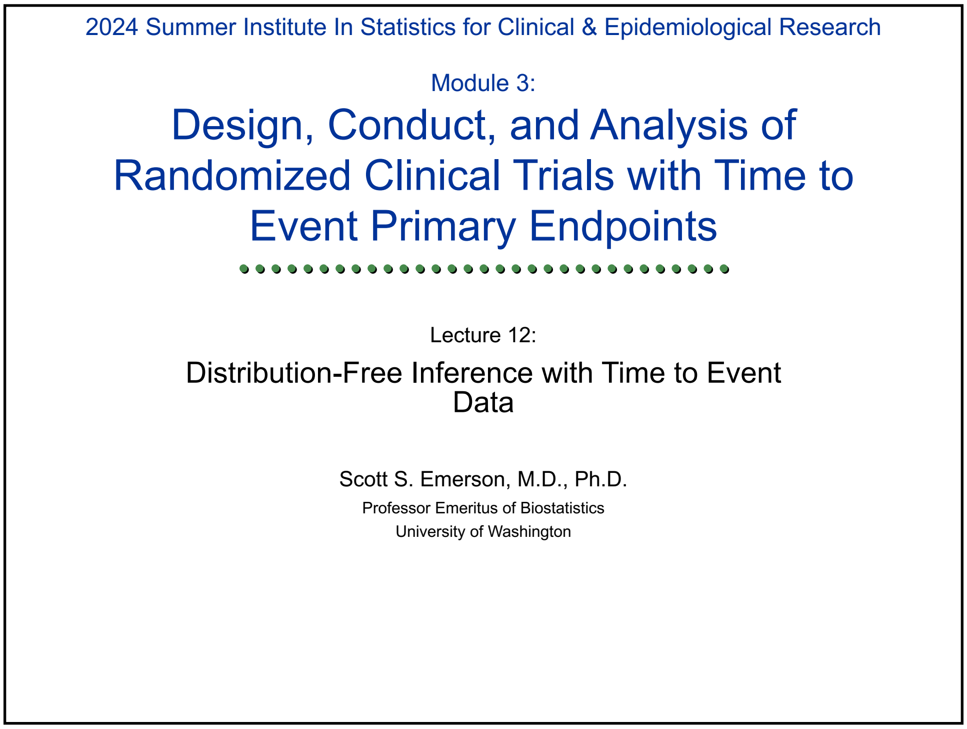


2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

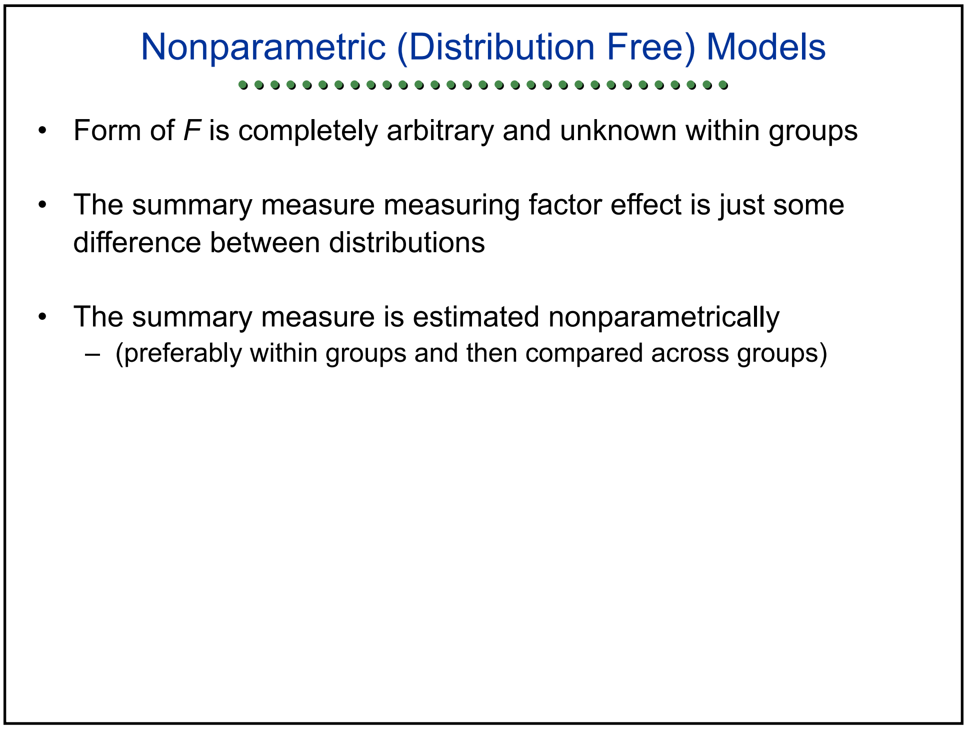
Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 12:

Distribution-Free Inference with Time to Event Data

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

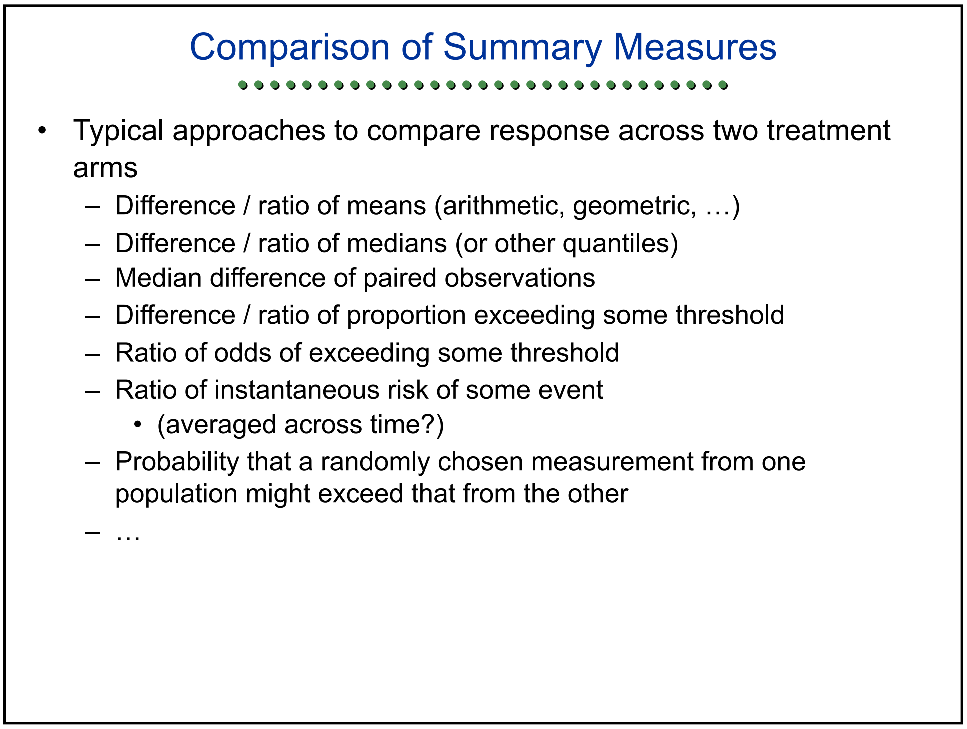
1



Nonparametric (Distribution Free) Models

- Form of *F* is completely arbitrary and unknown within groups
- The summary measure measuring factor effect is just some difference between distributions
- The summary measure is estimated nonparametrically
  - (preferably within groups and then compared across groups)

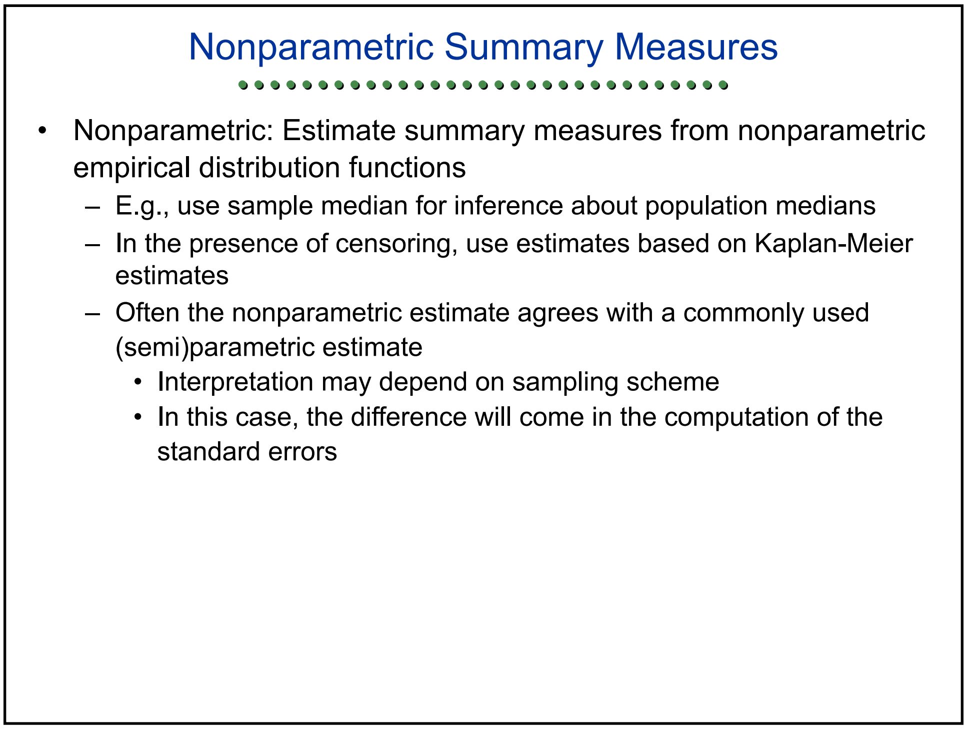
2



Comparison of Summary Measures

- Typical approaches to compare response across two treatment arms
  - Difference / ratio of means (arithmetic, geometric, …)
  - Difference / ratio of medians (or other quantiles)
  - Median difference of paired observations
  - Difference / ratio of proportion exceeding some threshold
  - Ratio of odds of exceeding some threshold
  - Ratio of instantaneous risk of some event
    - (averaged across time?)
  - Probability that a randomly chosen measurement from one population might exceed that from the other
  - …

3



Nonparametric Summary Measures

- Nonparametric: Estimate summary measures from nonparametric empirical distribution functions
  - E.g., use sample median for inference about population medians
  - In the presence of censoring, use estimates based on Kaplan-Meier estimates
  - Often the nonparametric estimate agrees with a commonly used (semi)parametric estimate
    - Interpretation may depend on sampling scheme
    - In this case, the difference will come in the computation of the standard errors

