SESSION 3:
ADDITIONAL TWO-SAMPLE TESTS

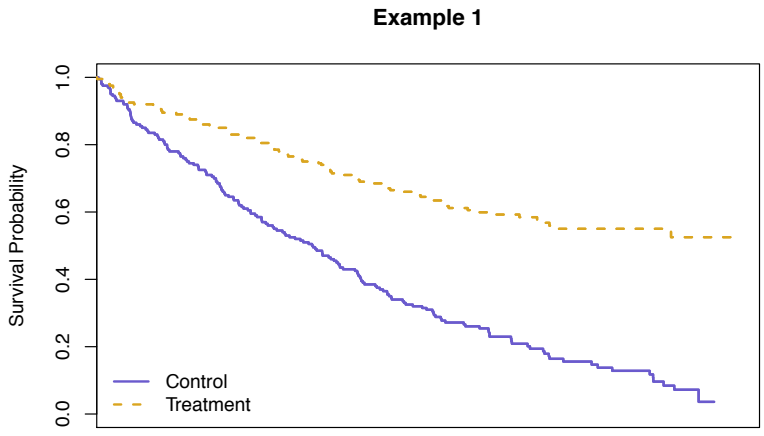
Module 12: Survival Analysis in Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2017

Barbara McKnight, Ph.D.
Professor
Department of Biostatistics
University of Washington

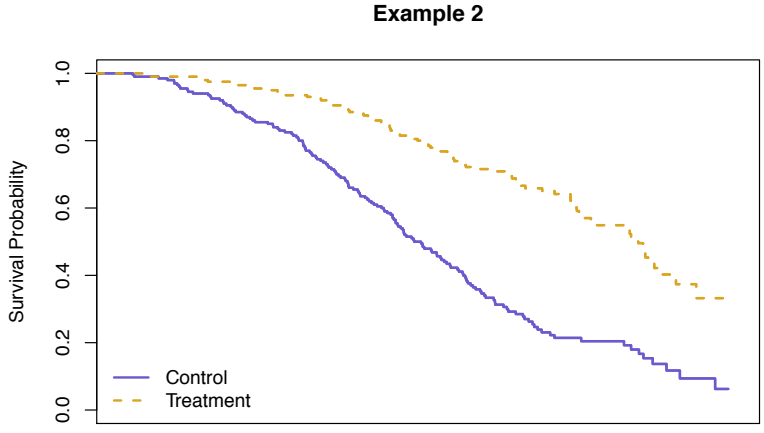
OUTLINE

- Limitations of proportional hazards
- Other contrasts based on functionals of S(t)
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- Other metrics to describe the distance between survival curves
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

