SESSION 2:
ONE-SAMPLE METHODS

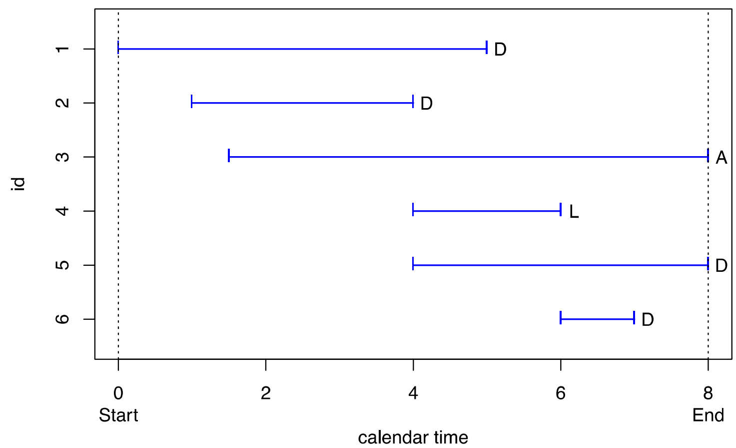
Module 12: Introduction to Survival Analysis
Summer Institute in Statistics for Clinical Research
University of Washington
June, 2017

Susanne May, Ph.D.
Professor
Department of Biostatistics
University of Washington

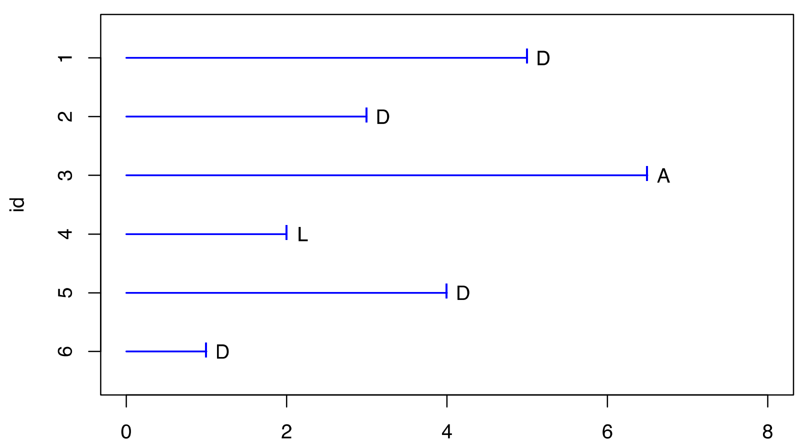
OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and CIs

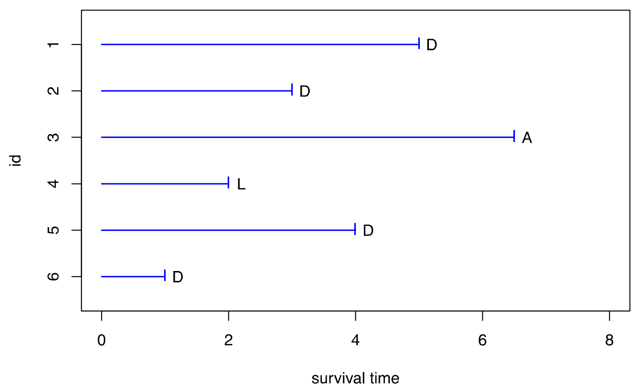
CLINICAL TRIAL



CENSORED DATA



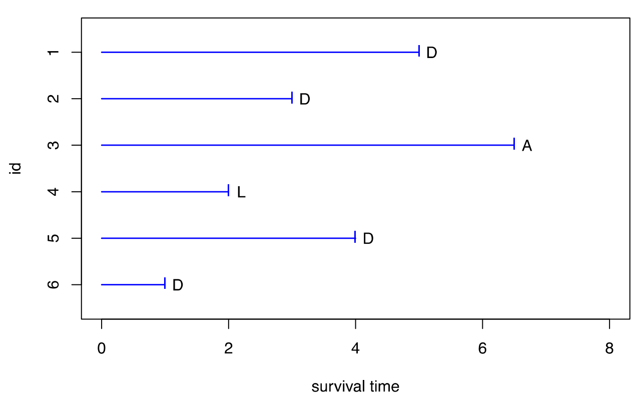
CENSORED DATA



id	Y	δ
1	5	1
2	3	1
3	6.5	0
4	2	0
5	4	1
6	1	1

"Censored" observations give some information about their survival time.