4

Nonparametric Summary Measures

Using Kaplan - Meier survival estimate $\hat{S}(t)$

Mean : $$\hat{\theta} = \int_0^\infty \hat{S}(u) du$$

Median : $$\hat{\theta} = \hat{S}^{-1}(0.5)$$

Proportion above threshold : $$\hat{\theta} = \hat{S}(a)$$

Weighted average of hazard : $$\hat{\theta} = \int_0^\infty w(u) \hat{\lambda}(u) du$$

5

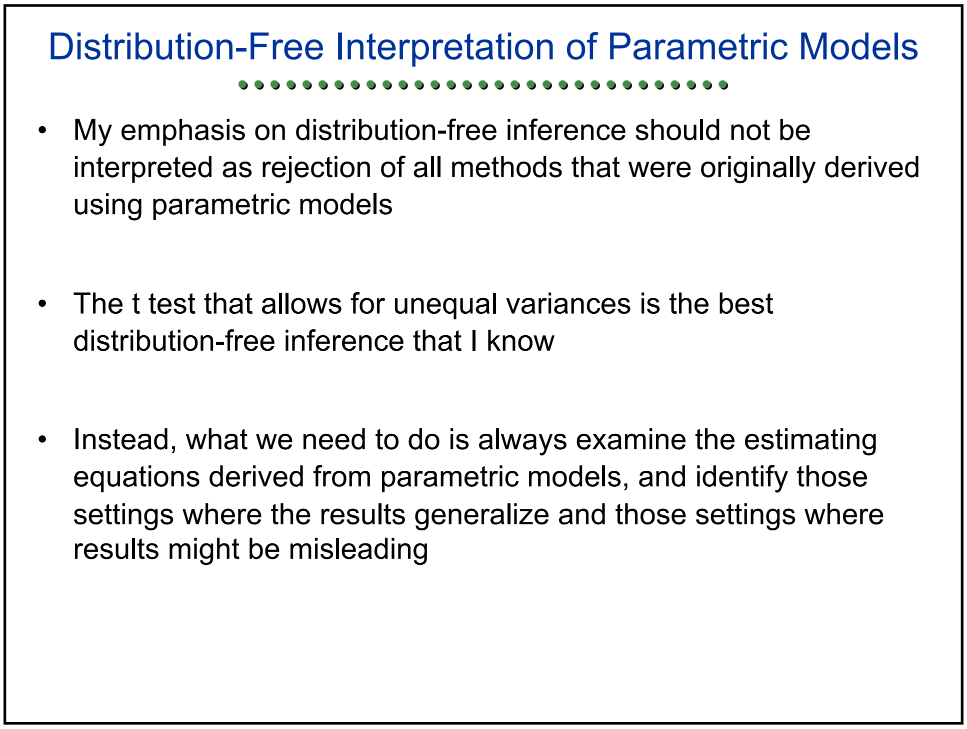
Nonparametric Summary Measures

- Depending on the censoring scheme, not all summary measures are estimable
- The support of the censoring distribution may preclude estimation of the mean and some quantiles
- Can instead use the mean of the truncated distribution
  - "Average increase in days alive during first 5 years"

Mean of truncated distribution : $$\hat{\theta} = \int_0^a \hat{S}(u) du$$

- In most cases, variance estimates can be obtained from the asymptotic theory of the Kaplan-Meier estimates

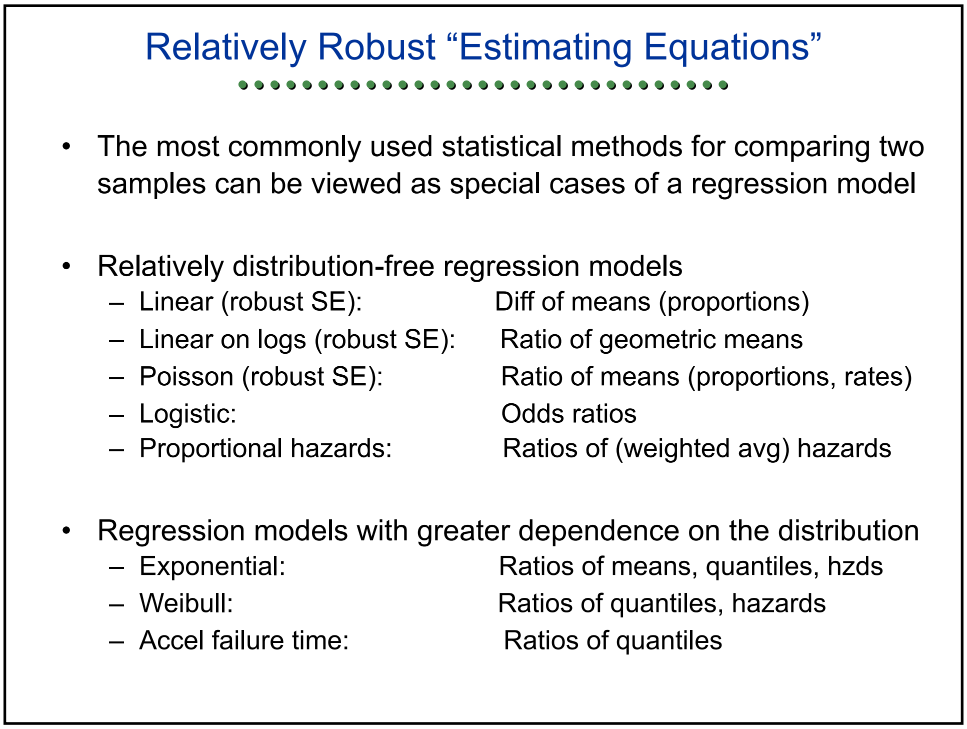
6



Distribution-Free Interpretation of Parametric Models

- My emphasis on distribution-free inference should not be interpreted as rejection of all methods that were originally derived using parametric models
- The t test that allows for unequal variances is the best distribution-free inference that I know
- Instead, what we need to do is always examine the estimating equations derived from parametric models, and identify those settings where the results generalize and those settings where results might be misleading

7



Relatively Robust “Estimating Equations”

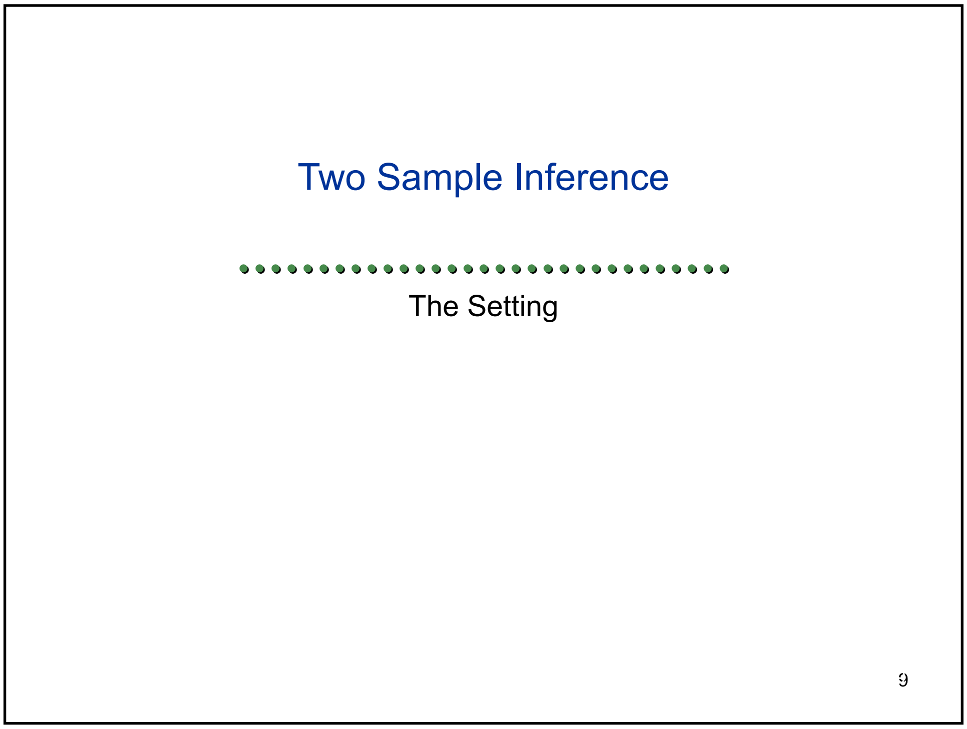
- The most commonly used statistical methods for comparing two samples can be viewed as special cases of a regression model
- Relatively distribution-free regression models

– Linear (robust SE):	Diff of means (proportions)
– Linear on logs (robust SE):	Ratio of geometric means
– Poisson (robust SE):	Ratio of means (proportions, rates)
– Logistic:	Odds ratios
– Proportional hazards:	Ratios of (weighted avg) hazards

- Regression models with greater dependence on the distribution

– Exponential:	Ratios of means, quantiles, hzds
– Weibull:	Ratios of quantiles, hazards
– Accel failure time:	Ratios of quantiles

