Topics in Biostatistics Applicable to Cancer Research – Topics in Survival Analysis

Zhigang Zhang

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

New York, NY

April 18, 2014

Outline

I. Introduction

II. Fundamental Tools: Kaplan-Meier, Log-Rank, Cox Model

III. (Left) Truncation

IV. Violation of Proportional Hazards Assumption

V. Competing (and Semi-Competing) Risks Analysis

VI. More Complicated Data Structures

VII. Softwares

VIII. Homework Questions

I. Introduction

- In cancer research (and most medical/clinical researches), the "time to event", a nonnegative random variable, is of primary interest.
  - Overall survival, local recurrence free survival, distant metastasis free survival, progression free survival, recurrence free survival, disease specific survival, etc.
  - Physicians often call them "clinical endpoints".
  - Usually the calculation starts at date of diagnosis, starting date of treatment, end date of treatment, or date of surgery, etc.
  - Almost always, they are subject to ***censoring***: withdrawal, lost to follow-up, event-free at last follow-up, etc.
  - Truncation may sometimes occur as well.

- ***Example***: Right Censored Overall Survival Data.

ID	OSTime	OSStatus	Gender	Age	Treatment	Stage
1	12.4	1	F	65	Surgery	IV
2	5.6	1	F	74	Surgery	III
3	8.7	0	M	52	Radiation	II
...	...	...	...	...	...	

  - "OSStatus" indicates whether this patient was dead at the last follow-up, with 1 indicating death and 0 otherwise.
  - Let T be the true, sometimes unobserved survival time, C the potential censoring time, what we observe is $X = \min(T, C)$ and $\Delta = I(T \leq C)$ (later we will introduce more complicated data structure).
  - We want to know everything about T.

- Some mathematical statistics concepts.
  - $F(t) := P(T \leq t)$ is the cumulative distribution function (CDF), $f(t)$ is the probability density function (PDF) or probability mass function (PMF). To simplify the situation let us assume from now on T is absolutely continuous with PDF $f(t)$.
  - $S(t) := P(T > t) = 1 - F(t)$ is called the survival function of T.
  - A unique function often used in survival analysis is the hazard function $\lambda(t) := \lim_{h \to 0} \frac{P(t<T \leq t+h | T>t)}{h}$. The cumulative hazard function is defined as $\Lambda(t) = \int_0^t \lambda(s)\, ds$.
  - It can be shown that $S(t) = \exp(-\Lambda(t))$ and $\lambda(t) = f(t)/S(t)$ whenever $S(t) > 0$. Therefore, the hazard function $\lambda(t)$ is as good as the PDF in determining the probabilistic behavior of

the random variable T.

* Why are physicians interested in the hazard function? Because it provides some sort of "conditional probability". Rigorously speaking, it is a conditional density. However, since the condition is not fixed, it does not integrate to 1. As a matter of fact, $\Lambda(\infty) = \infty$ if T is a proper random variable, i.e., $P(T = \infty) = 0$, meaning that there is no "cure" fraction.

* Statistical inference is typically made by comparing survival or hazard functions of different treatment groups, or via regression analysis (in general, one needs to model the hazard function).

II. Fundamental Tools

1. Kaplan-Meier (KM) Estimator of $S(t)$

- Without censoring, we can use $1-$ empirical CDF $= \sum_{i=1}^{n} I(T_i > t)/n$.
- When (right) censoring is present, the size of the "risk set" (i.e., how many subjects are still under observation and may potentially produce an event) may decrease without failures.
- Suppose there are n subjects in the sample. Let $t_1 < t_2 < ... < t_k$ $(k \leq n)$ be the distinct, ordered failure time points (i.e., not censored). Estimate $\lambda(t_i), i = 1, ..., k$ by $\hat{\lambda}(t_i) = d_i/n_i$, where d_i is the number of failures at time t_i and n_i is the number of subjects at risk just prior to t_i (i.e., size of the risk set).

- Now we can define the KM estimator $\hat{S}(t) = \prod_{i:t_i \leq t}(1 - \hat{\lambda}(t_i))$, a right-continuous step function (by convention). Note that the KM estimator is undefined after the largest observed failure time. So DO NOT extrapolate!!!
- The KM estimator can also be derived from another two interesting approaches. One is to re-distribute the "mass" of censored observations to the right. The other is to use the inverse-probability-censoring-weighting technique. It is the non-parametric maximum likelihood estimator (NPMLE) for $S(t)$, and the non-parametric likelihood is written as $\prod_{i=1}^{n} S(X_i)^{1-\Delta_i} f(X_i)^{\Delta_i}$.

2. Log-Rank Test

- Sometimes we may want to compare the survival functions derived

from various groups. That is, we wish to test (suppose there are J groups) the null hypothesis $S_1(t) = ... = S_J(t)$ for all t.

- To illustrate the approach, we consider $J = 2$ (e.g., treatment vs. placebo). The general case $J > 2$ can be tackled in a similar way.
- Combine the two samples and let $t_1 < t_2 < ... < t_k$ be the distinct, ordered observed failure time points from the combined sample. Let d_i and n_i denote the number of observed failures and number at risk, respectively, for the combined group at time t_i, $i = 1, ..., k$. Also let d_{1i} and n_{1i} denote the number of observed failures and number at risk, respectively, for the first group at time $t_i, i = 1, ..., k$. Under the null hypothesis, d_{1i}/n_{1i} and d_i/n_i should be close to each other.
- Quantitatively, given d_i, n_i and n_{1i}, d_{1i} has a hypergeometric dis-

tribution with mean $e_{1i} = d_i n_{1i}/n_i$ and variance $v_{1i} = d_i n_i (n_i - n_{1i})(n_i - d_i)/n_i^2/(n_i - 1)$. Thus we may define the test statistic as

$$W = \sum_{i=1}^{k} \Big(d_{1i} - e_{1i} \Big) \Big(\sum_{i=1}^{k} v_{1i} \Big)^{-1/2} .$$

It can be shown that W approximately follows a standard normal distribution when sample sizes are large. Note that the test statistic W is essentially a quantity of standardized observed-minus-conditionally-expected number of events.

- The test statistic (numerator) W has another more intuitive expression:

$$W = c \int_0^{\infty} \frac{n_1(t) n_2(t)}{n(t)} \Big\{ d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t) \Big\} .$$

- The log-rank test is not sensitive to survival functions that cross.

3. The Proportional Hazards (PH) Model

- What if we want a regression model (suppose we have observed some covariates Z)? $T = \beta' Z + \epsilon$ does not sound reasonable. Moreover, T is not always observed due to censoring.
- It appears that working on the hazard function will again bring us technical convenience. The proportional hazards model (often called the Cox model) specifies that

$$\lambda_z(t) := \lambda(t|Z = z) = \lambda_0(t)e^{\beta z} \ ,$$

  where $\lambda_0(t)$ is an unspecified baseline hazard function (that's why it is called "semi-parametric").
- As a simple example, let Z be binary, 0 vs 1. Then we have $\lambda_1(t)/\lambda_0(t) = e^{\beta}$, where e^{β} is the so-called "hazard ratio".

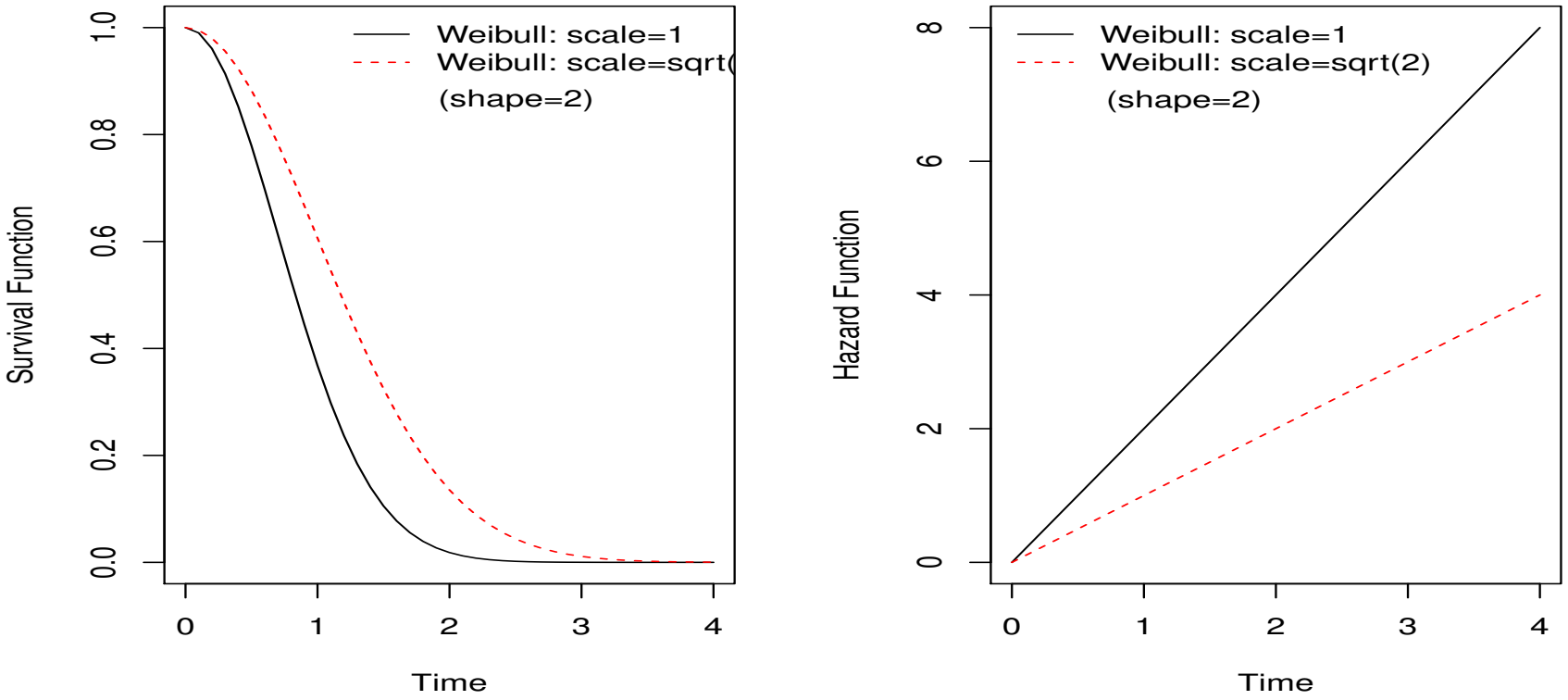


Figure 1: A two-sample example: Weibull distributions with PH.