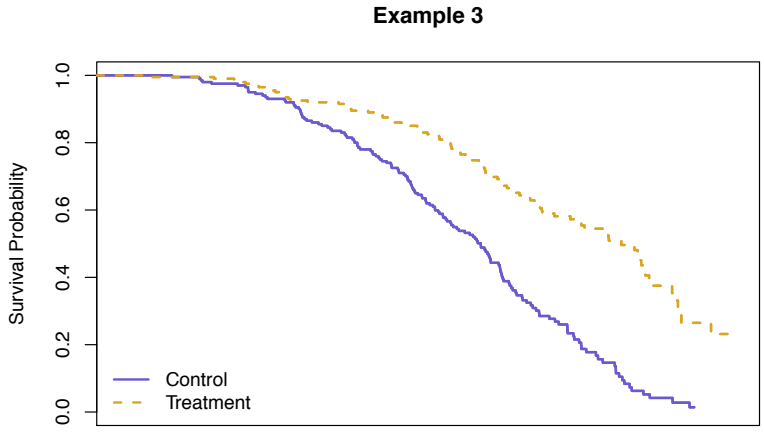
PROPORTIONAL HAZARDS EXAMPLES



PROPORTIONAL HAZARDS EXAMPLES



PROPORTIONAL HAZARDS EXAMPLES

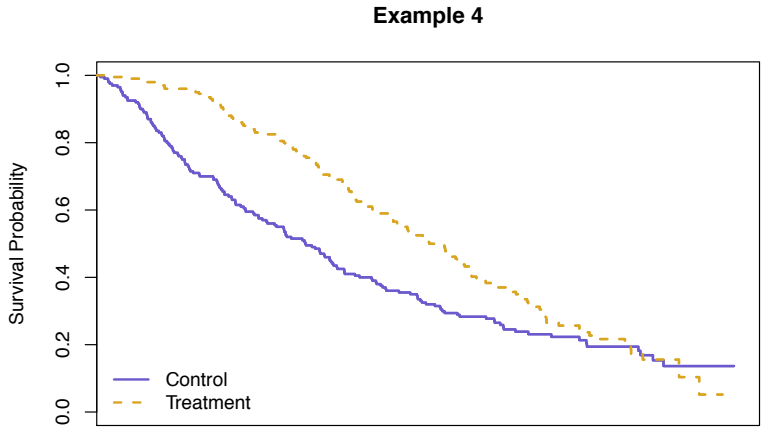


PROPORTIONAL HAZARDS EXAMPLES

Q: Which group has better survival in these examples?

A:

NON-PROPORTIONAL HAZARDS EXAMPLES



NON-PROPORTIONAL HAZARDS EXAMPLES

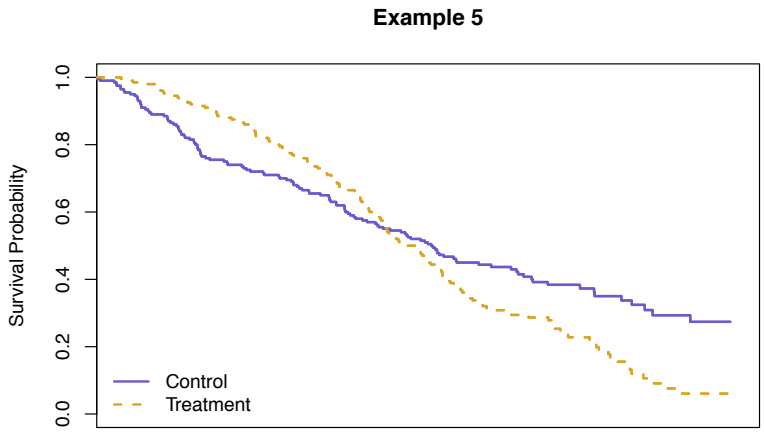
Q: Why does it appear the hazards are not proportional?

A:

Q: Which group has better survival?

A:

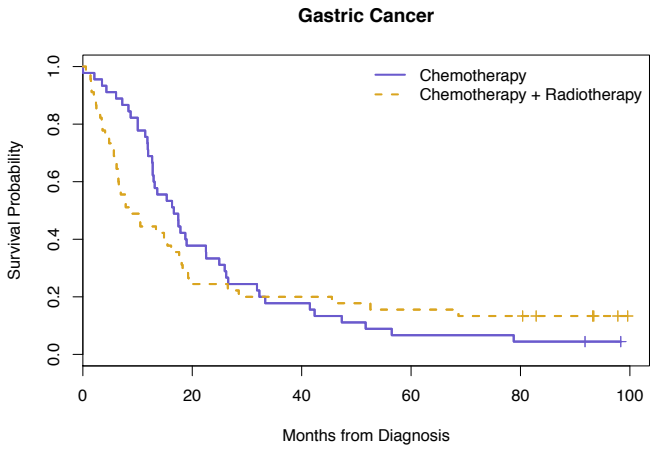
NON-PROPORTIONAL HAZARDS EXAMPLES



YOUR CHOICE

- Which group has better survival?
- You are a newly diagnosed patient. What would you want to know before choosing which treatment to take?

REAL DATA



Schein PS, Gastrointestinal Tumor Study Group. A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. Cancer. 1982 May 1;49(9):1771–1777.

HAZARD RATIO

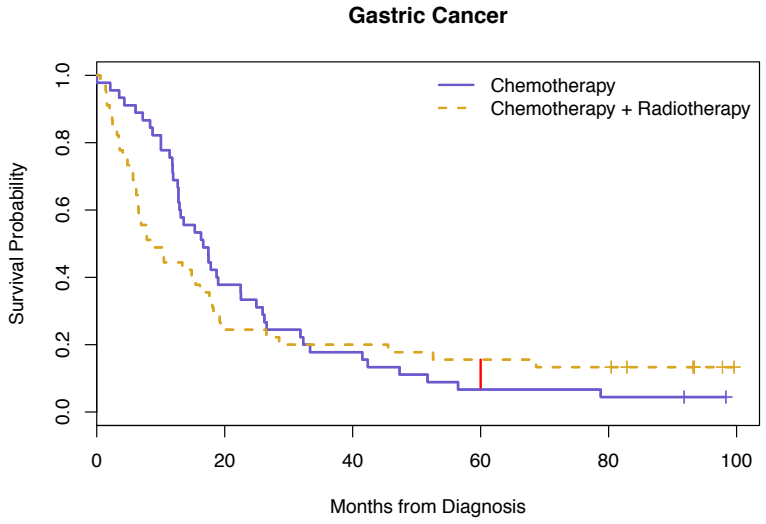
	Hazard Ratio	95% CI	P-value
Chemotherapy	1.0 (reference)	--	--
Chemotherapy + Radiotherpay	1.1	(0.72, 1.7)	.63