8

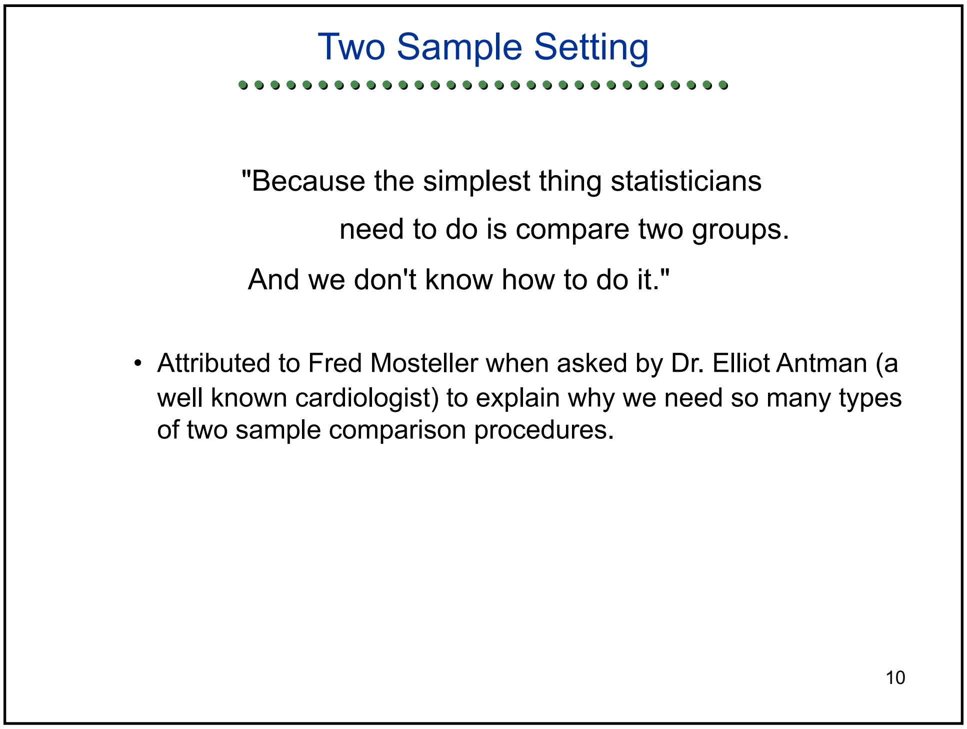


Two Sample Inference

The Setting

9

9



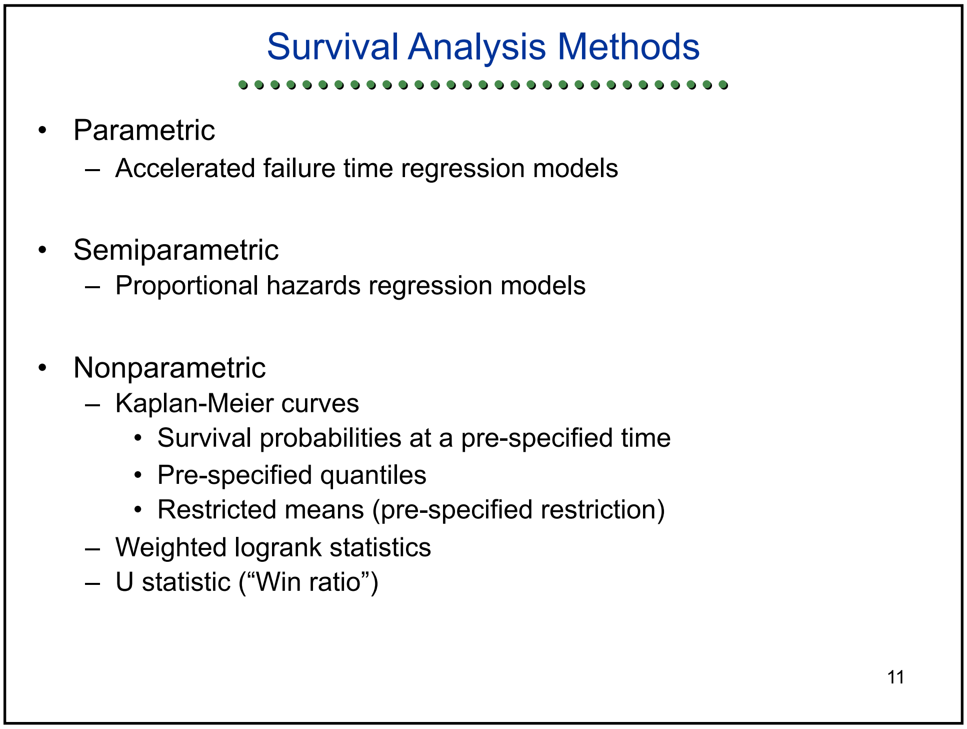
Two Sample Setting

"Because the simplest thing statisticians
need to do is compare two groups.
And we don't know how to do it."

- Attributed to Fred Mosteller when asked by Dr. Elliot Antman (a well known cardiologist) to explain why we need so many types of two sample comparison procedures.

10

10

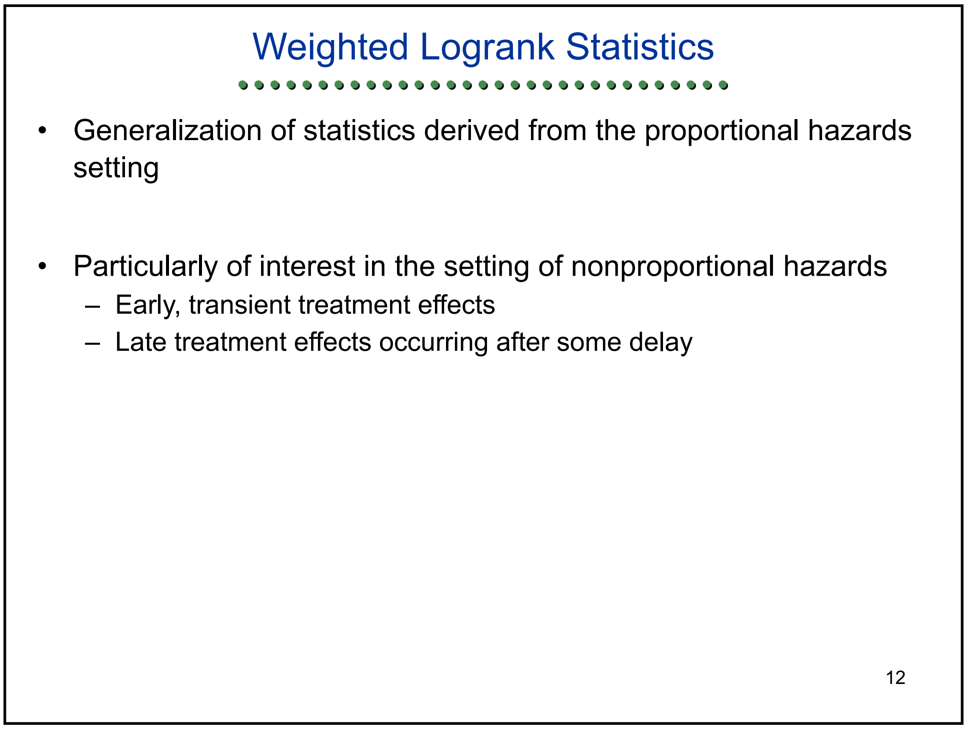


Survival Analysis Methods

- Parametric
  - Accelerated failure time regression models
- Semiparametric
  - Proportional hazards regression models
- Nonparametric
  - Kaplan-Meier curves
    - Survival probabilities at a pre-specified time
    - Pre-specified quantiles
    - Restricted means (pre-specified restriction)
  - Weighted logrank statistics
  - U statistic (“Win ratio”)

11

11

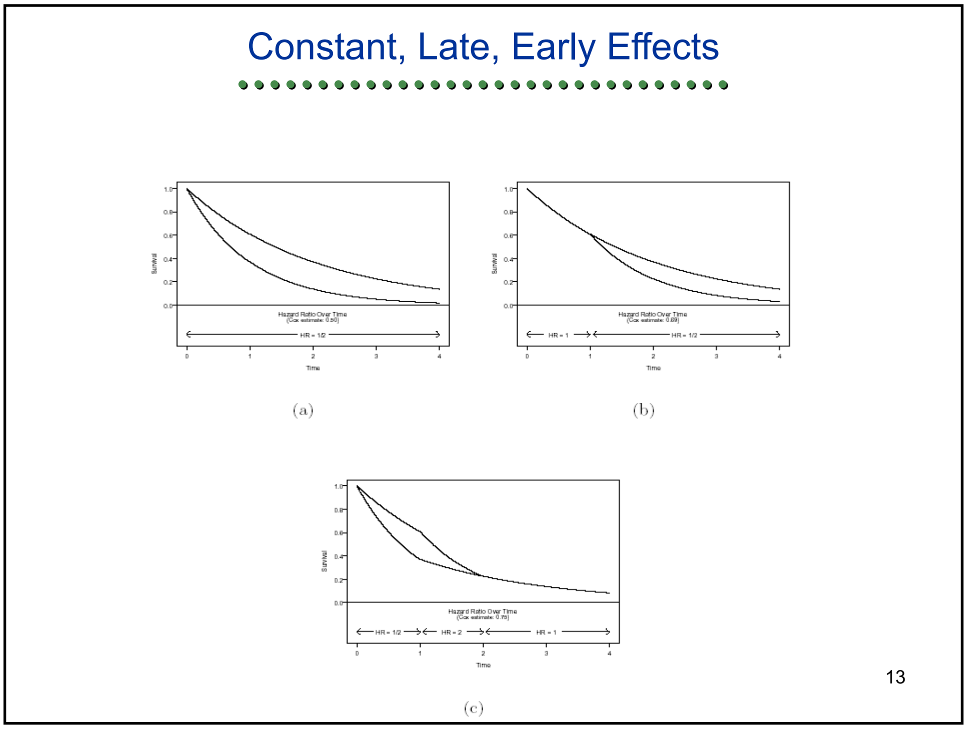


Weighted Logrank Statistics

- Generalization of statistics derived from the proportional hazards setting
- Particularly of interest in the setting of nonproportional hazards
  - Early, transient treatment effects
  - Late treatment effects occurring after some delay

12

12



Constant, Late, Early Effects

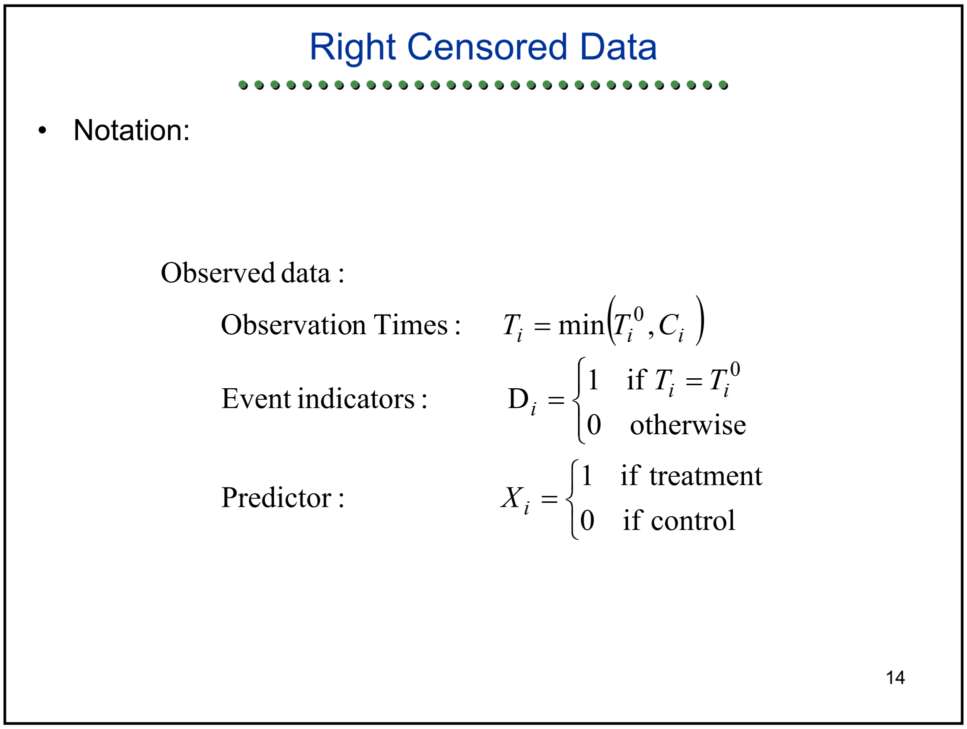
(a)