- The PH assumption requires that the two hazard functions are proportional to each other (no cross-over). The two corresponding survival functions cannot have cross-over either.
- Theoretically the PH model is very suitable for right-censored data,

and has inspired some important technical tools (e.g., partial likelihood, martingale and counting process theory) for statistical inference.

- Estimation of β and Λ_0 can be done in the following way. Recall that we observe $\{X_i, \Delta_i, Z_i\}, i = 1, ..., n$. Define the counting process $N_i(t) = \Delta_i I(X_i \leq t)$ and the risk process $Y_i(t) = I(X_i \geq t)$. Let $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$ (this function will not be used here but later). The partial likelihood for β is

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp\{\beta' Z_i\}}{\sum_{j \in \mathcal{R}_i} \exp\{\beta' Z_j\}} \right\}^{\Delta_i}.$$

In the above, $\mathcal{R}_i$ is the risk set at time X_i, i.e., $\mathcal{R}_i := \{j : X_j \geq X_i\}$. Thus, the denominator (i.e., the summation) can also be written as $\sum_{j:X_j \geq X_i} \exp\{\beta' Z_j\}$ or $\sum_{j=1}^n Y_j(X_i) \exp\{\beta' Z_j\}$. Taking

the logarithm, we have $l(\beta) := \log\{L(\beta)\} = \sum_{i=1}^{n} \Delta_i \Big\{ \beta' Z_i - \log\Big[\sum_{j=1}^{n} Y_j(X_i) \exp\{\beta' Z_j\} \Big] \Big\}$. Taking derivative with respect to β yields $U(\beta) := \partial l(\beta)/\partial\beta = \sum_{i=1}^{n} \Delta_i \Big\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(X_i) Z_j \exp\{\beta' Z_j\}}{\sum_{j=1}^{n} Y_j(X_i) \exp\{\beta' Z_j\}} \Big\}$

$= \sum_{i=1}^{n} \int_0^{\infty} \Big\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(t) Z_j \exp\{\beta' Z_j\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\beta' Z_j\}} \Big\} dN_i(t).$

- The above is essentially an observed-minus-expected value type of estimating equation, which can be solved numerically. The asymptotic properties can be obtained by using empirical process theory, or more simply, martingale theory.
- For the baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$, the Breslow estimator is defined $\hat{\Lambda}_0(t) = \sum_{i: X_i \le t} \frac{\Delta_i}{\sum_{j \in \mathcal{R}_i} \exp\{\hat{\beta}' Z_j\}} =$

$\sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s) \exp\{\hat{\beta}' Z_j\}}$.

- The Cox model is a flexible and efficient semi-parametric model. Many extensions and generalizations exist in the literature.
  - Cox model is rank-invariant.
  - The covariate Z can be time-dependent (i.e, $Z(t)$).
  - The regression parameter β can be time-dependent (i.e., $\beta(t)$), in which case we no longer have a PH model.
  - Frailty terms can be introduced to incorporate random effects from clusters.
  - Stratified Cox models can be used when different groups may have different baseline hazard function.

III. (Left) Truncation

- Roughly speaking, truncation means that the data are not representative. Technically, this means we have a biased sample and need to consider some conditional arguments.
- For example, "time 0" is date of diagnosis and we are examining overall survival. But due to some administrative (or other) reasons, those patients who died within 2 years cannot be observed. In other words, we only observe patients who can survive more than two years.
- Note that truncation is different from censoring. Censoring means that you still get some "incomplete" information. But truncation means that you don't get any information at all, hence bringing in bias if not taken into account. Below we shall see how it changes

the likelihood.

- Now our observation is $\{D_i, X_i, \Delta_i, Z_i\}, i = 1, ..., n$, where D_i denotes the truncation time for each subject (could be same for all i). Then the likelihood is $\prod_{i=1}^{n} S(X_i)^{1-\Delta_i} f(X_i)^{\Delta_i} / S(D_i)$. If D_i is zero for all i, then this likelihood reduces to the non-truncated case we discussed before.
- We can still derive the NPMLE although the algorithm is a little more complex. Also some (often reasonable) additional assumptions on D_i may enhance the efficiency.
- Truncation does not happen as often as censoring, but results could be misleading if not taken into account.

IV. Violation of PH Assumption

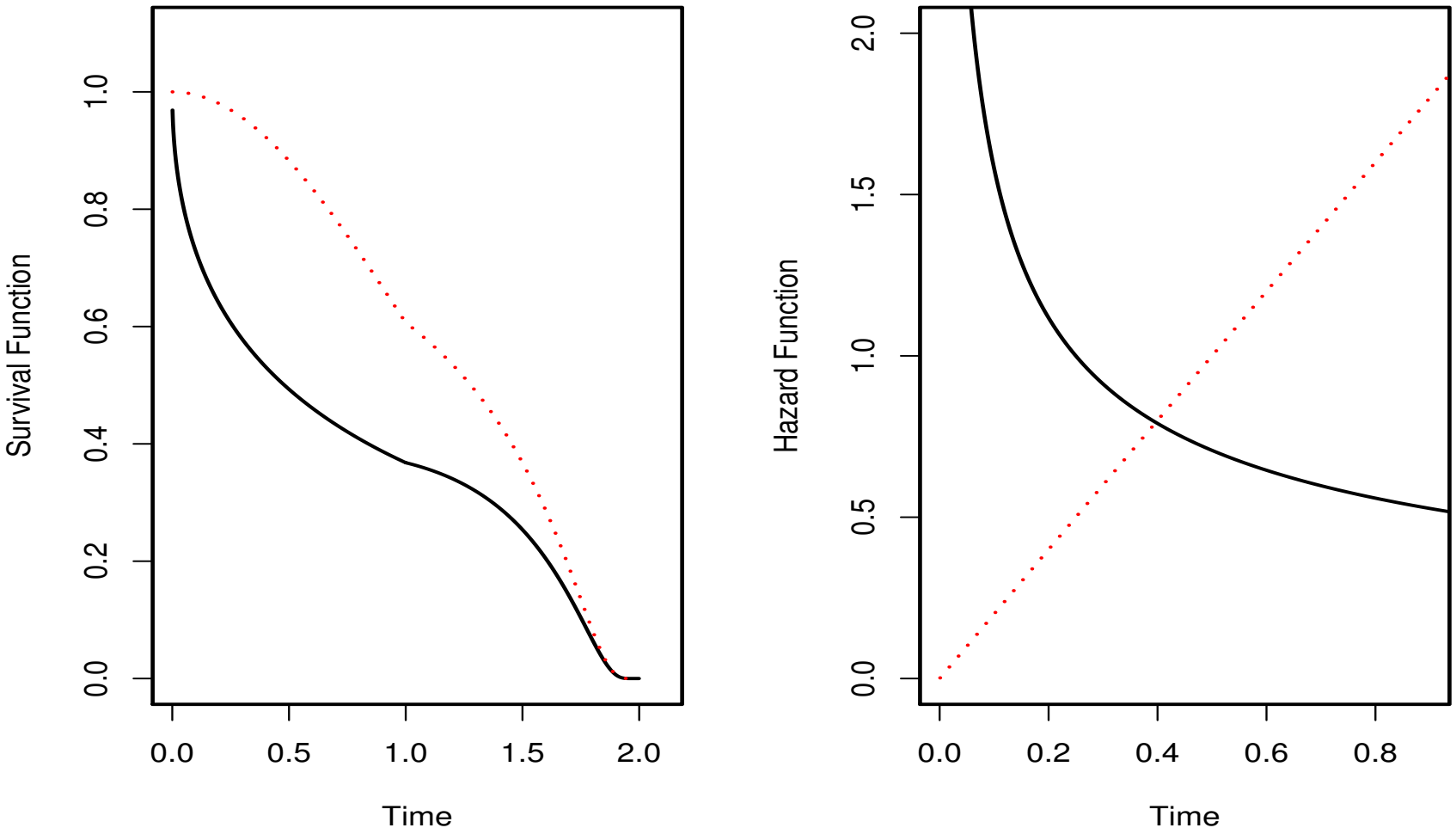


Figure 2: Hazard functions cross but not survival.