CENSORED DATA



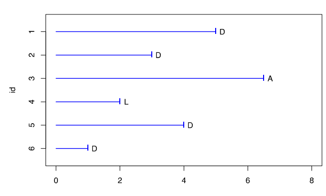
id	Y	δ
1	5	1
2	3	1
3	6.5	0
4	2	0
5	4	1
6	1	1

"Censored" observations give some information about their survival time.

ESTIMATION

- Can we use the partial information in the censored observations?
- Two off-the-top-of-the-head answers:
  - Full sample: Yes. Count them as observations that did not experience the event ever and estimate S(t) as if there were not censored observations.
  - Reduced sample: No. Omit them from the sample and estimate S(t) from the reduced data as if they were the full data.

CENSORED DATA



Problem: How to estimate:

	$\Pr[T > 3.5]$	$\Pr[T > 6]$
Full Sample:	$\frac{4}{6} = .67$	$\frac{2}{6} = .33$
Reduced Sample:	$\frac{2}{4} = .5$	$\frac{0}{4} = 0$

CENSORED DATA

Based on the data and estimates on the previous page,

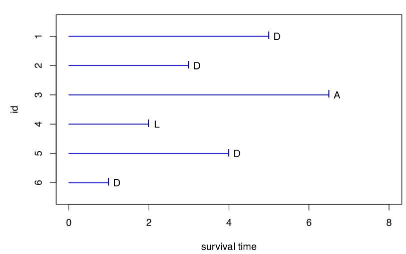
Q: Are the Full Sample estimates biased? Why or why not?

A:

Q: Are the Reduced Sample estimates biased? Why or why not?

A:

CENSORED DATA



Problem: How to estimate:

	$\Pr[T > 3.5]$	$\Pr[T > 6]$	
Full Sample:	$\frac{4}{6} = .67$	$\frac{2}{6} = .33$	← too high
Reduced Sample:	$\frac{2}{4} = .5$	$\frac{0}{4} = 0$	← too low

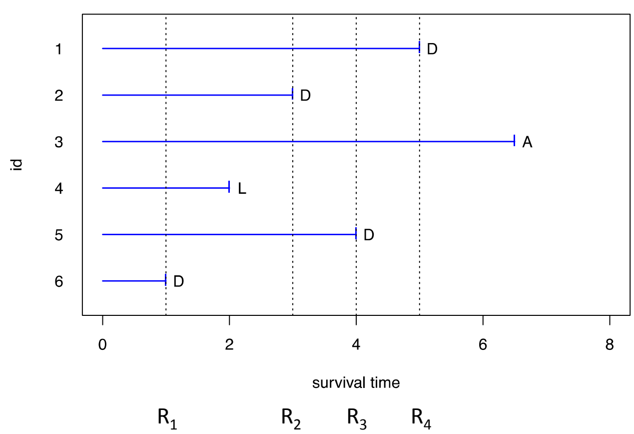
Need a good way to use the partial information in the censored observations.

IMPORTANT ASSUMPTION: Subjects who are censored at time t are representative of all subjects at risk of dying at time t.