(b)

(c)

13

13



Right Censored Data

- Notation:

Observed data :

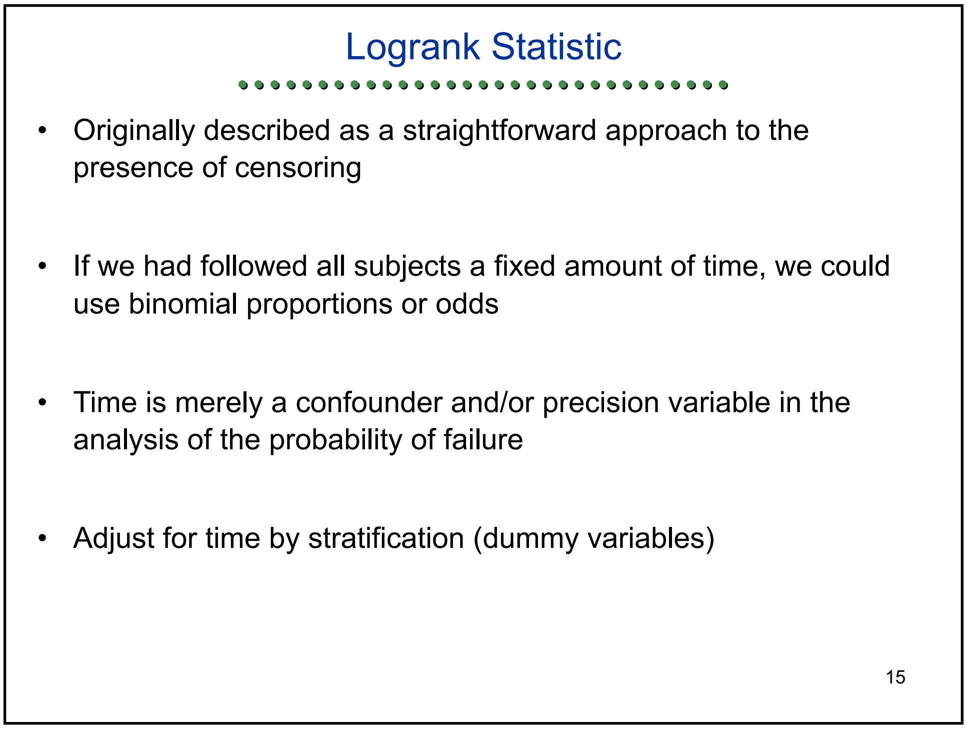
Observation Times : $T_i = \min\left(T_i^0, C_i\right)$

Event indicators : $\mathrm{D}_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

Predictor : $X_i = \begin{cases} 1 & \text{if treatment} \\ 0 & \text{if control} \end{cases}$

14

14



Logrank Statistic

- Originally described as a straightforward approach to the presence of censoring

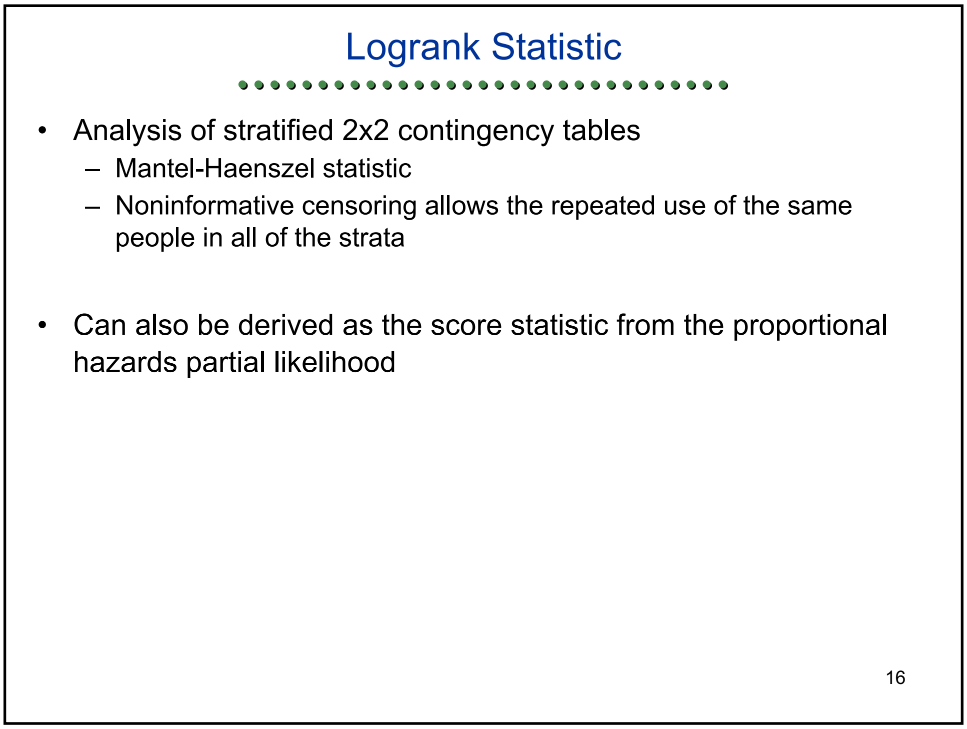
- If we had followed all subjects a fixed amount of time, we could use binomial proportions or odds

- Time is merely a confounder and/or precision variable in the analysis of the probability of failure

- Adjust for time by stratification (dummy variables)

15

15



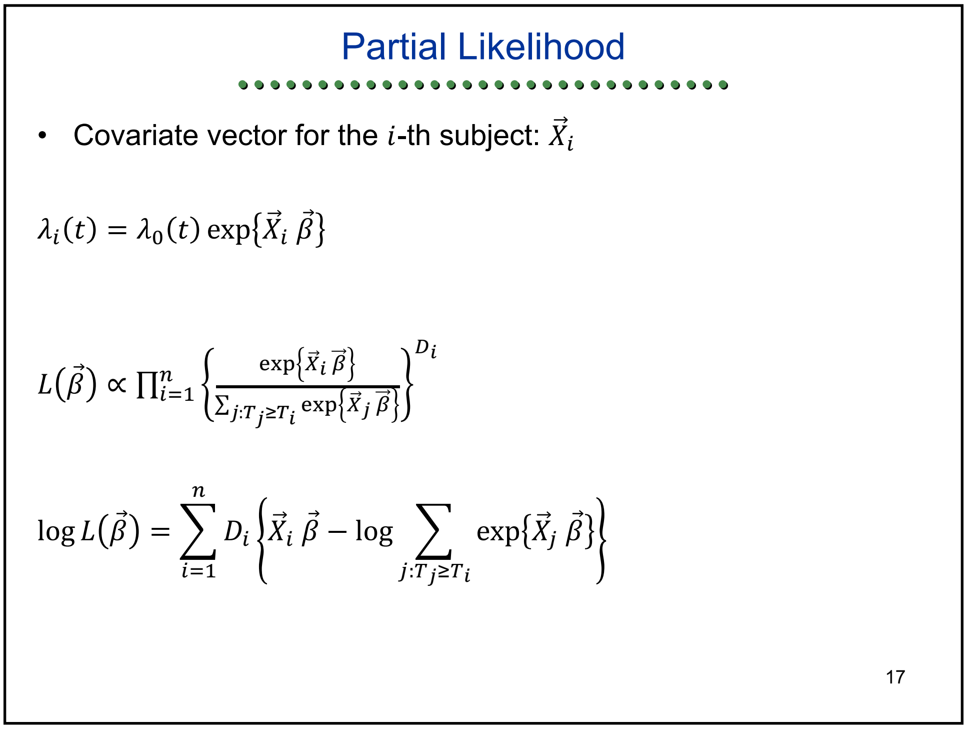
Logrank Statistic

- Analysis of stratified 2x2 contingency tables
  - Mantel-Haenszel statistic
  - Noninformative censoring allows the repeated use of the same people in all of the strata

- Can also be derived as the score statistic from the proportional hazards partial likelihood

16

16



Partial Likelihood

- Covariate vector for the i-th subject: $\vec{X}_i$

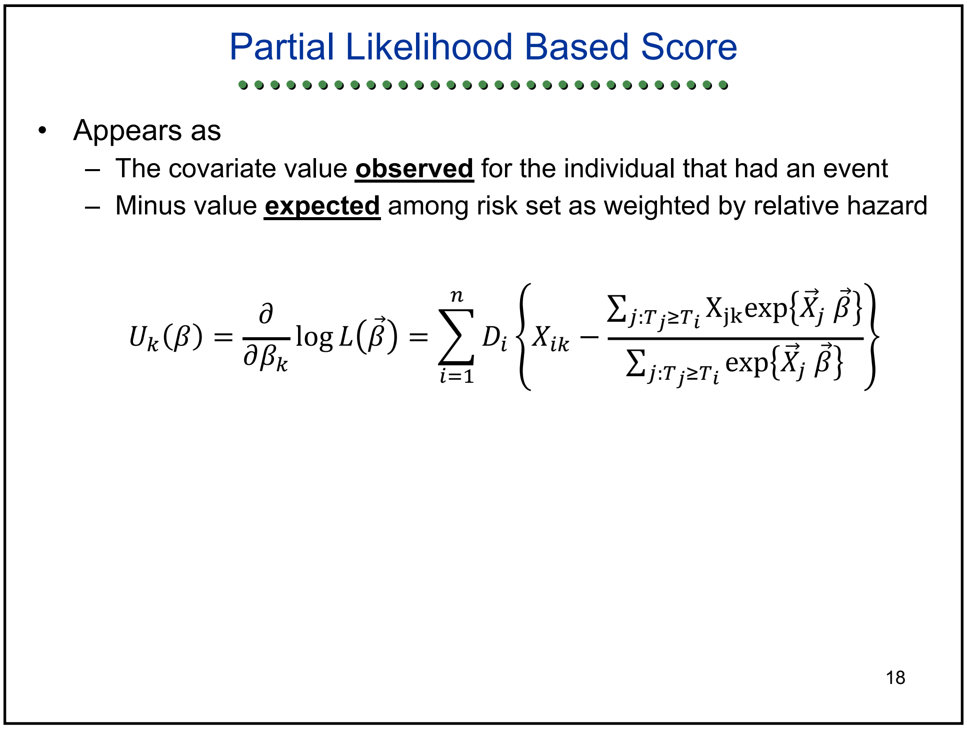
$$\lambda_i(t) = \lambda_0(t)\exp\{\vec{X}_i\,\vec{\beta}\}$$

$$L(\vec{\beta}) \propto \prod_{i=1}^{n}\left\{\frac{\exp\{\vec{X}_i\,\vec{\beta}\}}{\sum_{j:T_j \geq T_i}\exp\{\vec{X}_j\,\vec{\beta}\}}\right\}^{D_i}$$

$$\log L(\vec{\beta}) = \sum_{i=1}^{n} D_i\left\{\vec{X}_i\,\vec{\beta} - \log\sum_{j:T_j \geq T_i}\exp\{\vec{X}_j\,\vec{\beta}\}\right\}$$

