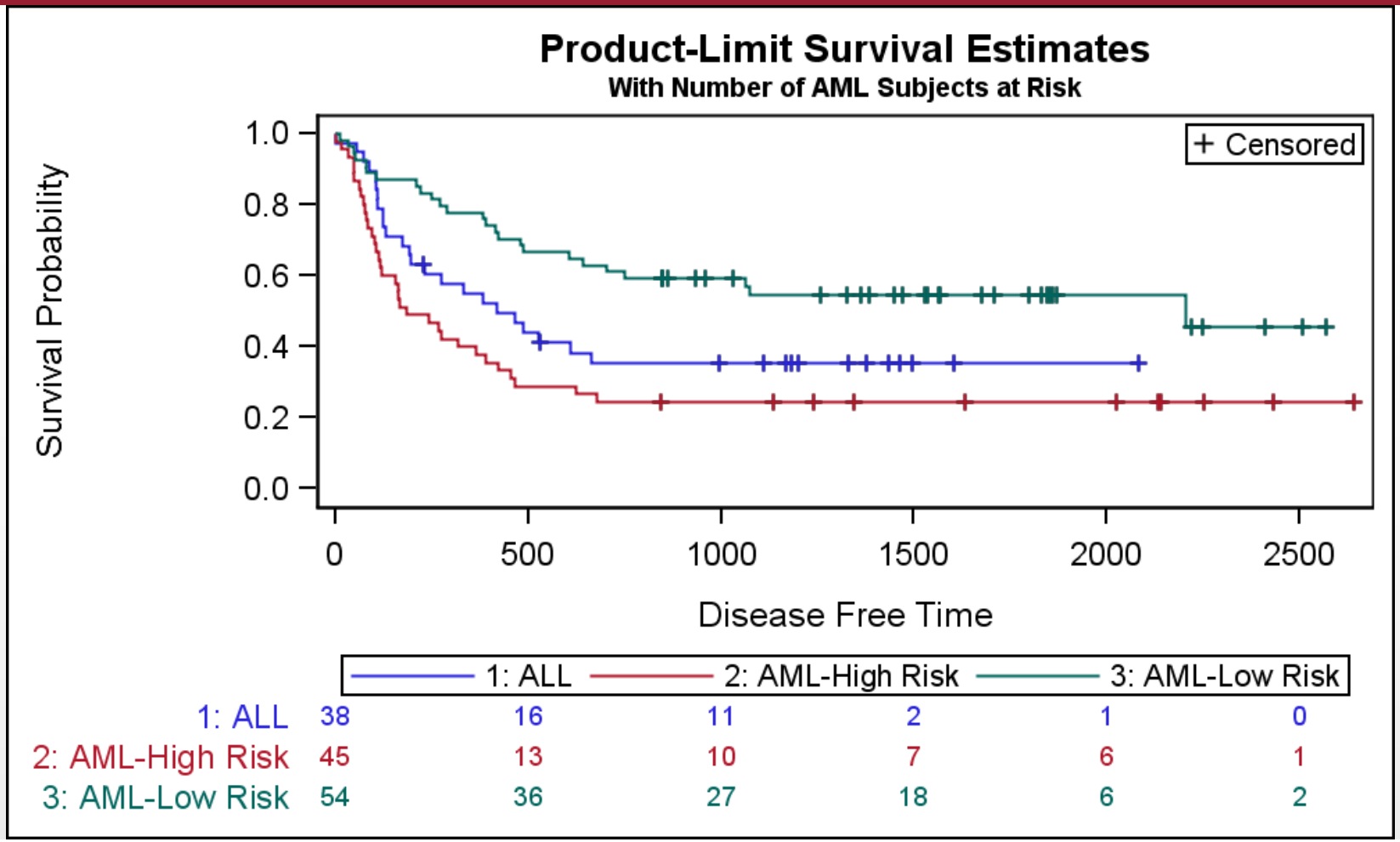
- A function of time
- $h(t)=\lim_{\Delta t \to 0} P(t \leq T < t+\Delta t)$
- Super clear, right?
- Relating the two:
  - The hazard is the slope of the survival function at *t*, divided by S(*t*)
- A constant hazard results in an exponentially distributed survival function

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
  - Survival analysis is a corner of epidemiology where everyone loves non-parametric approaches
- Often can only compare stratified groups
  - There are ways of controlling for many variables using inverse probability weights