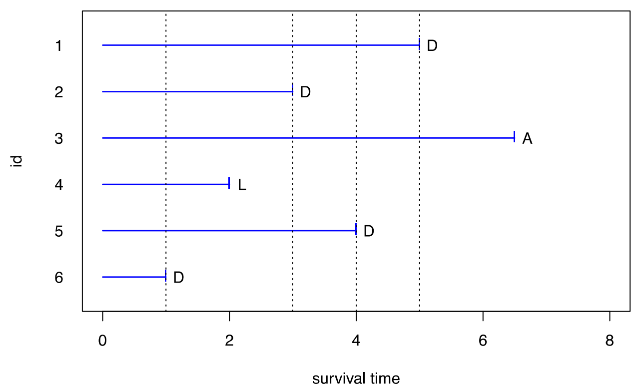
OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and CIs

RISK SETS



RISK SETS



R_1 {1,2,3,4,5,6} R_2 {1,2,3,5} R_3 {1,3,5} R_4 {1,3}

CENSORED DATA ASSUMPTION

- Important assumption: subjects who are censored at time t are at the same risk of dying at t as those at risk but not censored at time t.
  - When would you expect this to be true (or false) for subjects lost to follow-up?
  - When would you expect this to be true (or false) still alive at the time of the analysis?

CENSORED DATA ASSUMPTION

- Important assumption: subjects who are censored at time t are at the same risk of dying at t as those at risk but not censored at time t.
- This means the risk set at time t is an unbiased sample of the population still alive at time t.
- Can use information from the unbiased risk sets to estimate S(t) using the method of Kaplan and Meier (Product-Limit Estimator).

OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and CIs

USING RISK SETS INFO TO ESTIMATE S(t)

- Repeatedly use the fact that for $t_2 > t_1$,

  $\Pr[T > t_2] = \Pr[T > t_2 \text{ and } T > t_1] = \Pr[T > t_2 | T > t_1]\Pr[T > t_1]$

- An observation censored between t_1 and t_2 can contribute to the estimation of $\Pr[T > t_2]$ by its unbiased contribution to estimation of $\Pr[T > t_1]$.

PRODUCT-LIMIT (KAPLAN-MEIER) ESTIMATE

Notation: Let $t_{(1)}, t_{(2)} \ldots, t_{(J)}$ be the ordered failure times in the sample in ascending order.

$t_{(1)} =$ smallest Y_i for which $\delta_i = 1$ ($t_{(1)} = 1$)
$t_{(2)} = 2^{nd}$ smallest Y_i for which $\delta_i = 1$ ($t_{(2)} = 3$)
$\vdots$
$t_{(J)} =$ largest Y_i for which $\delta_i = 1$ ($t_{(4)} = 5$)

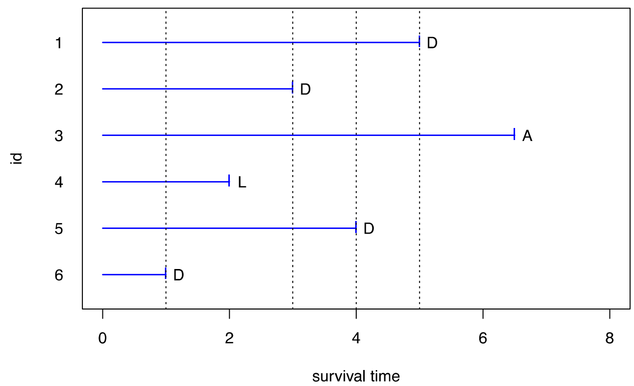
Q: Does J = the number of observed deaths in the sample?

A:

Q:When does J = n?