CROSSING HAZARDS

When the proportional hazards assumption doesn't hold:

- Cox model will give weighted-average of time-specific hazard ratios (weights depend on censoring distribution)
- log rank test will test whether a weighted-average difference of hazards is zero
  - statistic numerator $= \sum_j \frac{n_{1j}n_{2j}}{(n_{1j}+n_{2j})}(\frac{d_{1j}}{n_{1j}} - \frac{d_{2j}}{n_{2j}})$
  - More weight at earlier times when number at risk is larger
- May not be the quantity on which you want to base inference (estimation and testing)

FIVE-YEAR SURVIVAL



FIVE-YEAR SURVIVAL

- Compares only at a single point in time
- Ignores earlier survival differences, which may be important to some patients, given that in this example survival to 5 years in either group is low

S(t) AT A CHOSEN TIME t

- Choose time t for comparison at design stage.
- Compare $\hat{S}_1(t)$ to $\hat{S}_2(t)$ using

$$\frac{\hat{S}_1(t) - \hat{S}_2(t)}{\sqrt{\widehat{\text{var}}(\hat{S}_1(t)) + \widehat{\text{var}}(\hat{S}_2(t))}}$$

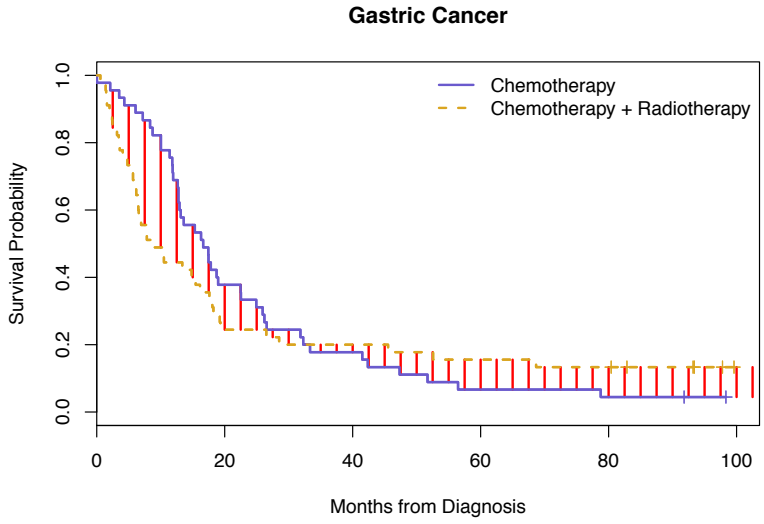
where $\widehat{\text{var}}(\hat{S}_2(t))$ is computed using Greenwood's formula or another large-sample formula such as the one based on the complementary log-log of $\hat{S}(t)$.