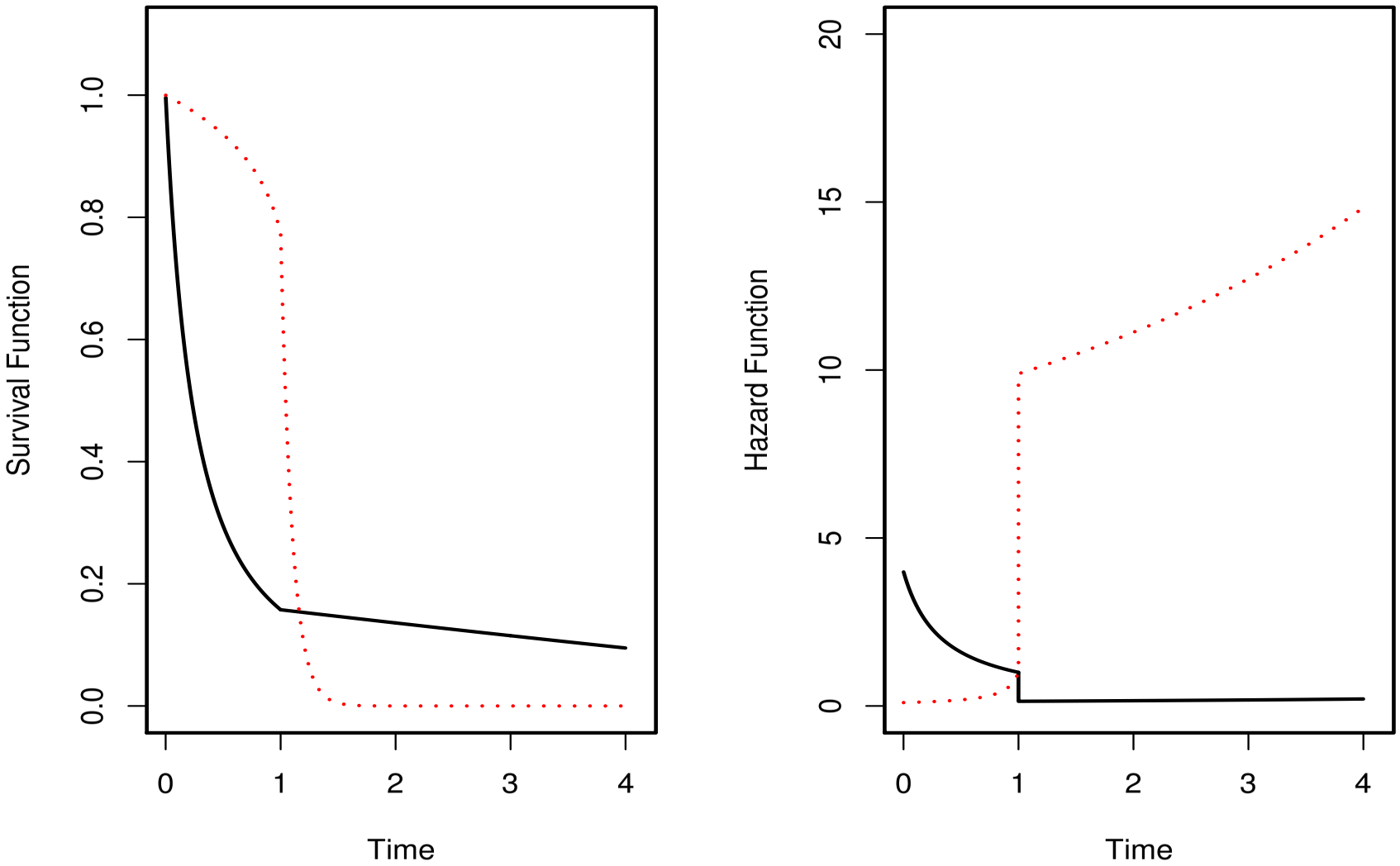


Figure 3: Both survival and hazard functions cross.

- Use the PH model anyway, which may not be too bad, if...
  - There is no cross-over, although hazards are not perfectly proportional.
  - Only interested in showing superiority, and quantification is a secondary objective.
  - Wish to approximately assess the hazard ratio (a crude average might be ok).
- If (say, a two-sample case) the survival curves cross, then we may have to abandon the PH model and try to find the answers directly from the Kaplan-Meier estimates.
- But if we wish to more precisely quantify the difference and the PH assumption fails, then we need to be more cautious. In such situations, the first question we need to ask is: **What do we want?**

We might be interested in:

- Early survival (say, 1-yr OS rate).
- Late survival (say, 10-yr OS rate).
- Mean/Median survival.
- The entire dynamic pattern.
- Something else – not hazard but other functions?

- Different objectives will lead to different solutions. The bottom line is: When the PH assumption fails, things are usually much more complicated, so we must be clear about our research question.

Solutions

- Stratified log-rank tests: Only works for categorical variables.

- Cox models with time-dependent coefficients:

$$\lambda_z(t) = \lambda_0(t)e^{\beta(t)z} \ .$$

We observe that (still using the simple case that Z is binary).

$$\lambda_1(t)/\lambda_0(t) = e^{\beta(t)} \ ,$$

where the hazard ratio is $e^{\beta(t)}$ that depends on time. Inference procedures are in general complicated and may require smoothing.

References: Peng & Huang (*Biometrika*, 2007); Martinussen & Scheike (Springer, 2006); Tian, Zucker & Wei (*JASA*, 2005).

- Alternative regression models:
  - Additive Hazards Model: $\lambda_z(t) = \lambda_0(t) + \beta z$. But it does not allow cross-over for either hazard or survival functions.
  - Accelerated Failure Time Model (AFTM): $\log(T) = \beta z + \epsilon$. To

be more transparent, let's again consider the two sample case ($Z = 0$ vs 1). The AFTM can be rewritten as $S_1(t) = S_0(t/e^{\beta})$ or $\lambda_1(t) = \lambda_0(t/e^{\beta})/e^{\beta}$. Thus AFTM allows for crossover for hazard but not survival functions.

* Accelerated Hazards Model: $\lambda_z(t) = \lambda_0(te^{\beta z})$. This model allows for crossover for both hazard and survival functions.
* Proportional Odds Model: $\log\{F_z(t)/S_z(t)\} = h(t) + \beta z$. This model requires "converging hazards".
* Linear Transformation Models: $h(T) = \beta z + \epsilon$. Similar to the AFTM.

- Quantile Regression: Instead of modelling the hazard function, we may focus on the (conditional) quantiles of the survival time:

$$Q_z(\tau) = \beta(\tau)Z\,,$$

where $Q_z(\tau) = \inf\{t : S_z(t) > 1 - \tau\}$. Quite challenging in both theory and computation under censoring.

References: Portnoy (*JASA*, 2003); Peng & Huang (*JASA*, 2008); Huang (*Ann. Stat.*, 2010).

- Restricted Mean Survival Time (RMST): Recall that $E(T) = \int_0^\infty S(t)dt$. But since $S(t)$ is rarely estimable near the end due to censoring, we may define the RMST as $E_\tau(T) = \int_0^\tau S(t)dt$ with some properly chosen τ.

  Equivalently, we can define the RMST in the following way: Let $Y = \min(T, \tau)$, then RMST is $E(Y)$. Caution: $E(Y) = E(T \wedge \tau) \neq E(T|T \leq \tau)$.

- Stratification: If the PH assumption fails on $Z_1 = 0, ..., J$ (here we partition the covariate $Z := (Z_1, Z_2)$, we may use Z_1 as a stratifier

rather than a covariate:

$$\lambda_z(t) = \lambda_{0z_1}(t)e^{\beta z_2} \ , z_1 = 0, ..., J \ ,$$

where Z_2 denotes other adjustment covariates. Note that now the quantity of interest is no longer β, but $\lambda_{0z_1}(t)$ or the cumulative hazard function $\Lambda_{0z_1}(t) = \int_0^t \lambda_{0z_1}(u)du$.

- Mean Residual Life (MRL) Models: MRL is defined as $m(t) = E(T - t|T > t)$. The ordering assumption on MRL is weaker than that of survival or hazard functions. In other words, it is possible that, even if two survival or hazard functions cross, the two corresponding MRL functions do not!

Restricted Mean Survival Time (RMST)

- $E_\tau(T) = \int_0^\tau S(t)dt$. By selecting small or large τ, we are able to

assess early or late effect.