A:

$t_{(j)}$



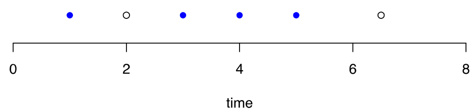
MORE NOTATION

For each $t_{(j)}$:

$D_{(j)}$ = number that die at time $t_{(j)}$
$S_{(j)}$ = number known to have survived beyond $t_{(j)}$
(by convention: includes those known to have been censored at $t_{(j)}$)
$N_{(j)}$ = number "at risk" of being observed to die at time $t_{(j)}$
(ie: number still alive and under observation just before $t_{(j)}$)

$S_{(j)} = N_{(j)} - D_{(j)}$

FOR EXAMPLE DATA



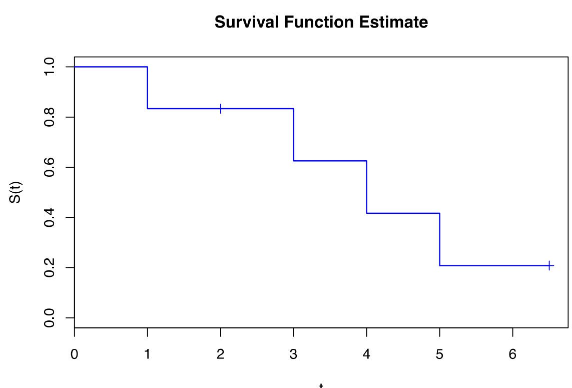
$t_{(j)}$	$N_{(j)}$	$D_{(j)}$	$S_{(j)}$
1	6	1	5
3	4	1	3
4	3	1	2
5	2	1	1

Product-limit (Kaplan-Meier) Estimator:

$$\hat{S}(t) = \Pi_{j:t_{(j)} \le t}(1 - \frac{D_{(j)}}{N_{(j)}}) = \Pi_{j:t_{(j)} \le t}(\frac{S_{(j)}}{N_{(j)}})$$

for t in	$\hat{S}(t)$
[0, 1)	1 (empty product)
$[1, 3)$	$1 \times \frac{5}{6} = .833$
$[3, 4)$	$1 \times \frac{5}{6} \times \frac{3}{4} = .625$
$[4, 5)$	$1 \times \frac{5}{6} \times \frac{3}{4} \times \frac{2}{3} = .417$
$[5, \infty)$	$1 \times \frac{5}{6} \times \frac{3}{4} \times \frac{2}{3} \times \frac{1}{2} = .208$

K-M ESTIMATOR



Note: does not descend to zero here (since last observation is censored).

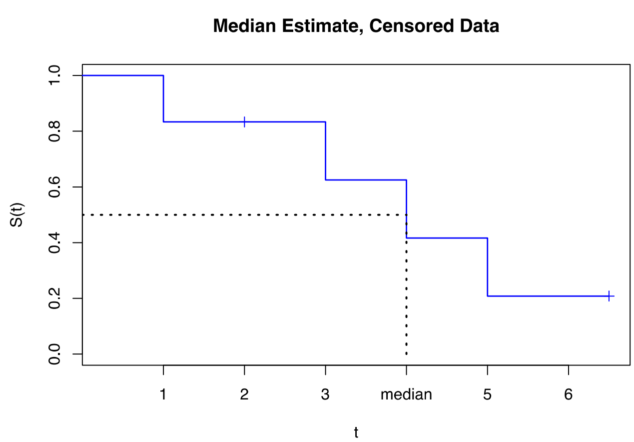
Q: Since the estimate jumps only at observed death times, how does information from the censored observations contribute to it?

A:

OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and CIs

MEDIAN SURVIVAL CENSORED DATA



OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and CIs

KM STANDARD ERRORS

Greenwood's Formula:

- $\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j:t_{(j)} \le t} \frac{D_{(j)}}{N_{(j)} S_{(j)}}$
- $\text{se}(\hat{S}(t)) = \sqrt{\widehat{Var}(\hat{S}(t))}$
- Pointwise CI: $(\hat{S}(t) - z_{\frac{\alpha}{2}} \text{se}(\hat{S}(t)), \quad \hat{S}(t) + z_{\frac{\alpha}{2}} \text{se}(\hat{S}(t)))$
  - Can include values < 0 or > 1.