17

17



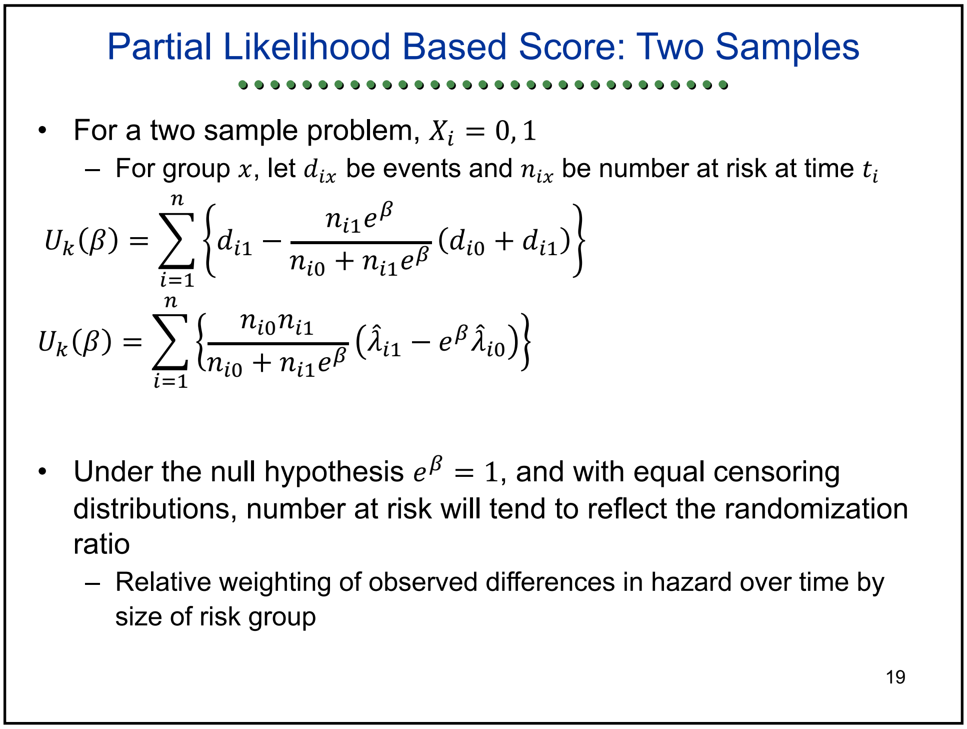
Partial Likelihood Based Score

- Appears as
  - The covariate value **observed** for the individual that had an event
  - Minus value **expected** among risk set as weighted by relative hazard

$$U_k(\beta) = \frac{\partial}{\partial \beta_k}\log L(\vec{\beta}) = \sum_{i=1}^{n} D_i\left\{X_{ik} - \frac{\sum_{j:T_j \geq T_i}\mathrm{X}_{\mathrm{jk}}\exp\{\vec{X}_j\,\vec{\beta}\}}{\sum_{j:T_j \geq T_i}\exp\{\vec{X}_j\,\vec{\beta}\}}\right\}$$

18

18



Partial Likelihood Based Score: Two Samples

- For a two sample problem, $X_i = 0, 1$
  - For group x, let d_{ix} be events and n_{ix} be number at risk at time t_i

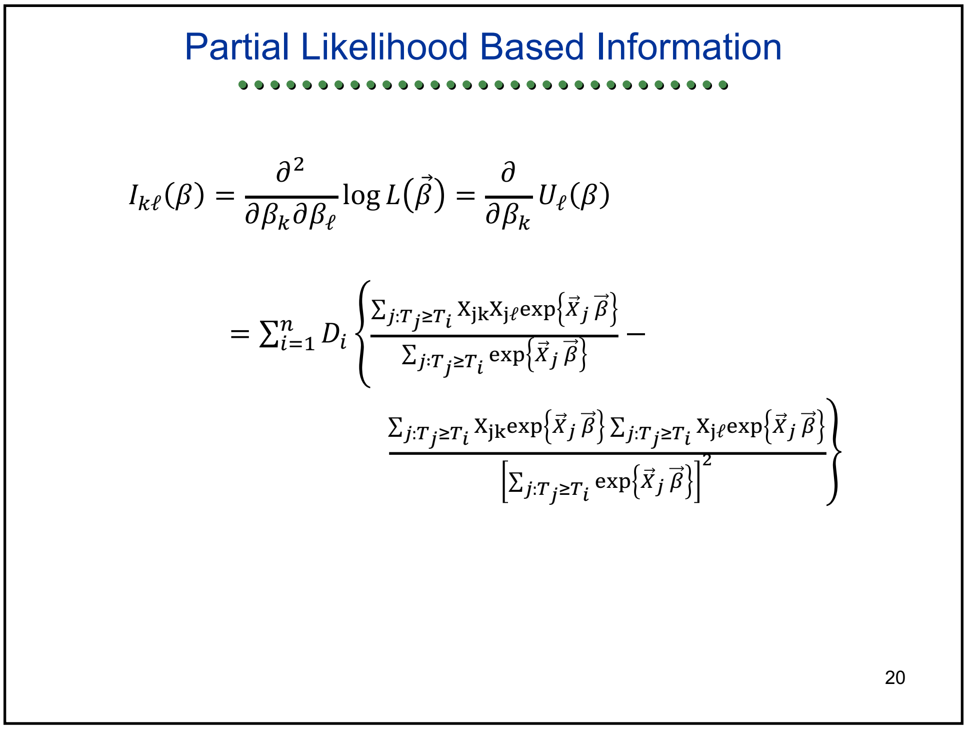
$$U_k(\beta) = \sum_{i=1}^{n} \left\{ d_{i1} - \frac{n_{i1} e^{\beta}}{n_{i0} + n_{i1} e^{\beta}} (d_{i0} + d_{i1}) \right\}$$

$$U_k(\beta) = \sum_{i=1}^{n} \left\{ \frac{n_{i0} n_{i1}}{n_{i0} + n_{i1} e^{\beta}} \left(\hat{\lambda}_{i1} - e^{\beta} \hat{\lambda}_{i0} \right) \right\}$$

- Under the null hypothesis $e^{\beta} = 1$, and with equal censoring distributions, number at risk will tend to reflect the randomization ratio
  - Relative weighting of observed differences in hazard over time by size of risk group

19

19



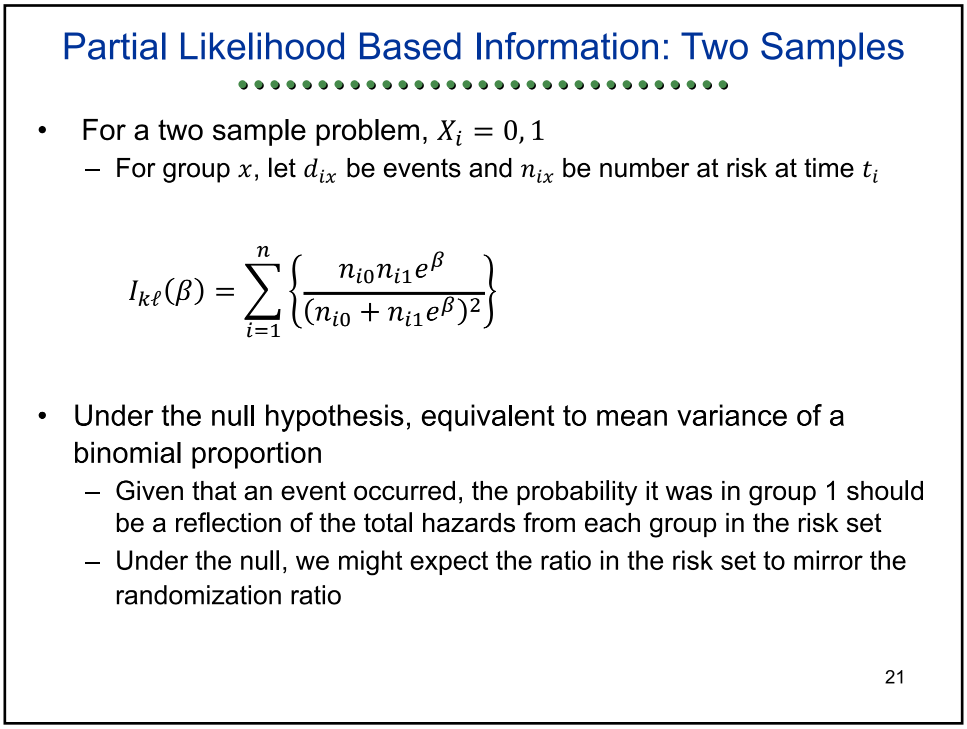
Partial Likelihood Based Information

$$I_{k\ell}(\beta) = \frac{\partial^2}{\partial \beta_k \partial \beta_\ell} \log L(\vec{\beta}) = \frac{\partial}{\partial \beta_k} U_\ell(\beta)$$

$$= \sum_{i=1}^{n} D_i \left\{ \frac{\sum_{j:T_j \geq T_i} \mathrm{X_{jk}} \mathrm{X_{j\ell}} \exp\{\vec{X}_j \vec{\beta}\}}{\sum_{j:T_j \geq T_i} \exp\{\vec{X}_j \vec{\beta}\}} - \frac{\sum_{j:T_j \geq T_i} \mathrm{X_{jk}} \exp\{\vec{X}_j \vec{\beta}\} \sum_{j:T_j \geq T_i} \mathrm{X_{j\ell}} \exp\{\vec{X}_j \vec{\beta}\}}{\left[\sum_{j:T_j \geq T_i} \exp\{\vec{X}_j \vec{\beta}\}\right]^2} \right\}$$