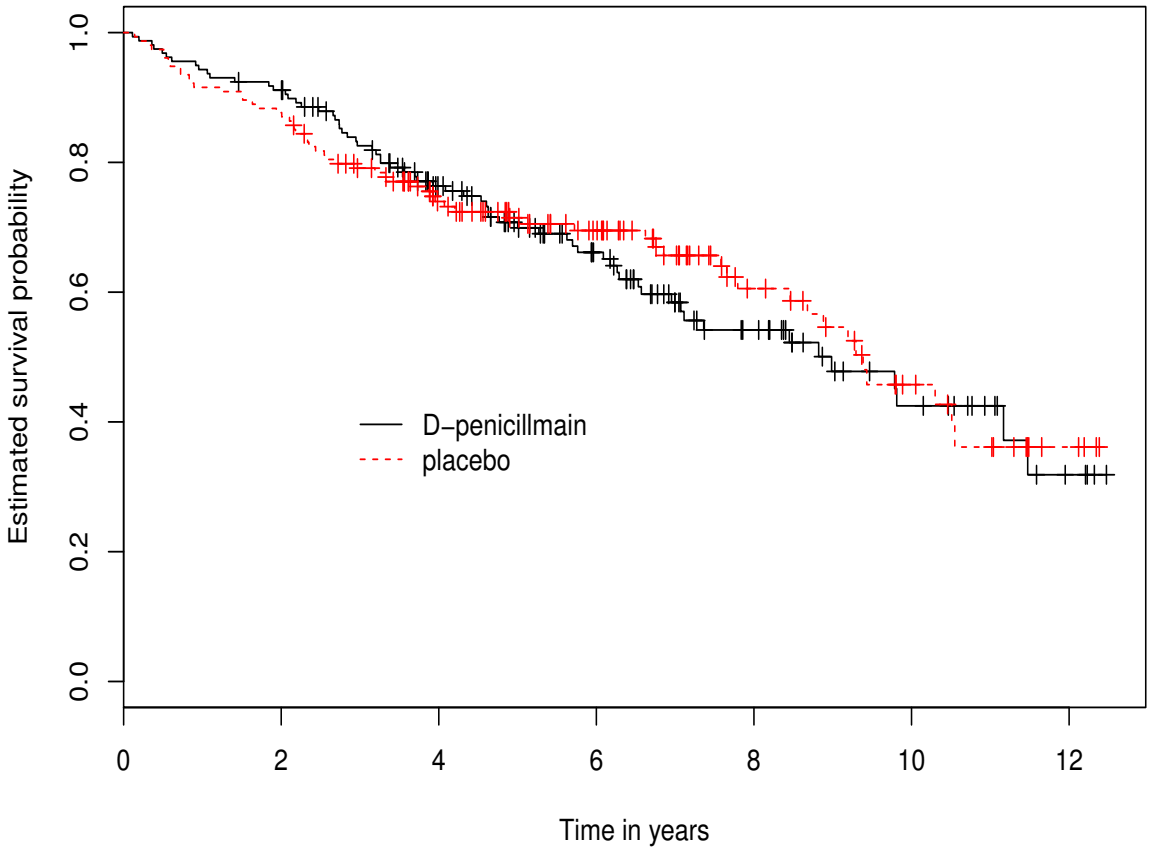


Figure 4: Kaplan-Meier estimates of the two arms.

- As an example, we look at the well-known Mayo primary biliary cirrhosis (PBC) randomized trial. For illustration purpose, let's focus on the treatment: D-penicillamine vs placebo. K-M estimates show that PH assumption may not hold. A log-rank test gives a p-value of 0.75.
- How about RMST? In Karrison (*Controlled Clinical Trials*, 1997), $\tau = 10$ was used. The D-penicillamine group had RMST as 7.13 (SE: 0.28) years while the placebo group had 7.29 (SE: 0.30). There is no significant difference.
- Note that we can choose a small τ, say, 3 years, to assess early RMST.
- It is possible to incorporate covariates for adjustment. There also exist multiple ways for estimating RMST.

- References: Irwin (*Journal of Hygiene*, 1949); Karrison (*JASA*, 1987); Karrison (*Controlled Clinical Trials*, 1997); Zucker (*JASA*, 1998); Anderson et al. (*LIDA*, 2004); Royston & Parmar (*Stat. Med.*, 2011).

Stratification: Cumulative Hazard

- The key idea is to use Z_1 as a stratifier:

$$\lambda_z(t) = \lambda_{0z_1}(t)e^{\beta z_2} \ , z_1 = 0, ..., J \ .$$

- To compare each treatment group to the reference group (say, $Z_1 = 0$), we may define

$$\theta_{z_1}(t) = \Lambda_{0z_1}(t)/\Lambda_{00}(t), z_1 = 0, ..., J \ .$$

  In other words, we are looking at a generalized "hazard ratio".

- Note that by the assumed model, effects of the adjustment covariates would be cancelled out.

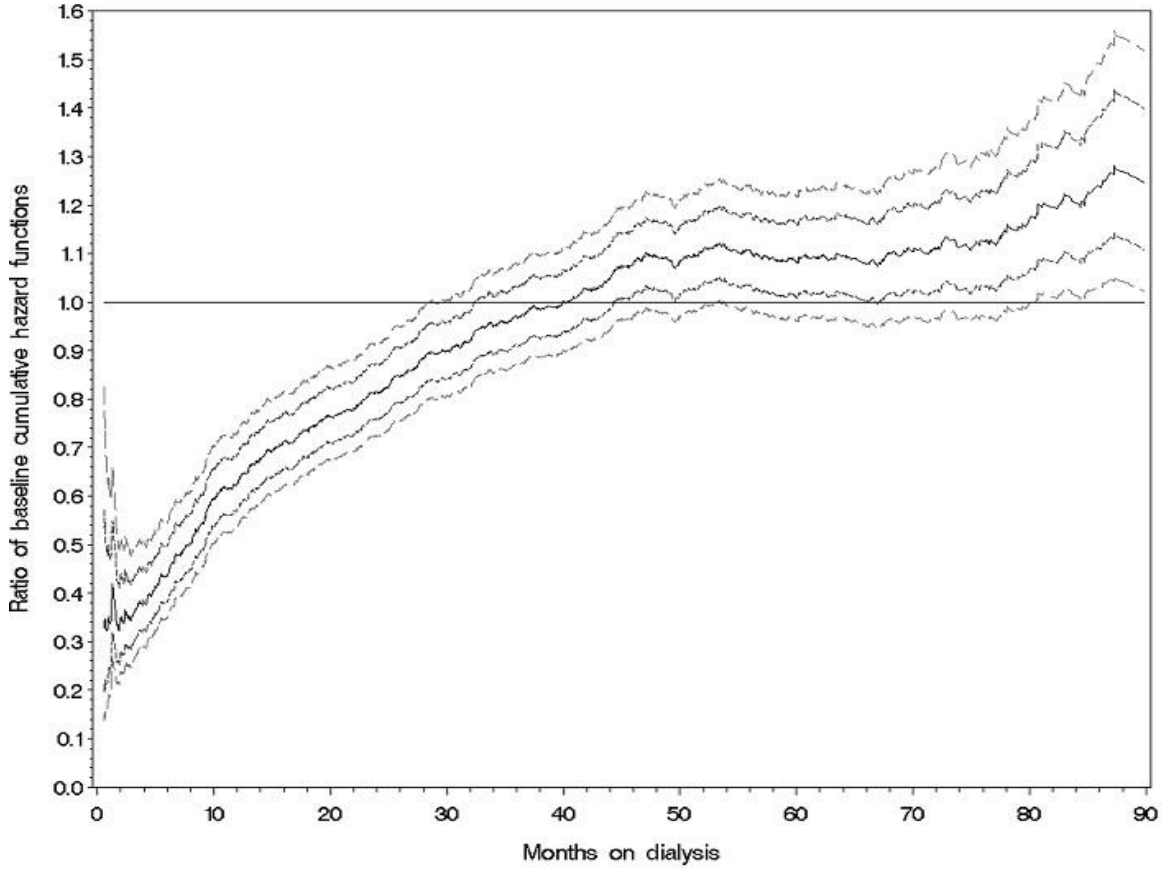


Figure 5: Generalized hazard ratio of PD/HD (Wei & Schaubel, *Biometrics*, 2008).

- Example (Wei & Schaubel, *Biometrics*, 2008): Analysis of the Dialysis Data from the Canadian Organ Replacement Register. To compare end-stage renal disease patients survival on hemodialysis (HD, reference level) and peritoneal dialysis (PD).

Mean Residual Life (MRL) Models

- MRL is defined as $m(t) = E(T - t|T > t)$. In words, it measures the remaining life expectancy of a subject when the subject has survived up to time t.
- When covariates are present, we may use the following transformed MRL model

  $$m(t|z_1, z_2) = E(T-t|T > t; Z_1 = z_1, Z_2 = z_2) = g\{\alpha(t)'z_1+\beta'z_2\} \ ,$$

  $*$ $g(\cdot)$ is a pre-specified nonnegative link function and assumed to

be twice continuously differentiable and strictly increasing;

* $Z_1 = (1, Z_1^{(1)}, \ldots, Z_1^{(q)})'$ and $Z_2 = (Z_2^{(1)}, \ldots, Z_2^{(p)})'$ be a $(q+1) \times 1$ and a $p \times 1$ vector of covariates, respectively;
* $\alpha(t)$ and β are conformal vectors of unknown time-dependent and time-independent regression parameters, respectively.

- Ordering restriction: For example, let Z be binary, we have

$$\lambda_1(t) \leq \lambda_0(t) \Longrightarrow m_1(t) \geq m_0(t)\,,$$

but the inverse is not true.

- Thus MRL has both theoretical and practical advantages. The disadvantage is that, as a mean measure, it may not be robust unless the sample size is large or some "restricted" version of MRL is considered.

- As an example, we consider an MSKCC retrospective trial for glioblastoma multiforme (GBM) patients.