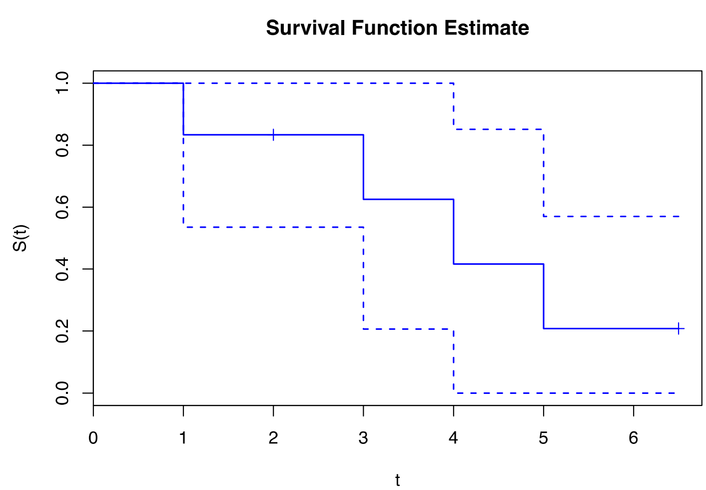
LOG –LOG KM STANDARD ERRORS

Use complementary log log transformation to keep CI within (0,1):

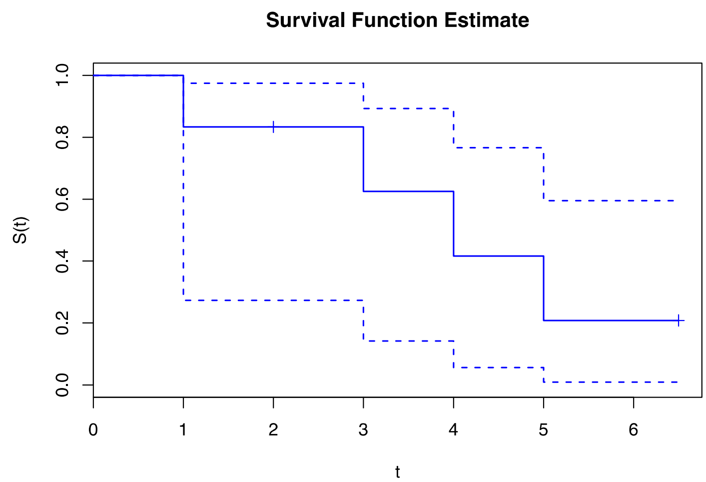
- $\widehat{Var}(\log(-\log(\hat{S}(t)))= \frac{\sum_{j:t_{(j)}\leq t} \frac{D_{(j)}}{N_{(j)}S_{(j)}}}{[\log(\hat{S}(t))]^2}$
- se = $\sqrt{\widehat{Var}(\log(-\log(\hat{S}(t)))}$
- CI for $\log(-\log(S(t)))$:

  $(\log(-\log(\hat{S}(t))) - z_{\frac{\alpha}{2}}\text{se},\quad \log(-\log(\hat{S}(t))) + z_{\frac{\alpha}{2}}\text{se})$
- CI for $\hat{S}(t)$: $([\hat{S}(t)]^{e^{z_{\alpha/2}\text{se}}}, [\hat{S}(t)]^{e^{-z_{\alpha/2}\text{se}}})$
  - CI remains within (0,1).

GREENWOOD'S FORMULA



COMPLEMENTARY LOG-LOG



MEDIAN CONFIDENCE INTERVAL

Confidence interval for the median is obtained by inverting the sign test of H_0 : median = M (Brookmeyer and Crowley, 1982).

- With complete data $T_1, T_2, \ldots, T_n$, the sign test of H_0 : median $= M$ is performed by seeing if the observed proportion, $\hat{P}[Y > M]$ is too big (Binomial Distribution or Normal Approximation).
- With censored data $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ giving incomplete data about $T_1, T_2, \ldots, T_n$, we cannot always tell whether $T_i > M$:

When $Y_i \leq M, \delta_i = 1$	observed death before M	we know $T_i \leq M$
When $Y_i > M$	death or censored after M	we know $T_i > M$
When $Y_i \leq M, \delta_i = 0$	censored before M	we don't know if $T_i \leq M$ or $T_i > M$

MEDIAN CONFIDENCE INTERVAL

Solution: Following Efron (self-consistency of KM), we estimate $\Pr[T > M]$ when $Y_i \leq M, \delta_i = 0$ using $\frac{\hat{S}(M)}{\hat{S}(Y_i)}$.