20

20



Partial Likelihood Based Information: Two Samples

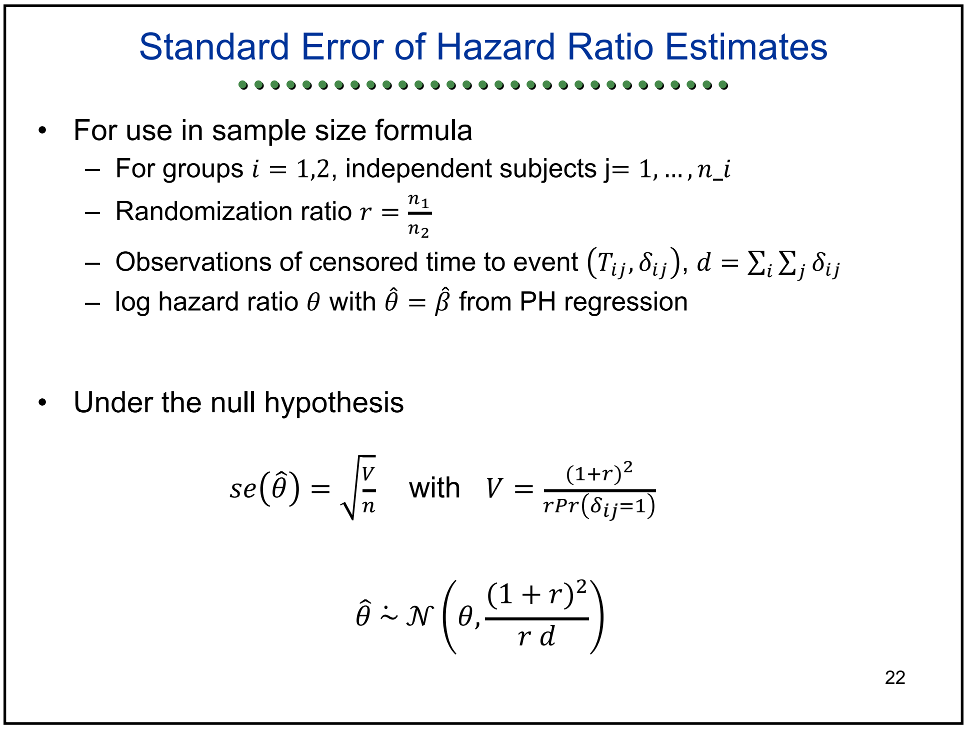
- For a two sample problem, $X_i = 0, 1$
  - For group x, let d_{ix} be events and n_{ix} be number at risk at time t_i

$$I_{k\ell}(\beta) = \sum_{i=1}^{n} \left\{ \frac{n_{i0} n_{i1} e^{\beta}}{(n_{i0} + n_{i1} e^{\beta})^2} \right\}$$

- Under the null hypothesis, equivalent to mean variance of a binomial proportion
  - Given that an event occurred, the probability it was in group 1 should be a reflection of the total hazards from each group in the risk set
  - Under the null, we might expect the ratio in the risk set to mirror the randomization ratio

21

21



Standard Error of Hazard Ratio Estimates

- For use in sample size formula
  - For groups $i = 1{,}2$, independent subjects j= $1, \ldots, n\_i$
  - Randomization ratio $r = \frac{n_1}{n_2}$
  - Observations of censored time to event $\left(T_{ij}, \delta_{ij}\right)$, $d = \sum_i \sum_j \delta_{ij}$
  - log hazard ratio θ with $\hat{\theta} = \hat{\beta}$ from PH regression

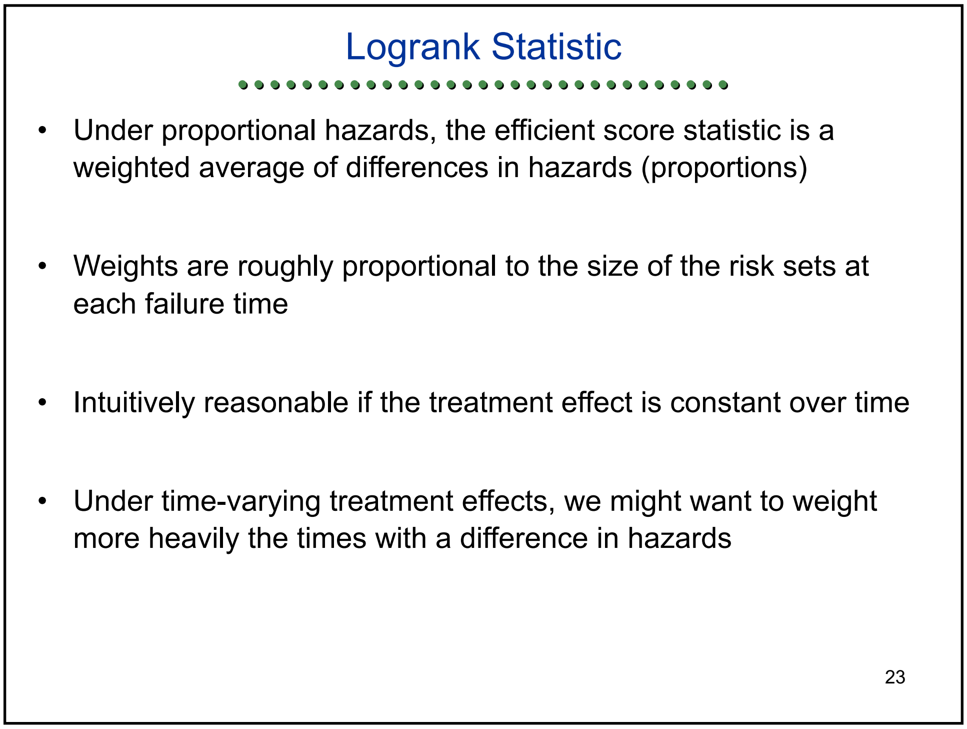
- Under the null hypothesis

$$se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} \quad \text{with} \quad V = \frac{(1+r)^2}{r Pr\left(\delta_{ij}=1\right)}$$

$$\hat{\theta} \mathrel{\dot{\sim}} \mathcal{N}\left(\theta, \frac{(1+r)^2}{r\, d}\right)$$

22

22

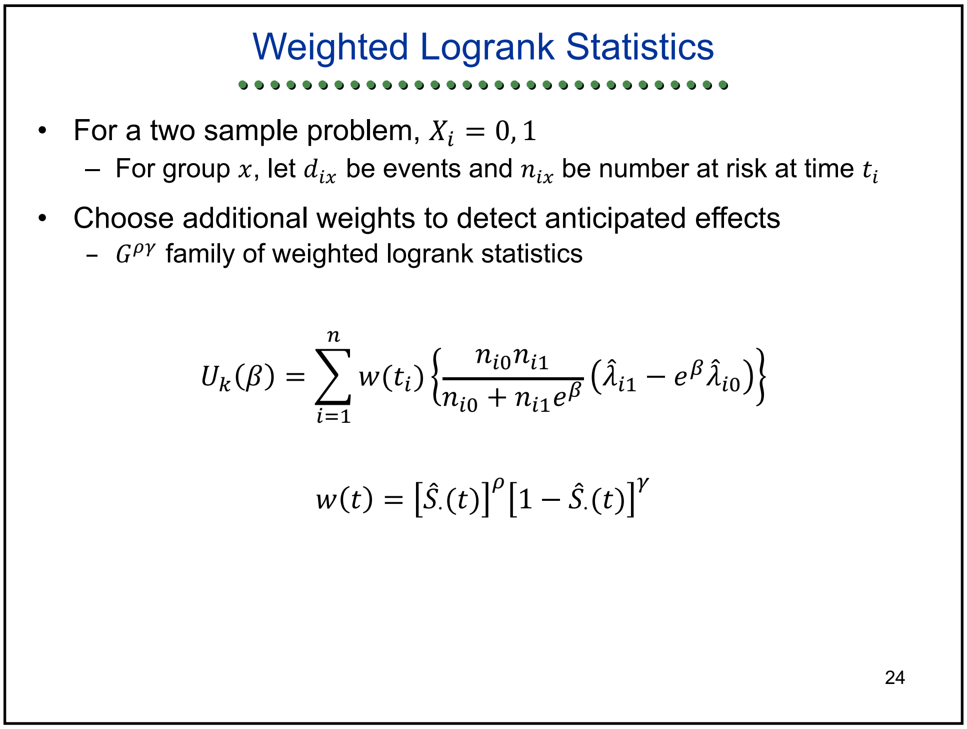


Logrank Statistic

- Under proportional hazards, the efficient score statistic is a weighted average of differences in hazards (proportions)
- Weights are roughly proportional to the size of the risk sets at each failure time
- Intuitively reasonable if the treatment effect is constant over time
- Under time-varying treatment effects, we might want to weight more heavily the times with a difference in hazards

23

23



Weighted Logrank Statistics

- For a two sample problem, $X_i = 0, 1$
  - For group x, let d_{ix} be events and n_{ix} be number at risk at time t_i
- Choose additional weights to detect anticipated effects
  - $G^{\rho\gamma}$ family of weighted logrank statistics

$$U_k(\beta) = \sum_{i=1}^{n} w(t_i)\left\{\frac{n_{i0}n_{i1}}{n_{i0} + n_{i1}e^{\beta}}\left(\hat{\lambda}_{i1} - e^{\beta}\hat{\lambda}_{i0}\right)\right\}$$

$$w(t) = \left[\hat{S}.(t)\right]^{\rho}\left[1 - \hat{S}.(t)\right]^{\gamma}$$

24

24

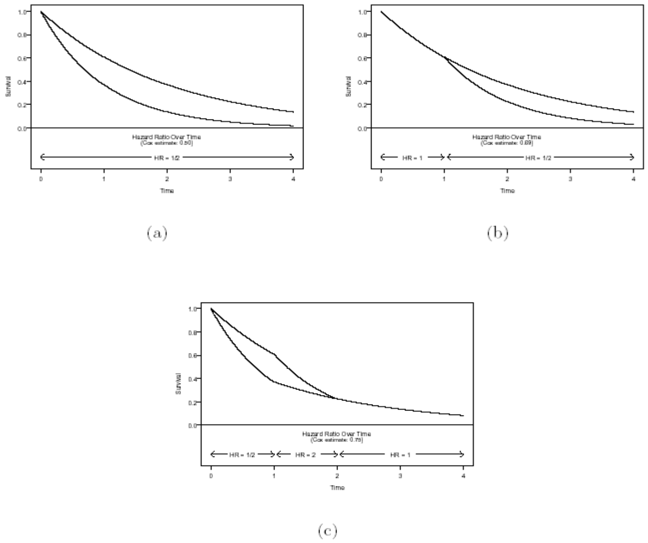
$G^{\rho\gamma}$ Family

- Fleming & Harrington:
  - Logrank statistic: ρ=0; γ=0
  - Wilcoxon statistic: ρ=1; γ=0
    - Weights early differences more heavily
      - “Early” defined relative to survivor function, not time
  - ρ=1; γ=1
    - Places greatest weight between 25th, 75th quantiles
  - ρ=0; γ=1
    - Weights late differences more heavily