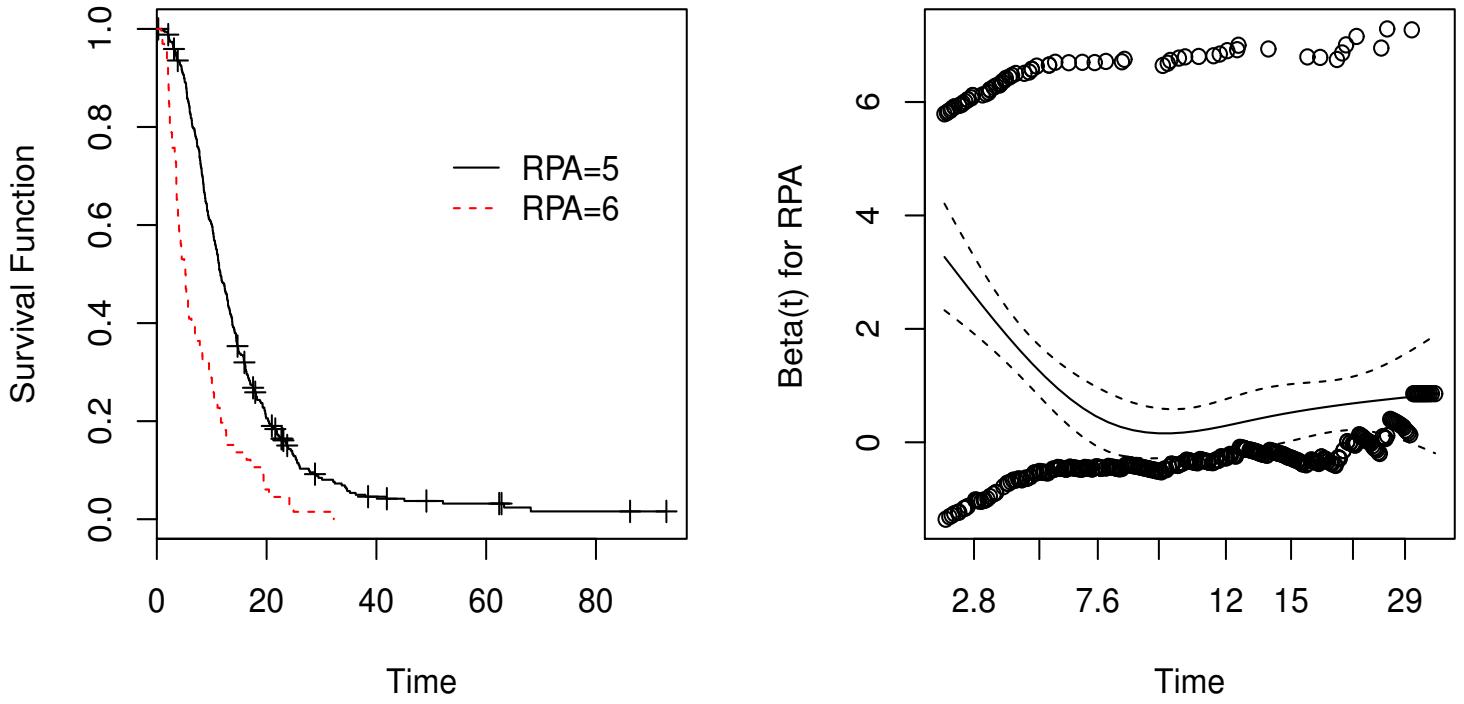


Figure 6: MSKCC GBM trial: effect of RPA on OS.

- We see that the PH assumption is questionable (p=0.0006).
- To illustrate, we consider the following three covariates: Gender, RPA (5 or 6) and extent of surgery (1=biopsy, 2=subtoal resection and 3=gross resection). Let $g(t) = t$ (i.e., additive MRL model) for simplicity (GOF p-value = 0.24).
- Some preliminary analysis shows that extent of surgery has a significant, time-independent effect. So we denote it by Z_2, and RPA (or Gender) by Z_1, respectively. The estimate of β is 4.206 (SE=1.194), implying a significant improvement with more extensive surgery. The estimate of $\alpha(t)$ is shown in the next figure. The effect of RPA is significantly time-varying (p<0.001) while Gender is not (p=0.2)
- References: Chen (*JASA*, 2007); Sun et al. (*Biometrika*, 2012).

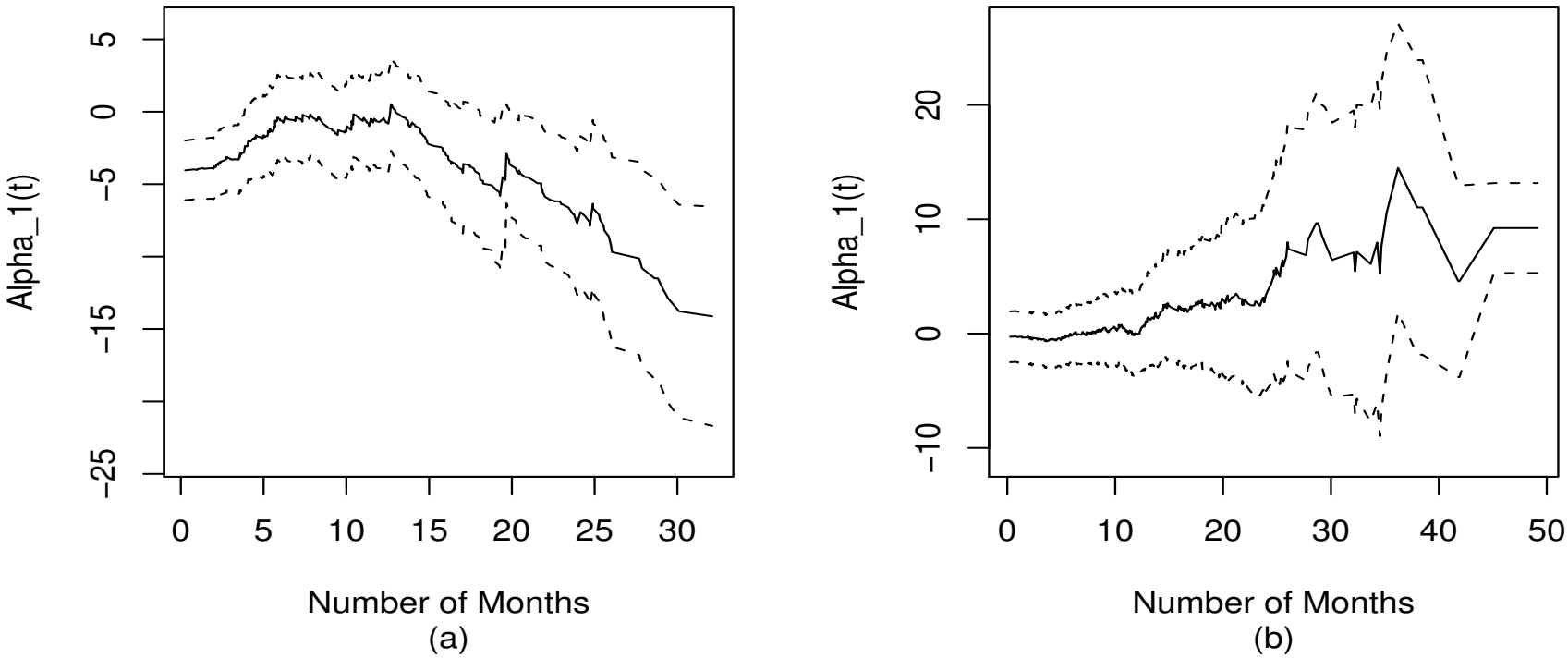


Figure 7: Effect of (a) RPA and (b) Gender on MRL.

Summary

- When the PH assumption fails, we should first examine our analysis objectives.

- The Cox model may still be used as an approximation. E.g., when the hazard functions don't cross.
- If the simple Cox model cannot be used, we may focus on the early effect or late effect or some weighting averages.
- We may switch to the generalized Cox model with time-varying coefficients, and examine the entire dynamic pattern.
- We may choose other semiparametric survival models (focusing on hazard or other functions), in which case results are usually interpreted quite differently.

V. Competing (Semi-Competing) Risks Analysis

- There are some hidden assumptions for the previously mentioned methods (all of them). First, it is assumed that there is no cure-fraction. Second, the censoring has to be independent of the survival time of interest.
- As an example, think about disease-specific survival. To simplify let's consider only two "types" of deaths, one is due to disease (type 1) and the other all other causes (type 2). It is obvious that the two types of deaths censor each other, and this censoring may not be independent (e.g., an overall health status may affect survival time due to to any cause). This is the so-called competing risks scenario.
- Practically we will have the following data:

Example: Competing Risks Survival Data.

ID	OSTime	OSStatus	Gender	Age	Treatment	Stage
1	12.4	1	F	65	Surgery	IV
2	5.6	1	F	74	Surgery	III
3	8.7	0	M	52	Radiation	II
4	19.2	2	M	71	Radiation	III
5	14.5	1	M	66	Surgery	I
...	...	...	...	...	...	

- Note that here "OSStatus" (Δ) may take three values, 0=censoring; 1=death due to cause 1; 2=death due to cause 2.
- We should not change 2 to 0, and thus we cannot use Kaplan-Meier estimator, log-rank test or Cox models.

- More notation: T_1 is time to death due to cause 1, T_2 is time to

death due to cause 2, C is the potential censoring time which is still assumed to be independent of both T_1 and T_2. Now what do we observe?

- First of all, even without censoring, we would only be able to observe EITHER T_1 OR T_2. As a matter of fact, we would get $T := T_1 \wedge T_2$ and $I(T_1 \leq T_2)$. With censoring, we have $X = \min(T, C)$. But we want to draw inference about T_1, not T.
- Unfortunately, it has been shown that $P(T_1 \leq t)$ is not identifiable non-parametrically without making assumptions about the correlation between T_1 and T_2 (which are often not testable). Therefore, we have to turn to the so-called cumulative incidence function (CIF).
- Define the CIF for cause 1 as $F_1(t) := P(T \leq t; \Delta = 1) = P(T_1 \leq$

$t; T_1 \leq T_2)$. $F_1(t)$ is identifiable (up to the last death of type 1). Again, let $Y_i(t) = I(X_i \geq t)$, $\hat{S}(t)$ be the KM estimator for T, $N_{i1} = I(\Delta_i = 1)I(X_i \leq t)$, then

$$\hat{F}_1(t) = \sum_{i:t_i \leq t} \hat{S}(t_{i-1}) \frac{d_{1i}}{n_i} = \int_0^t \hat{S}(u-) \frac{dN_{1i}(u)}{\bar{Y}(u)} \,.$$

where d_{1i} represents the number of deaths due to cause 1, and n_i is size of the risk set as before. An easier way to understand this algorithm is to observe that

$$F_1(0) = 0 \,, \qquad F_1(t) - F_1(t-) = S(t-)\lambda_1(t) \,,$$

with $\lambda_1(t)$ defined as a cause-specific hazard function, i.e., $\lambda_1(t) := \lim_{h \to 0} P(t < T \leq t + h, \Delta = 1 | T > t)/h$.