FIVE-YEAR SURVIVAL DIFFERENCE

Gastric Cancer

Difference	se(Difference)	Z Statistic	P-value
.0889	.0656	1.36	.1753

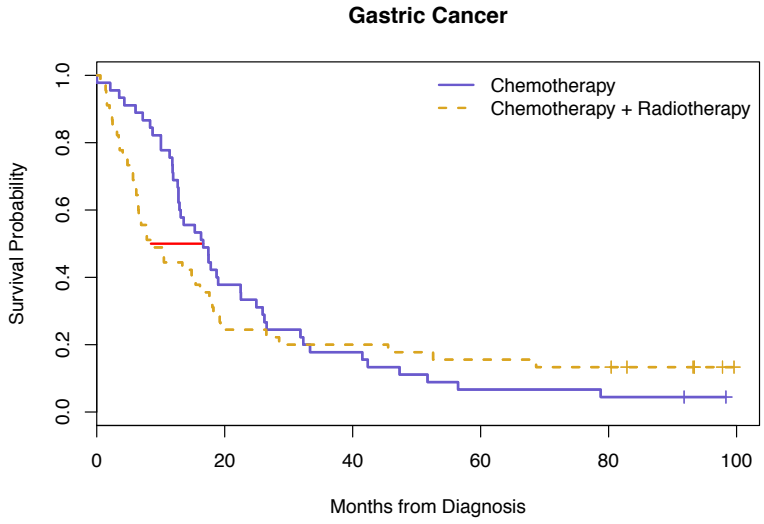
COMPARISON AT MORE THAN ONE TIME



AVERAGE DIFFERENCES

- Average difference between survival curves over time might be of interest
- In gastric cancer example, differences are of different signs at different times, so there would be cancellation
- Allows poorer survival after survival curves cross to detract from better survival before
- Interpretation?
- Also related to average quantile difference

MEDIAN SURVIVAL



MEDIAN SURVIVAL

- Compares only a single quantile
- Hard for some patients to interpret the difference in medians

MEDIAN TEST

Idea: Define $\hat{M}_1$ and $\hat{M}_2$ to be the median survival times in the two samples.

Then let the overall median survival time be defined by the weighted average.

$$\hat{M} = \frac{N_1}{N}\hat{M}_1 + \frac{N_2}{N}\hat{M}_2$$

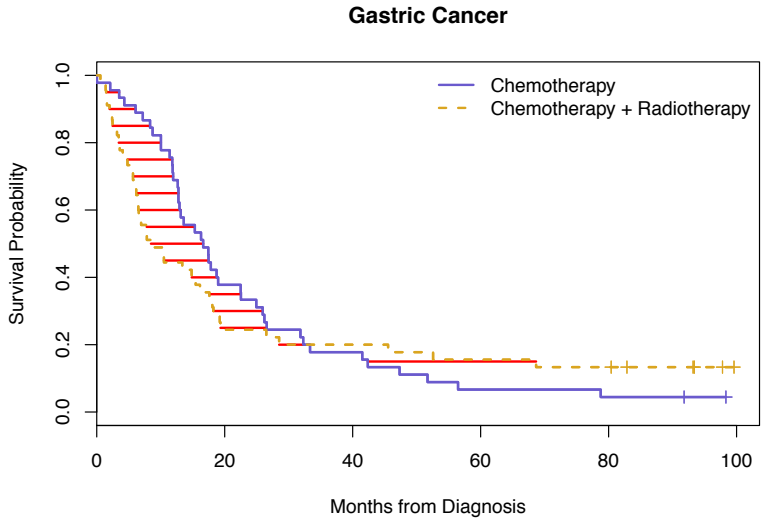
A test of $H_0 : M_1 = M_2$ can be performed by testing

$$H_0 : S_1(\hat{M}) = S_2(\hat{M})$$

Reference distribution based on joint asymptotic distribution of $(S_1(\hat{M}), S_2(\hat{M}))$.

Brookmeyer R, Crowley J. JASA 1982;77(378):433–440.

MORE THAN ONE QUANTILE



MEAN SURVIVAL TIME

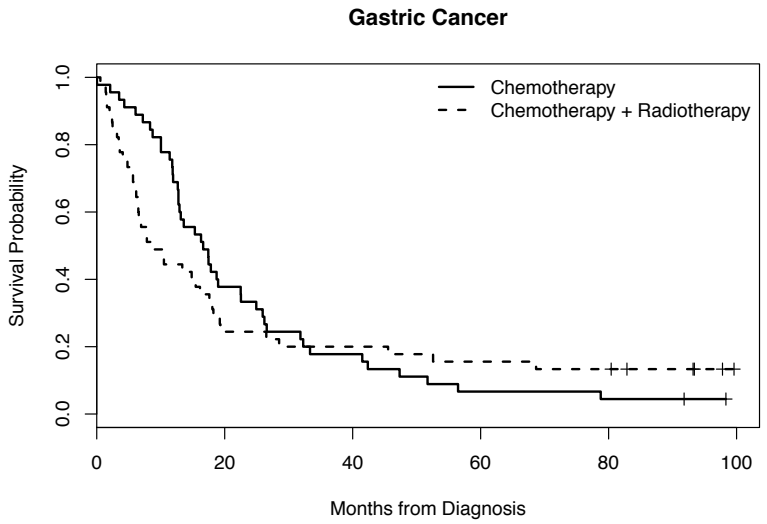
Useful Fact: $\int_0^\infty S(t)dt = E(T) = \int_0^\infty tf(t)dt$