- For complete data, we let $U_i = \begin{cases} 1 & T_i > M \\ 0 & T_i \leq M \end{cases}$

  and our test is based on $\sum_{i=1}^{n} U_i$.

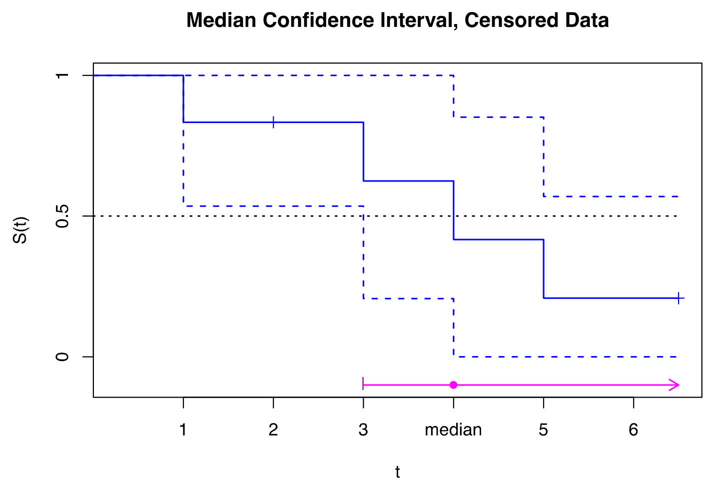
- For censored data, we let $U_i = \begin{cases} 1 & Y_i > M \\ \frac{\hat{S}(M)}{\hat{S}(Y_i)} & Y_i \leq M; \delta_i = 0 \\ 0 & Y_i \leq M; \delta_i = 1 \end{cases}$

  and our test is based on $\sum_{i=1}^{n} U_i$.

MEDIAN CONFIDENCE INTERVAL

- It turns out, this is the same as basing our test of H_0 : median $= M$ on a test of $H_0 : S(M) = \frac{1}{2}$.
- So a 95% CI for the median contains all potential M for which the test of $H_0 : S(M) = \frac{1}{2}$ cannot reject at $\alpha = .05$ (2 sided).
- Since $\hat{S}(M)$ only changes value at observed event times, the test need only be checked at $M = t_{(1)}, t_{(2)}, \ldots, t_{(J)}$.
- Originally proposed for Greenwood's formula CIs for $\hat{S}(M)$, but any good CIs are OK.
- Implemented in many software packages.