- Important: $F_1(t)$ is NOT a proper CDF! That is, $F_1(\infty) < 1$ if there exist (at least) two competing risks. But $F_1(t) + F_2(t) = F(t)$

is a proper CDF.

- K-sample comparison tools and regression models have also been developed focusing on $F_1(t)$.

Example: Estimated CIF Curves.

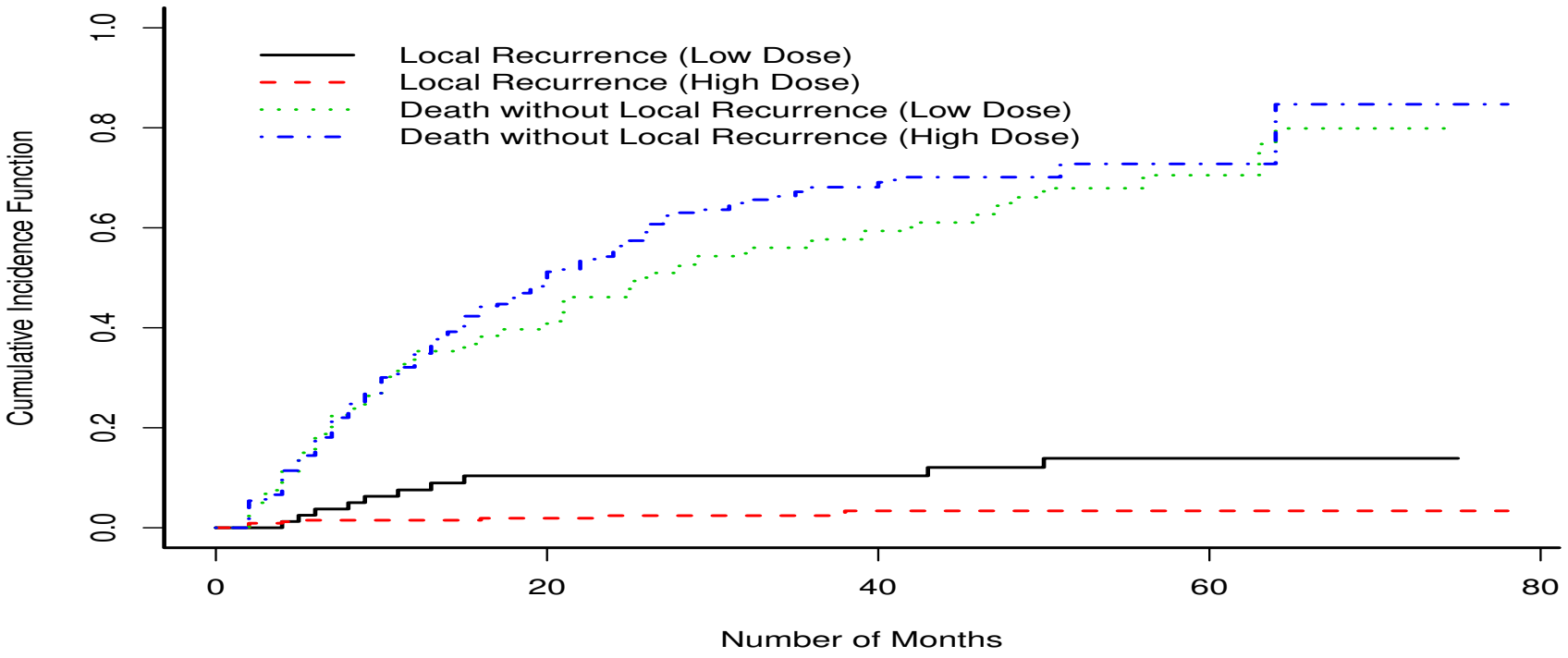


Figure 8: CIF estimates for two competing risks.

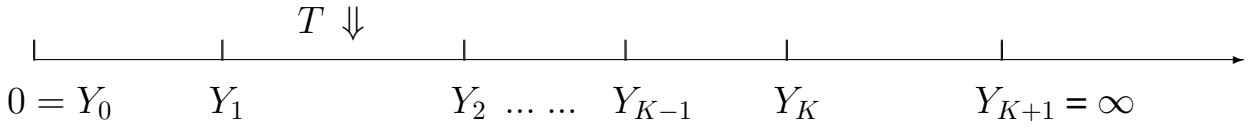
For this example the local recurrence CIFs are not significantly different (p=0.23). But the death without local recurrence CIFs are significantly different (p=0.02).

Semi-Competing Risks Analysis

- Rigorously speaking, the above example is a semi-competing risks data, because local recurrence does not "censor" death.
- Some (quite complicated) methods have been proposed to deal with this situation.
- One way to avoid competing (or semi-competing) risks analysis is to use the last follow-up status, not the death date (vital status).

VI. More Complicated Data Structure

- Suppose patients should pay regular visits to their physicians as post-treatment follow-ups once every 3 months for the first year, every 6 months for the second year, and once a year after that. The clinical endpoint is tumor progression time.
- The observation times are 3, 6, 9, 12, 18, 24, 36, 48, ..., etc (assume that patients comply with the follow-up schedule). Denote the progression time by T, an important fact is that T is not directly observed, but only known to lie between two adjacent follow-up times.



- For example, a patient visited his/her physician at months 3, 6 and

9 and was not found to have progression, but month 12 he/she had, then the true time to progression T fell in the interval $(9, 12]$.

- However, most analyses would use 12 as the time to progression (occasionally using grouped time analysis when all patients made all visits so that any two intervals are either identical or disjoint), rather than acknowledging the fact that $T \in (9, 12]$.
- Such an approximation introduces bias. And if patients skip some of the pre-schedule visits (or don't follow the schedule at all), the interval gets wider and more importantly, overlapping (for instance, one interval is $(5, 12]$ and the other $(9, 17]$). The approximation can be worse, while appropriate methods are needed to handel such data.
- An extreme case is that there is only one follow-up time (call it C).

Then all we know is that either $T \in (0, C]$ or $T \in (C, \infty)$. This is sometimes called the *current status data.*

- In the literature of survival analysis, this is generally referred to as *interval-censored data* (with current status data a special subtype).

Examples (Simulated Data)

- First we look at some simple simulations: T is generated from an exponential distribution with median around 3.5 and/or 2.3. The follow-up times are generated from uniform distributions with an upper bound of 15 (either single or multiple).
- We see that the Kaplan-Meier estimates (using "right-censored" data) are overestimating.
- The log-rank test with "right-censored" data may be misleading.

- The Cox model regression analysis using "right-censored" data results in an HR of 1.07 with a 95% CI as (0.85, 1.35), and a P-value of 0.55. A correct Cox model regression analysis accounting for interval censoring yields an HR of 1.79 with a 95% CI as (1.57, 2.05), and a P-value < 0.0001. The true HR is 1.50.
- As the follow-up visits are more frequent, the intervals tend to be more accurate, and the bias of using right censoring as an approximation gets smaller.
- How about using the "midpoint" imputation? May still have bias.
- The correct way to analyze such data is to use the interval censoring tools.