Proof: $\int_0^\infty S(t)dt = S(t)t|_0^\infty - \int_0^\infty t(-f(t))dt = \int_0^\infty tf(t)dt$

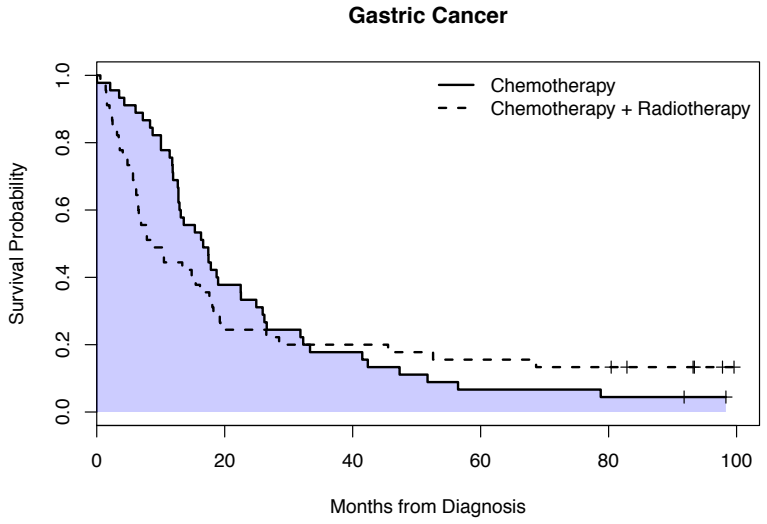
by integration by parts and

the fact that $E(T) < \infty \Rightarrow tS(t) \overset{t\to\infty}{\to} 0$.

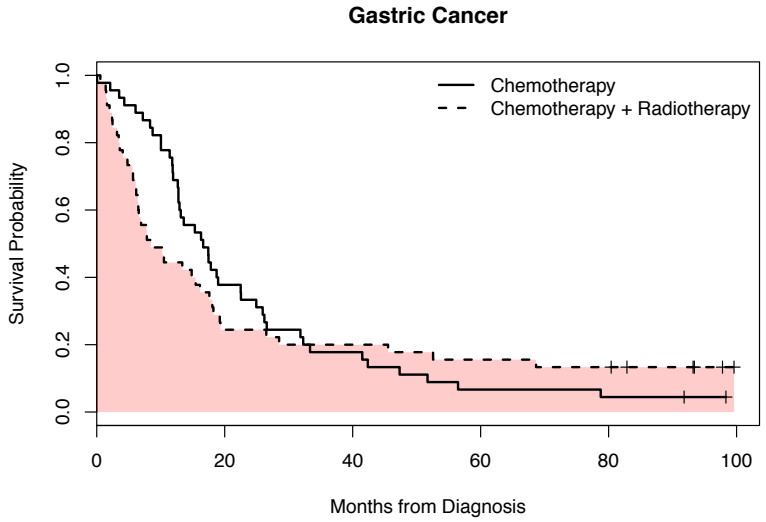
MEAN SURVIVAL TIME



MEAN SURVIVAL TIME



MEAN SURVIVAL TIME



MEAN SURVIVAL TIME

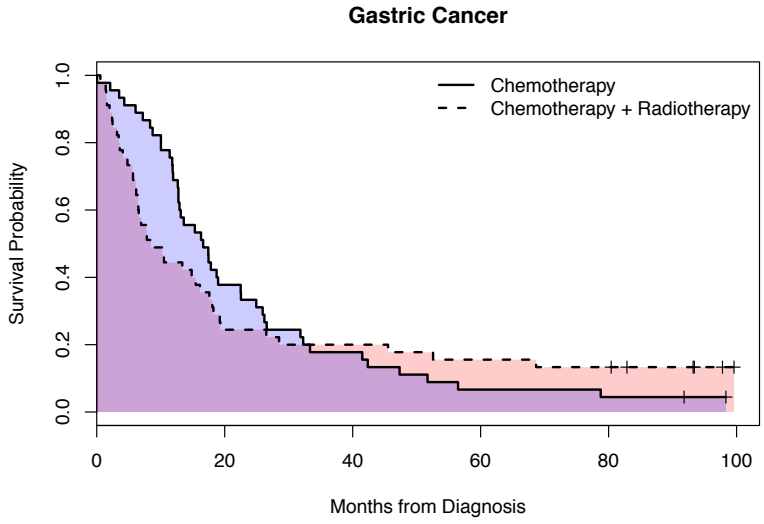
- Mean survival time $\mu = \int_0^\infty S(t)dt$
- Large sample (asymptotic) distribution proved by Gill in The Annals of Statistics. 1983;11(1):49–58.
- In finite samples, can be infinite if last time is a censoring
  - Integrate to last failure time only
  - Integrate to last observed time only