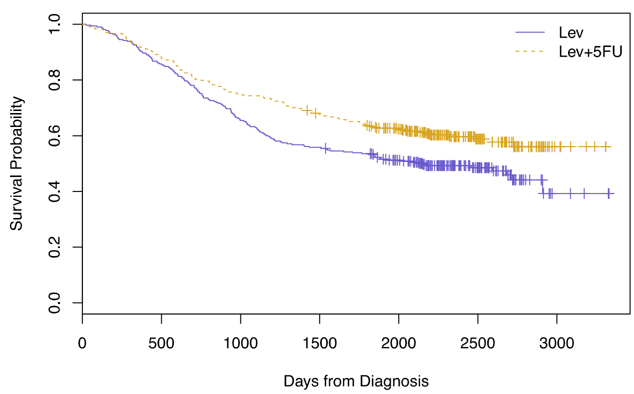
MEDIAN CONFIDENCE INTERVAL



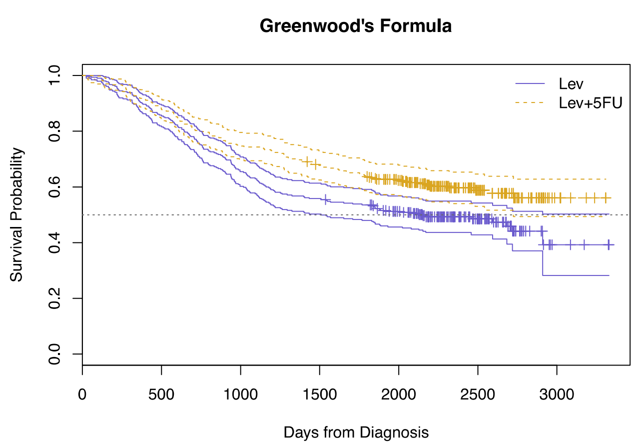
COLON CANCER EXAMPLE

- Clinical trial at Mayo Clinic (Moertel et al. (1990) NEJM)
- Stage B_2 and C colon cancer patients; adjuvant therapy
- Three arms
  - Observation only
  - Levamisole
  - 5-FU + Levamisole
- Stage C patients only
- Two treatment arms only

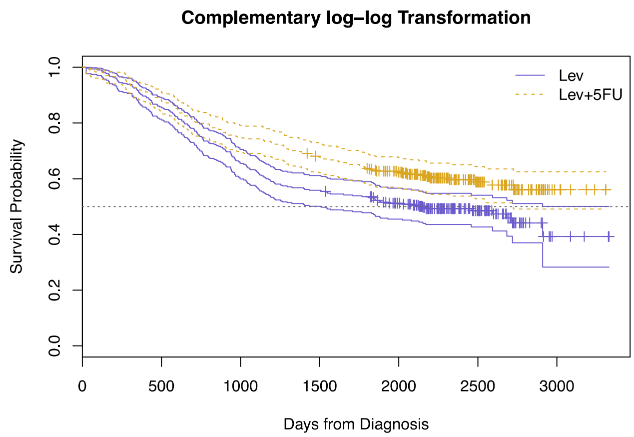
COLON CANCER EXAMPLE



COLON CANCER EXAMPLE



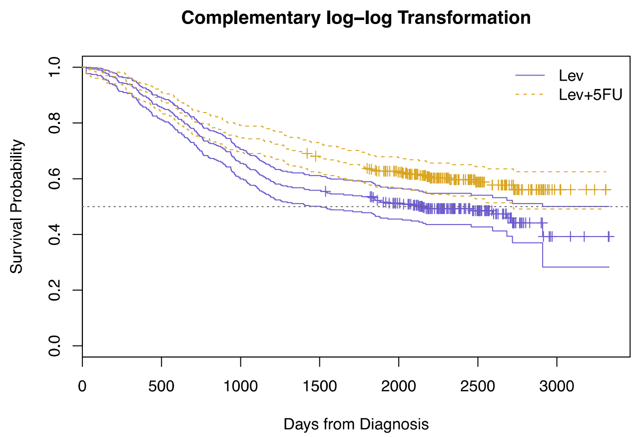
COLON CANCER EXAMPLE



PRESENTATION

	N	Events	Median (days)	95% CI
Levamisole Only	310	161	2152	(1509, ∞)
5FU + Levamisole	304	123	--	(2725, ∞)

COLON CANCER EXAMPLE



ESTIMATION

- Estimate S(t) using KM curve (nonparametric).
  - Pointwise standard errors and Cis
  - Almost always presented
  - Not appropriate when the event of interest happens only to some (more on this this tomorrow)
- Median: based on KM curve: often presented (too often?)

TO WATCH OUT FOR

- Mean survival time hard to estimate without parametric assumptions
  - Censoring means incomplete information about largest times
  - Mean over restricted time interval may be useful in some settings (some on this tomorrow)
- Median estimate more complicated than median of times
- Even with CIs, evaluating differences between curves visually is subjective
- Interpretation of survival function estimates depends on validity of censoring assumptions