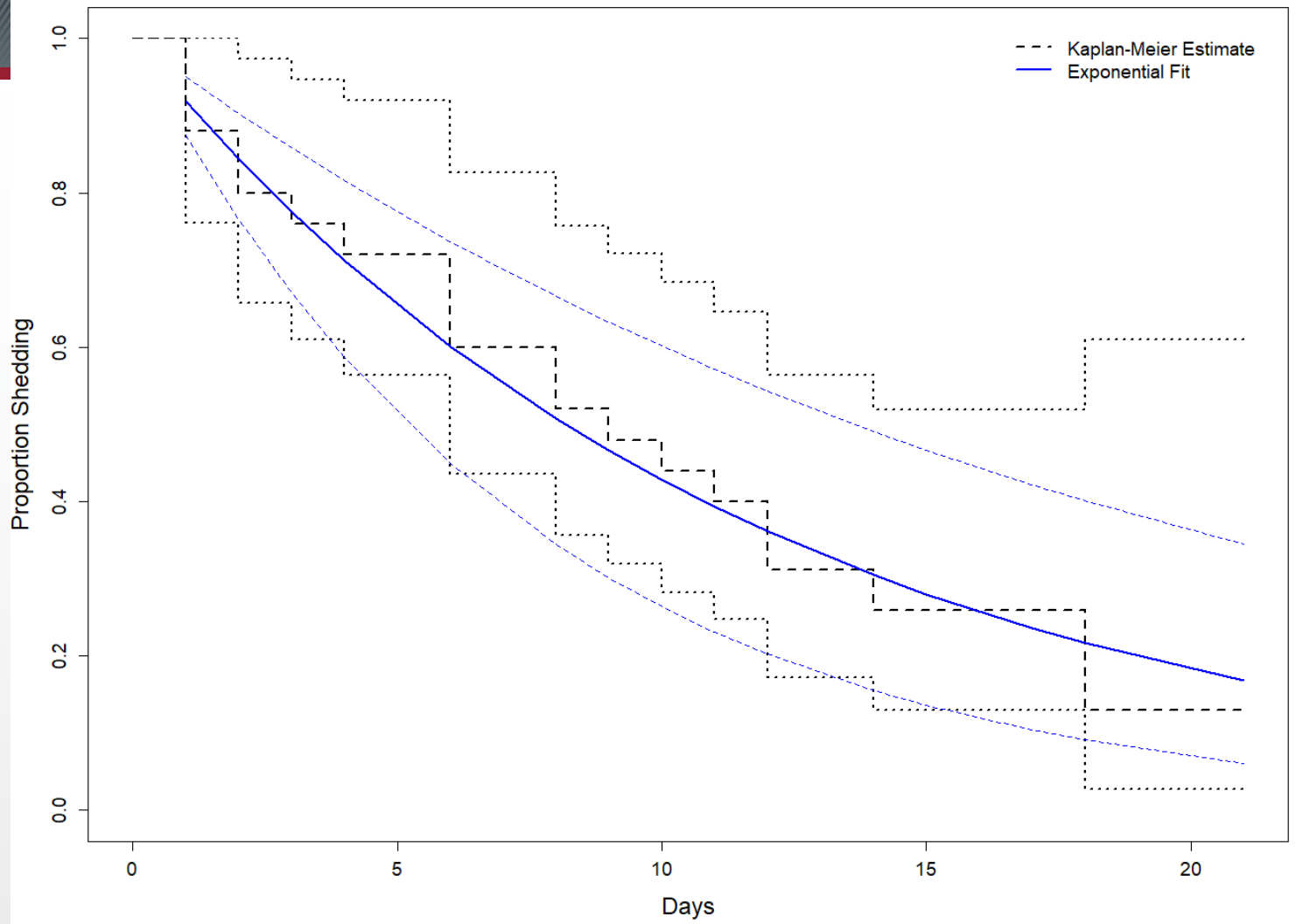
Crude Estimation of the Survival Function

- S(t) = # Participants with T > t / N
- This crude method has a problem – it throws away the partial information from censored individuals
- The K-M method gets around this using the "actuarial" or "life table" method, which estimates a series of conditional probabilities
  - If you're really interested in how to calculate this by hand, I will send you the appropriate reference


Product-Limit Survival Estimates
With Number of AML Subjects at Risk
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
+ Censored
0
500
1000
1500
2000
2500
Disease Free Time
1: ALL
2: AML-High Risk
3: AML-Low Risk
1: ALL 38 16 11 2 1 0
2: AML-High Risk 45 13 10 7 6 1
3: AML-Low Risk 54 36 27 18 6 2

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


Kaplan-Meier Estimate
Exponential Fit
Proportion Shedding
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Days

Problems…

- These estimate relative *time* not relative *hazard*
- Not comparable to a RR
- Indeed, they are directly opposite to RR and related measures in terms of their interpretation
  - Estimates below 1 are bad, above 1 are good
- Exponential and Weibull distributions have transformations, more complex distributions do not

Hazard Ratios

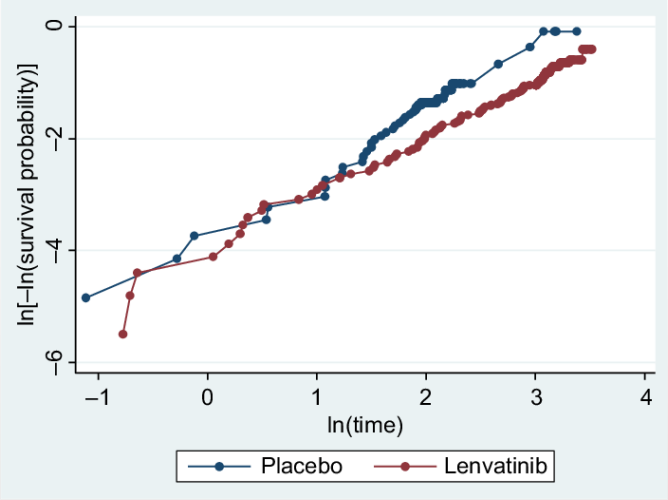
- $HR = h_1(t) / h_0(t)$
- $\exp(\boldsymbol{\beta}) = h_1(t) / h_0(t)$

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
  - This is a stronger assumption than people think it is

Check Your Assumptions

- log-log S(t) over time should be parallel
- Schoenfeld residuals should be relatively constant over time



Dealing with Rate Data

- The *actual* reason to use Poisson regression
- Works like any other regression model we've talked about so far
- Count data works out of the box
- Rate data requires an *offset* term
  - Sometimes called an *exposure* term, which is very confusing
  - This is essentially accounting for how much time these counts had the opportunity to arise in