MEAN SURVIVAL TIME

	Mean Survival*	SE
Chemotherapy	24.1 months	3.3 months
Chemotherapy + Radiotherapy	24.3 months	4.8 months

* Up to 99.6 months (last observed time in either group)

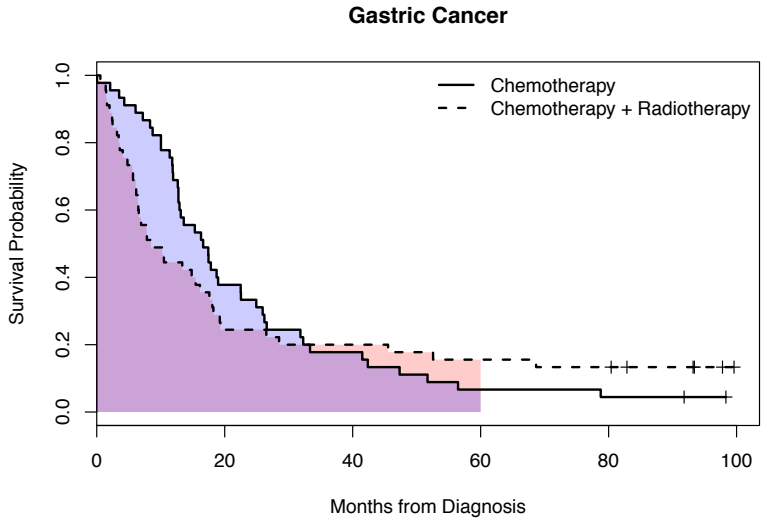
MEAN SURVIVAL TIME



MEAN SURVIVAL TIME DIFFERENCE

- Average of survival function differences over time
- Average of survival quantile differences over quantiles
- Allows cancellation
- Not much information at late times where few are at risk.
- Infinite estimate if KM curve doesn't descend to zero
- May want to truncate to a shorter interval, restricting to times where *S(t)* estimates are precise

RESTRICTED MEAN SURVIVAL TIME



RESTRICTED MEAN SURVIVAL TIME

- Interpretation: average time lived in the interval $[0, \tau]$.
- Interpretation for differences: on average, the amount more time lived in $[0, \tau]$ on treatment A than on treatment B.
- Some asymptotically equivalent ways to estimate it:
  - $\hat{\mu} = \int_0^\tau \hat{S}(t)dt$
  - $\frac{1}{n}\sum_{i=1}^{n} \frac{d_i y_i}{\hat{S}_{c(y_i)}}$ where $\hat{S}_{c(y_i)}$ is the KM estimated survival function of the censoring distribution
  - Using pseudo-observations based on the jackknife.

$$\hat{\mu} = \sum_{i=1}^{n} \hat{\mu}_i,$$

RESTRICTED MEAN SURVIVAL DIFFERENCE

- Standard estimation and testing:
  - $\hat{\mu}_k = \int_0^\tau \hat{S}_k(t)dt$
  - $\widehat{\text{var}}(\hat{\mu}_k) = \sum_{j=1}^{J} [\int_{t_j}^{\tau} \hat{S}_K(t)dt]^2 \frac{D_{jk}}{N_{jk}(N_{jk}-D_{jk}))}$
  - Compare test statistic:

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\widehat{\text{var}}(\hat{\mu}_1) + \widehat{\text{var}}(\hat{\mu}_2)}}$$

to standard normal distribution (asymptotic).