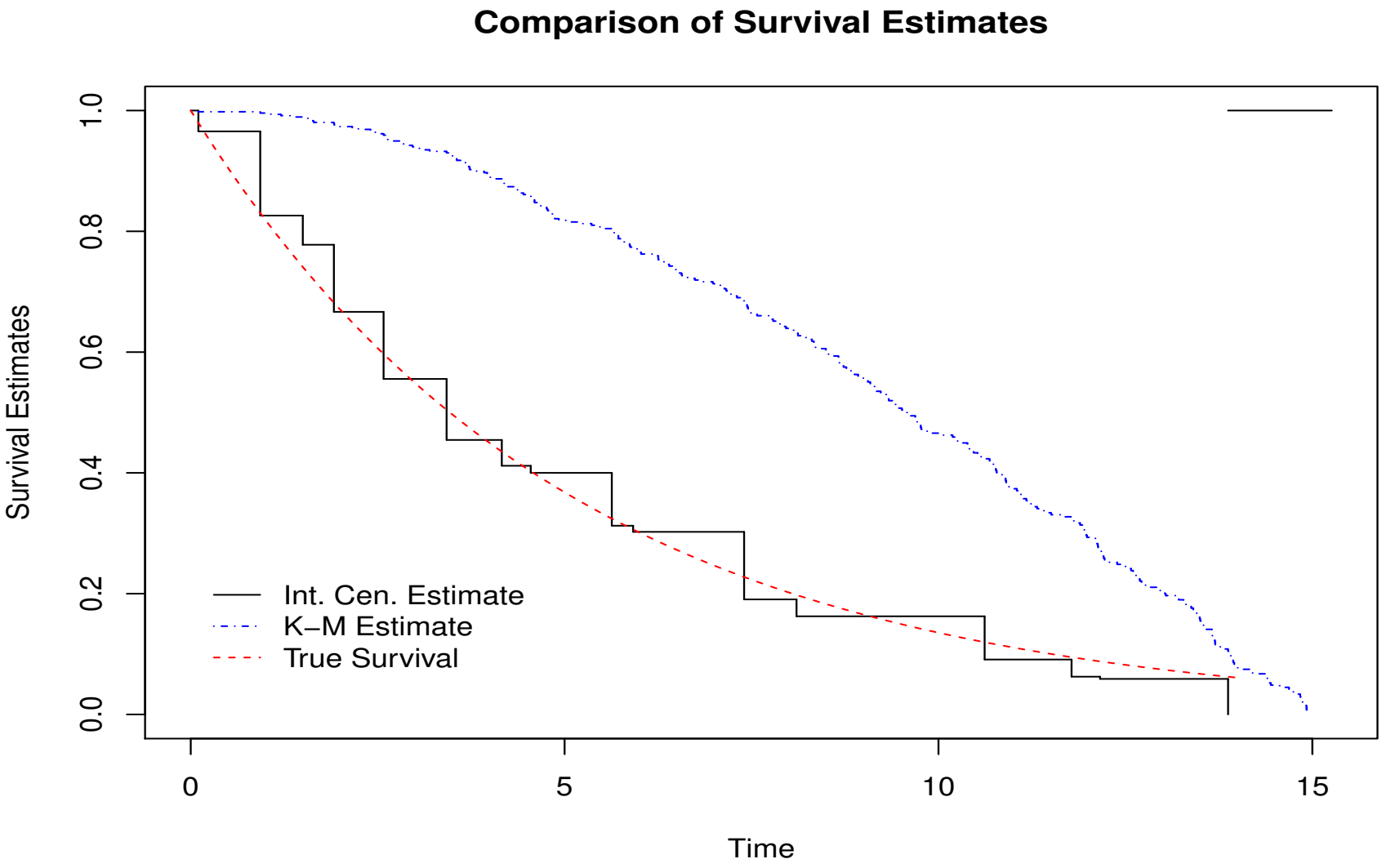


Figure 9: Comparison of Survival Estimates (Single F/U).

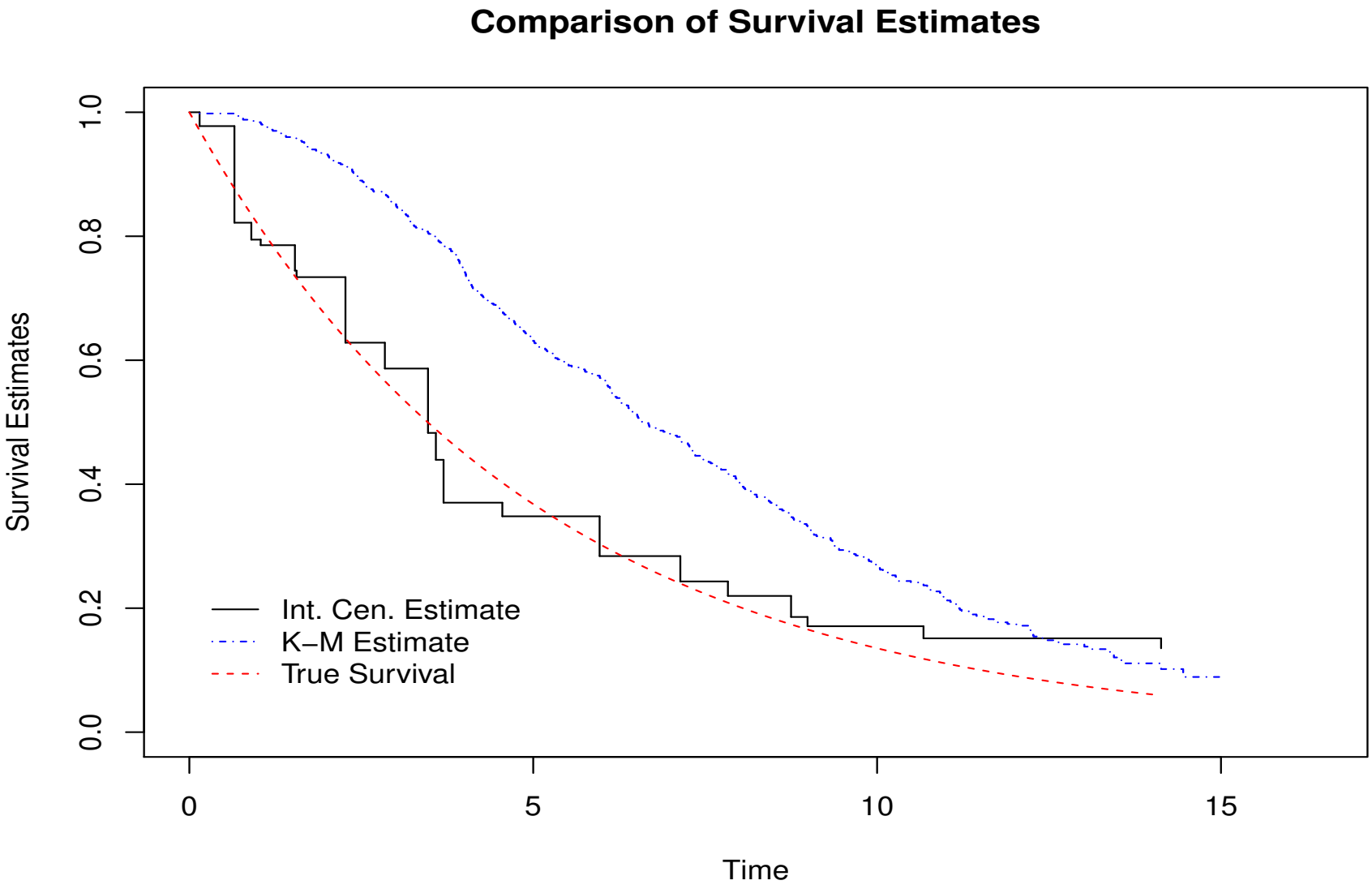


Figure 10: Comparison of Survival Estimates (Multiple F/Us).

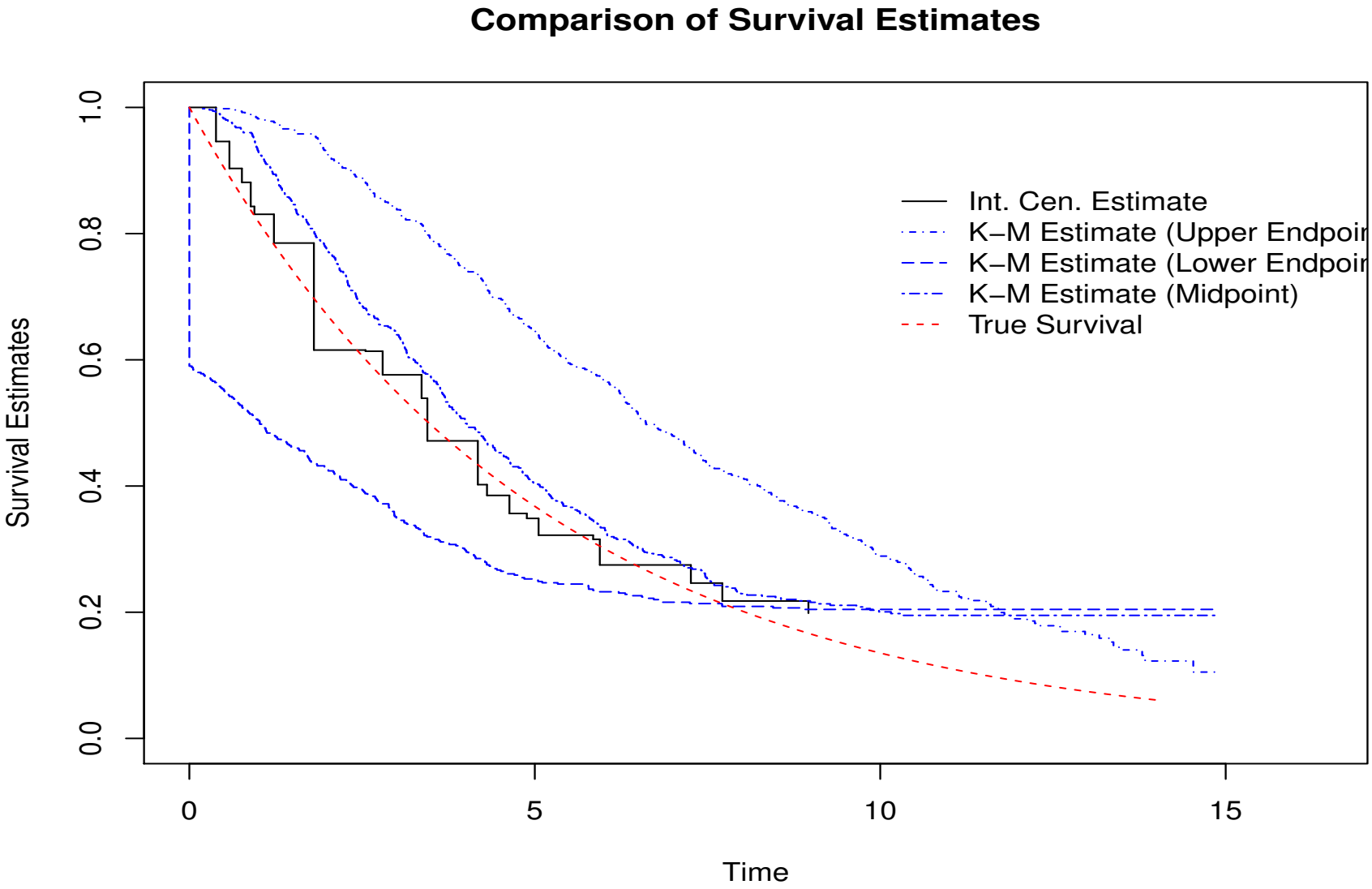


Figure 11: Comparison of Survival Estimates (Different Endpoints).