25

25

Constant, Late, Early Effects



(a)

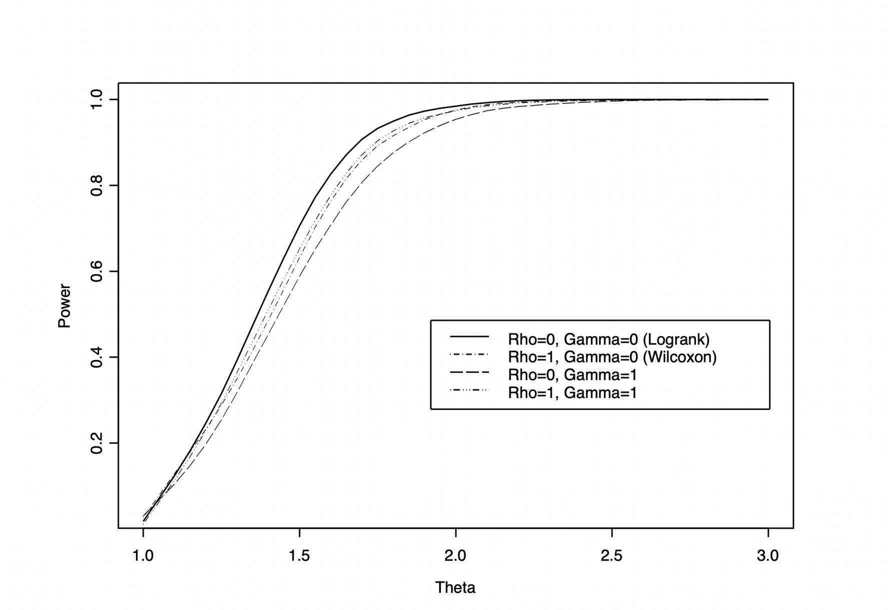
(b)

(c)

26

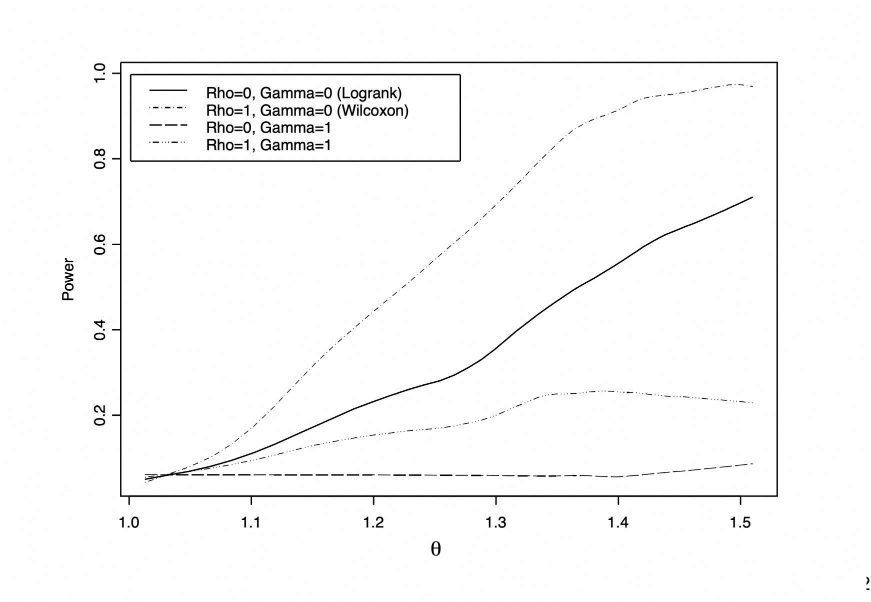
26

Constant (PH) Effects: Power



27

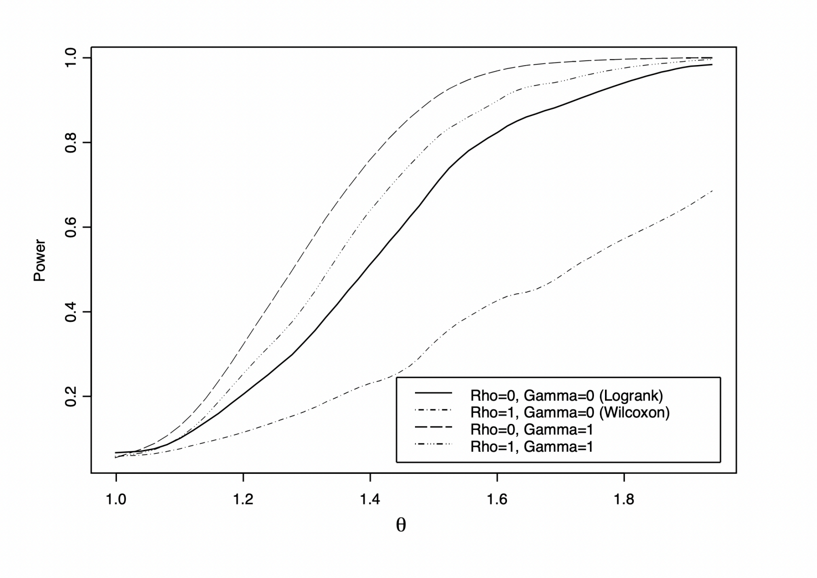
Early Effects: Power



28

28

Late Effects: Power



29

29

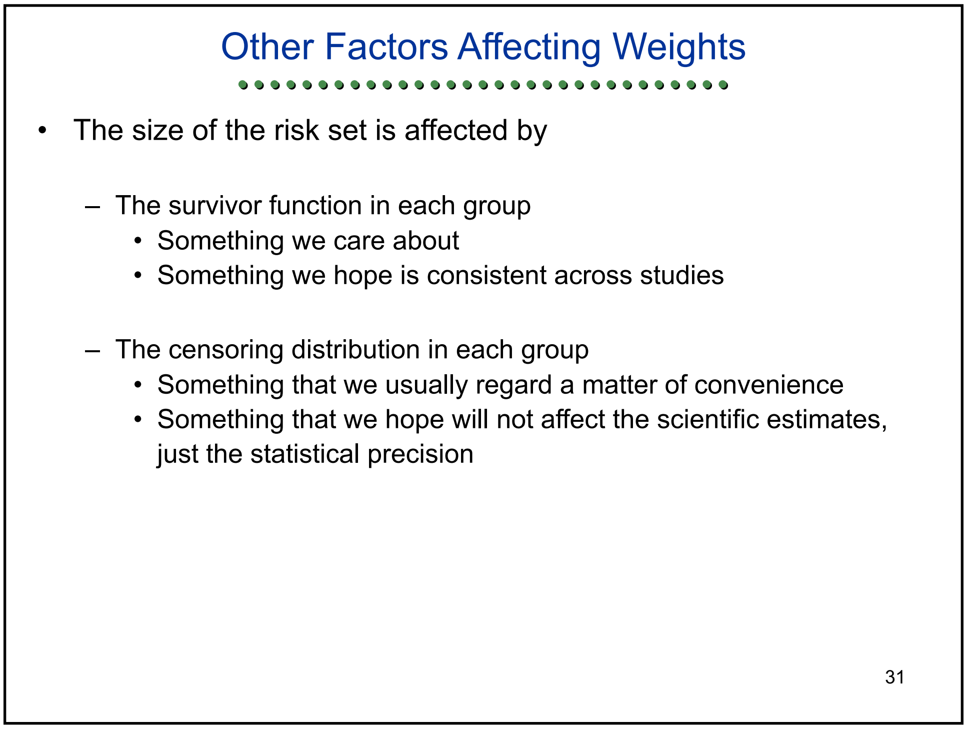
Caveats

- The scientific interpretation of these weighted logrank statistics is difficult in the presence of nonproportional hazards
  - (And why use them when we have PH?)

- The weights we specify are only part of the story
  - The size of the risk sets at each failure time also affects the inference

30

30

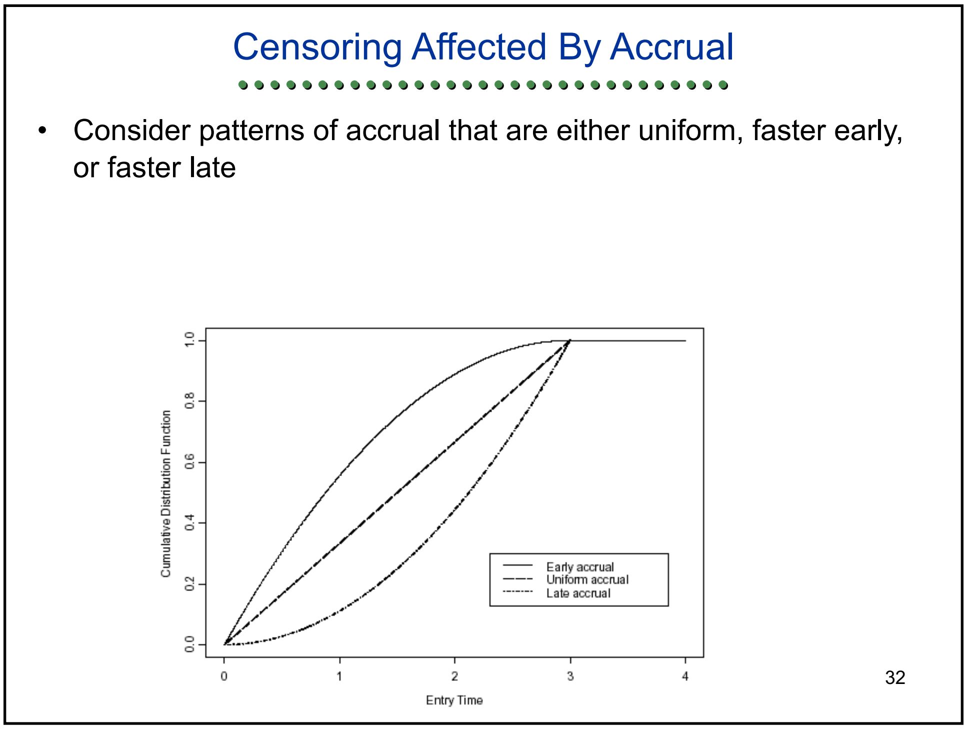


Other Factors Affecting Weights

- The size of the risk set is affected by
  - The survivor function in each group
    - Something we care about
    - Something we hope is consistent across studies
  - The censoring distribution in each group
    - Something that we usually regard a matter of convenience
    - Something that we hope will not affect the scientific estimates, just the statistical precision