RESTRICTED MEAN SURVIVAL TIME

$$E[\min(T,\tau)] = \widehat{E[Y]} = \int_0^{\tau} \hat{S}(t)dt$$

Several approaches to variance estimation:

- Asymptotic
- Random perturbation resampling method (Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostat. 2014 Apr 1;15(2):222–233.)
- Variance of pseudo observations

PSEUDO OBSERVATIONS

- There are a number of other less direct ways to estimate $\mu_k = \int_0^{\tau} \hat{S}_k(t)dt$ that make generalizing to regression models easier.
- One appealing method based on creating pseudo-observations based on the jackknife.
  - Group means computed in the usual way from pseudo-observations
  - Standard errors computed from pseudo-observations in the usual way.
  - Test statistic based on two-sample t-test (unequal variances) with pseudo-observations.

PSEUDO OBSERVATIONS

Estimation of μ using pseudo-observations based on the jackknife.

$$\hat{\mu} = \sum_{i=1}^{n} \hat{\mu}_i,$$

where $\hat{\mu}_i = n\hat{\mu} - (n-1)\hat{\mu}_{-i}$.

- $\hat{\mu}$ is computed by the first method from the pooled sample, and
- $\hat{\mu}_{-i}$ is computed the same way but leaving out the i^{th} observation.
- Andersen et al. Lifetime Data Anal. 2004;10(4):335–350.
- Functions available in Stata, R and SAS.

RESTRICTED MEAN SURVIVAL TIME

	Restricted Mean Survival (2000 days)	SE
Chemotherapy	673	77.8
Chemotherapy + Radiotherapy	599	101.1

Comparison Method	P-value
Asymptotic	.560
Pseudo observations	.566

DESIGN AND INFERENCE ISSUES

- Not much information / precision available at late times when few subjects are at risk
  - If a restricted mean over an interval $[0, \tau]$ is of interest, important to follow subjects enough longer than τ to have an adequate number still at risk at time τ.

EXAMPLE

- Schermerhorn et al. (2015) compared survival in a matched cohort of 39,966 pairs of Medicare patients who received either endovascular or open repair of an abdominal aortic aneurism.
  - Perioperative mortality and complication rates were higher in those given open repair: 5.2% vs 1.6% for mortality and 12.9% vs 3.8%
  - The estimated hazard ratio for death comparing endovascular to open repair varied over time:
    - HR = .32 (95% CI: .29 - .35) over the first 30 days
    - HR = .64(95% CI: .58 -.71) for 30 – 90 days
    - HR = 1.17(95% C: I 1.13 – 1.21) for 90 days – 4 years
    - HR = 1.05 (95% CI: 1.00 - 1.09) after 4 year.

Schermerhorn ML, Buck DB, O'Malley AJ et al. NEJM 2015 Jul 23;373(4):328–338.

EXAMPLE

- Because of non-proportional hazards they estimated differences in restricted mean survival using the pseudo observation approach of Andersen et al with the matched-pair data.
  - Over the first 4 years, the endovascular group lived an average of 12.4 days longer (95% CI 9.0 – 15.6)
  - Over the first 7 years, the endovascular group lived an average of 8.2 days longer (95% CI: 1.5-14.4)
  - The authors concluded that the advantage of endovascular repair persisted to 7 years.
- The pseudo-observation approach makes it easy to accommodate the matched design.

SCREENING TRIAL

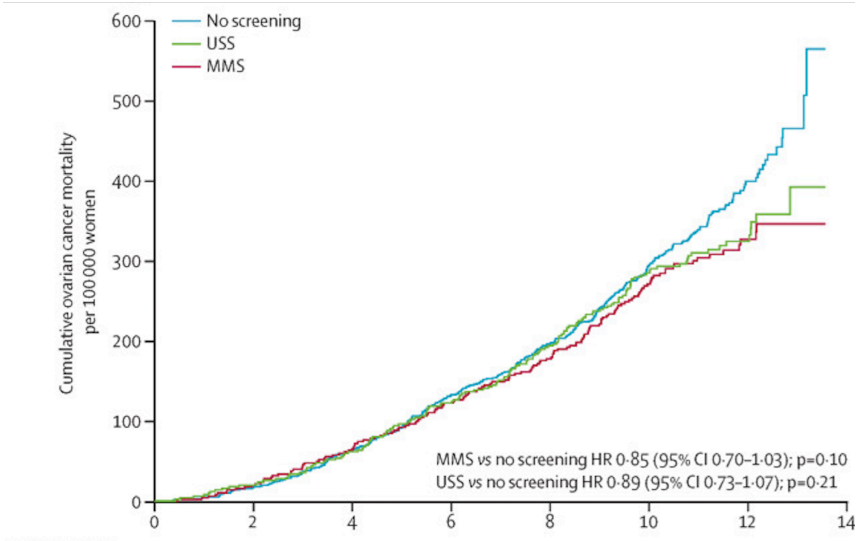
- 202,546 women 50-72 years of age, England, Wales, Northern Ireland
- Randomized to one of three arms in 1:1:2 ratio between June 1, 2001 and Oct 21, 2005.
  - Annual multimodal screening (serun CA 125 + algorithm)
  - Annual transvaginal ultrasound
  - No screening
- Screening ended Dec 31, 2011.
- Not blinded
- Primary outcome: death from ovarian cancer (by end of 2014)