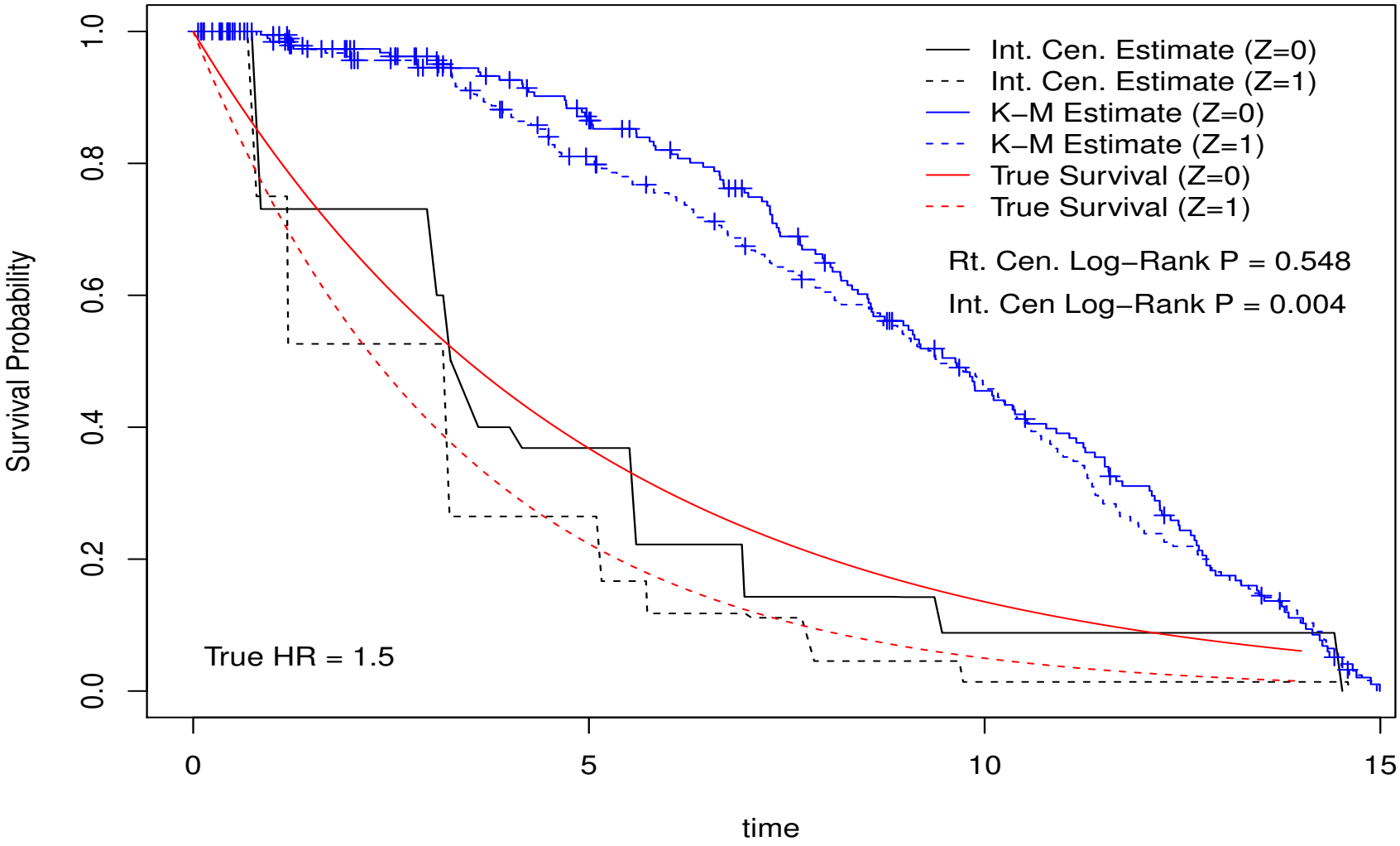


Figure 12: Comparison of Test Results (Multiple F/Us).

Example: Case I Interval Censoring (Lindsay & Ryan, 1994)

ED_{01} data: control/experimental dose group

	Bladder				Lung			
Month	DNT	DWT	SNT	SWT	DNT	DWT	SNT	SWT
4	0	1	0	0	1	0	0	0
5	1	1	0	0	1	1	0	0
.	.	.	.	.	.	.	.	.
32	1	0	0	0	1	0	0	0

	Bladder				Lung			
Month	DNT	DWT	SNT	SWT	DNT	DWT	SNT	SWT
1	3	0	0	0	2	1	0	0
2	0	1	0	0	0	1	0	0
.	.	.	.	.	.	.	.	.
33	0	1	0	0	0	1	0	0

Note: DNT(SNT): death (sacrifice) without tumor; DWT(SWT): death (sacrifice) with tumor.

Example: Case II Interval Censoring (Finkelstein and Wolfe, 1985)

Breast cosmesis data

Radiotherapy Alone			Radio- and Chemotherapy		
$(45, \infty)$	$(25, 37]$	$(37, \infty)$	$(8, 12]$	$(0, 5]$	$(30, 34]$
$(6, 10]$	$(46, \infty)$	$(0, 5]$	$(0, 22]$	$(5, 8]$	$(13, \infty)$
$(0, 7]$	$(26, 40]$	$(18, \infty)$	$(24, 31]$	$(12, 20]$	$(10, 17]$
$(46, \infty)$	$(46, \infty)$	$(24, \infty)$	$(17, 27]$	$(11, \infty)$	$(8, 21]$
...	...	...	...	...	...

Note: A right endpoint ∞ indicates observation is right-censored.

VII. Softwares (R Packages and Codes...)

Right-Censored Data

- Surv, survfit, survdiff and coxph from library **survival**.
  - survfit(Surv(OSTime, OSStatus)∼Treatment, data=YourDataName)
  - survfit(Surv(OSTime, OSStatus)∼1, data=YourDataName)
  - survdiff(Surv(OSTime, OSStatus)∼Treatment, data=YourDataName)
  - survdiff(Surv(OSTime, OSStatus)∼Treatment+strata(Stage), data=YourDataName)
  - coxph(Surv(OSTime, OSStatus)∼Treatment+Stage+Gender+Age, data=YourDataName)
  - coxph(Surv(OSTime, OSStatus)∼factor(Treatment)+Stage+Gender+Age, data=YourDataName)
- cuminc and crr from library **cmprsk**
  - cuminc(YourDataName$OSTime, YourDataName$OSStatus, YourDataName$Treatment)
  - crr(YourDataName$OSTime, YourDataName$OSStatus, cbind(YourDataName$Treatment, YourDataName$Age))

Interval-Censored Data

- icfit from the R packages interval and EMICM from the package Icens both compute the NPMLE of the survival function of a random variable that is interval-censored.
- ictest from the package interval performs the weighted log-rank tests for interval-censored data.
- intcox from the package intcox performs the Cox regression anlaysis for interval-censored data (without variance estimates). survreg from the package survival performs parametric regression using the accelerated failure time model.
- Other R packages: Epi, MLEcens, smoothSurv, dblcens.
- SAS macros: EMICM, ICSTEST, ICE. The SAS procedure LIFEREG also performs parametric regression using the accelerated failure time model.

VIII. Homework Questions

In all the questions below, T and C denote (proper) continuous, nonnegative random variables (i.e., the survival time and the right censoring time, respectively). We always assume that T and C are independent. Also let $X = T \wedge C$ and $\Delta = I(T \leq C)$.

1. The expectation of T is defined as $\int_0^\infty t\, d\, F(t)$ where $F(t)$ is the CDF for T. Let $S(t)$ be the survival function for T. Show that $E(T)$ can also be expressed as $E(T) = \int_0^\infty S(t)dt$ whenever $E(T) < \infty$.

2. Let $G(s)$ be the survival function for C. Show that $E\left[\frac{\Delta X}{G(X)}\right] = E(T)$. (This is the well-known "inverse-probability-censoring-weighting" formula)

3. The linear transformation model $h(T) = \beta Z + \epsilon$ can be viewed as a "generalized" Cox model. In this model $h : \mathcal{R}^+ \to \mathcal{R}$ ($\mathcal{R}$ denotes the real line and $\mathcal{R}^+$ the positive half real line) is an unknown strictly increasing function, β and Z are the regression parameter and covariate, and $\epsilon \in \mathcal{R}$ is a random error with a known, strictly decreasing survival function $D(\epsilon)$. Show that (i) $D^{-1}\{S(t|Z = z)\} = h(t) - \beta z$, where D^{-1} is the inverse of D; (ii) The linear transformation model reduces to the Cox model when $D(\epsilon) = e^{-e^\epsilon}$.

4. Prove equation (12) on page 191 of the seminal paper by Cox (1972, *Journal of the Royal Statistical Society, Series B*, page 187–220). You may assume that there are no "tied" observations. (If you have time, read the discussion of Kalbfleisch and Prentice at the end of the paper)

5. Two survival data sets are presented below:

Data Set 1			Data Set 2		
ID	OSTime	OSStatus	ID	OSTime	OSStatus
1	3	1	1	3	1
2	6	1	2	6	1
3	15	1	3	9	1
4	10	0	4	10	0
5	18	0	5	18	0

Note that the only difference between these two data is that the 3rd subject has a larger observed survival in the 1st data. Thus it might look intuitive that the KM curve for the 1st data should be above that of the 2nd. That is, it seems reasonable that the 1st survival is universally better than the 2nd survival. Use any software of your choice (or you may even compute by hand and then plot the estimates) to draw the two KM curves in the same plot and show that this is not the case. Also give a heuristic argument why it is so (you do not need to mathematically prove it).