Jacobs IJ, Menon U, Ryan A, et al. (2016) The Lancet. 387(10022):945–956.

OVARIAN CANCER SCREENING TRIAL

- Primary analysis: Cox regression (proportional hazards)
  - MMS vs. no screening: Mortality reduction =
    (1 – HR)100 = 15% (95% CI: -1% – 33%) P = .10
  - USS vs. no screening: Mortality reduction =
    (1 – HR) 100 = 11% (95% CI: -7% - 27%) P = .21

OVARIAN CANCER SCREENING TRIAL



Number at risk

No screening	101299	100720	99662	98238	96632	75582	25252
MMS	50624	50343	49846	49176	48345	37758	12592
USS	50623	50338	49838	49192	48363	37768	12689

OVARIAN CANCER SCREENING TRIAL

- Secondary analyses, excluding prevalent cases:
- Post-hoc Weighted* logrank test:
  - MMS mortality reduction = 22% (3-38%) P = .023
  - USS mortality reduction = 20% (0 – 35%) P = .049

  * by pooled cumulative mortality

ANOTHER OPTION: METRICS

- Tests based on detecting consistent differences between survival curves or hazard across time lose power when the hazards or survival curves cross.
- Weighting can focus on a time period when direction of differences is consistent.
- Other metrics can measure distance between survival functions or hazard functions in a way that does not require the direction of differences to be consistent
- Tests based on them can have more power when survival functions or hazards cross.

METRICS

- Supremum: Tests based on the supremum of a difference of cumulative weighted hazard functions over $[0, t_m]$:

$$\sup_{t\in[0,t_m]} \sum_{i:t_i<t} W_i \frac{n_{1i}n_{2i}}{n_{1i}+n_{2i}}\left(\frac{d_{1i}}{n_{1i}} - \frac{d_{1i}}{n_{1i}}\right)$$

  - Gill, R.D. (1980). Censoring and stochastic integrals. Math. Centre Tracts 124, Mathematisch Centrum Amsterdam.
  - Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Biometrics. 1980;36(4):607–625.
  - Fleming TR, Harrington DP, O'Sullivan M. JASA. 1987;82(397):312–320.

METRICS

- l^2: Tests based on the integrated squared difference of survival or cumulative hazard functions over $[0, t_m]$:

$$\sum_{t_i:t_i\le t_m, \delta_i=1} (\hat{S}_2(t_i) - \hat{S}_1(t_i))^2 d(-\hat{S}(t_i))$$

  or

$$\sum_{t_i:t_i\le t_m, \delta_i=1} ((\hat{S}_2(t_i) - \hat{S}_1(t_i))W_i)^2 d(\hat{H}(t_i))$$

  where the weight function W_i and H are functions of the asymptotic covariance of the cumulative hazard estimator at different times.

  - Koziol Biom. J. 1978;20(6):603–608.
  - Koziol, Yuh . Biom. J. 1982;24(8):743–750.
  - Schumacher. International Statistical Review 1984;52(3):263–281.

ISSUE

- Hard to think of a good scientific hypothesis that specifies which of these metrics and associated tests is consistent with the hypothesis.
- Large temptation to choose the type of test after looking at the data and noticing crossing hazards or crossing survival functions in the search for a powerful test.
- Scientific hypotheses more likely to be consistent with a difference between functionals of the survival function S(t).

TO WATCH OUT FOR

- Base quantity to be compared (weighted sum for logrank, time, quantile or restricted mean) on what would be meaningful in the context of the trial.
- Important to choose it before looking at the data.