31

31



Censoring Affected By Accrual

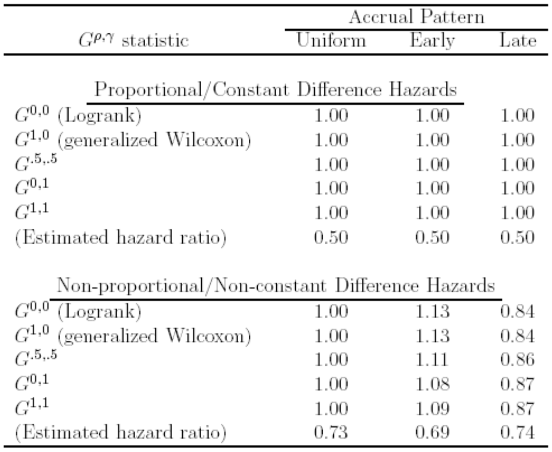
- Consider patterns of accrual that are either uniform, faster early, or faster late

32

32

Inference for PH, Late Tx Effects

- For the same survival curves, different accrual patterns greatly affect the asymptotic behavior of the weighted logrank statistics

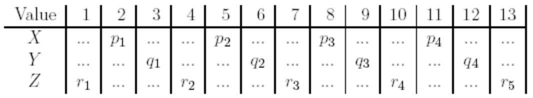


$G^{\rho,\gamma}$ statistic	Accrual Pattern		
	Uniform	Early	Late
Proportional/Constant Difference Hazards			
$G^{0,0}$ (Logrank)	1.00	1.00	1.00
$G^{1,0}$ (generalized Wilcoxon)	1.00	1.00	1.00
$G^{.5,.5}$	1.00	1.00	1.00
$G^{0,1}$	1.00	1.00	1.00
$G^{1,1}$	1.00	1.00	1.00
(Estimated hazard ratio)	0.50	0.50	0.50
Non-proportional/Non-constant Difference Hazards			
$G^{0,0}$ (Logrank)	1.00	1.13	0.84
$G^{1,0}$ (generalized Wilcoxon)	1.00	1.13	0.84
$G^{.5,.5}$	1.00	1.11	0.86
$G^{0,1}$	1.00	1.08	0.87
$G^{1,1}$	1.00	1.09	0.87
(Estimated hazard ratio)	0.73	0.69	0.74

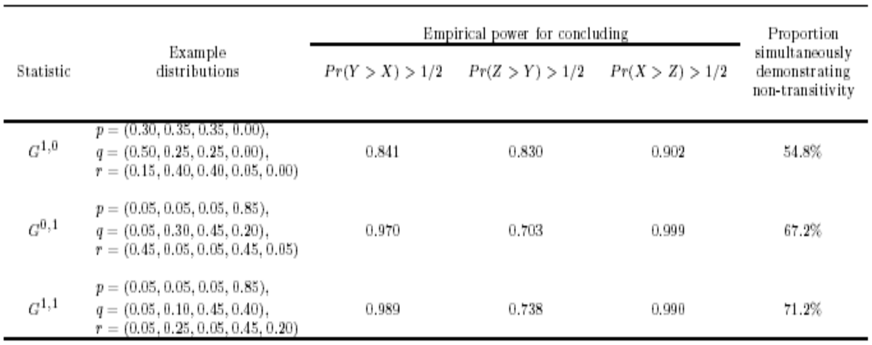
Transitivity

- The weighting scheme used in the weighted logrank statistics also introduces intransitivity to studies
  - (Generally less of an issue with unweighted logrank statistic)
- The weights are stochastically determined from
  - Each group’s survivor function
  - The censoring distribution
- Hence we can obtain A > B > C > A
  - Very distressing to regulatory agencies, if not all scientists

Demonstrating Intransitivity



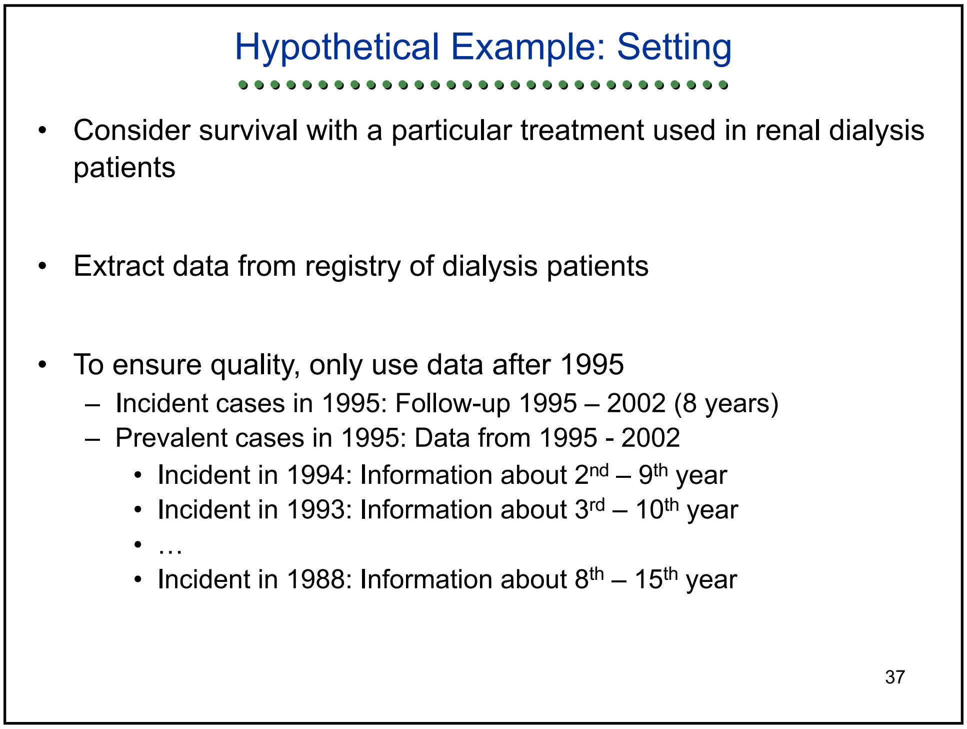
Value	1	2	3	4	5	6	7	8	9	10	11	12	13
X	...	p_1	...	...	p_2	...	...	p_3	...	...	p_4	...	...
Y	...	...	q_1	...	...	q_2	...	...	q_3	...	...	q_4	...
Z	r_1	...	...	r_2	...	...	r_3	...	...	r_4	...	...	r_5



Statistic	Example distributions	Empirical power for concluding			Proportion simultaneously demonstrating non-transitivity
		$Pr(Y > X) > 1/2$	$Pr(Z > Y) > 1/2$	$Pr(X > Z) > 1/2$	
$G^{1,0}$	$p = (0.30, 0.35, 0.35, 0.00)$, $q = (0.50, 0.25, 0.25, 0.00)$, $r = (0.15, 0.40, 0.40, 0.05, 0.00)$	0.841	0.830	0.902	54.8%
$G^{0,1}$	$p = (0.05, 0.05, 0.05, 0.85)$, $q = (0.05, 0.30, 0.45, 0.20)$, $r = (0.45, 0.05, 0.05, 0.45, 0.05)$	0.970	0.703	0.999	67.2%
$G^{1,1}$	$p = (0.05, 0.05, 0.05, 0.85)$, $q = (0.05, 0.10, 0.45, 0.40)$, $r = (0.05, 0.25, 0.05, 0.45, 0.20)$	0.989	0.738	0.990	71.2%

Effect of Censoring on Inference

- The estimates of treatment benefit can vary even more markedly according to the censoring distribution
- With “crossing hazards”, changes in censoring can make any of the weighted logrank statistics qualitatively differ from each other
- And it is possible for the conclusion drawn from the statistic to differ markedly from the conclusion suggested by the survival curves

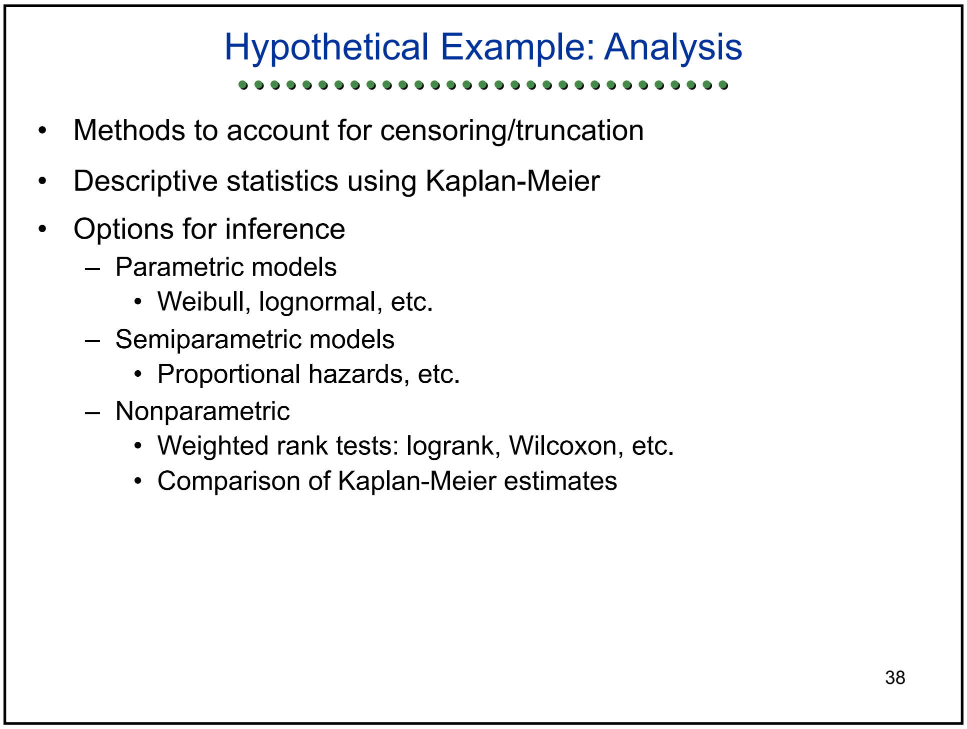


Hypothetical Example: Setting

- Consider survival with a particular treatment used in renal dialysis patients
- Extract data from registry of dialysis patients
- To ensure quality, only use data after 1995
  - Incident cases in 1995: Follow-up 1995 – 2002 (8 years)
  - Prevalent cases in 1995: Data from 1995 - 2002
    - Incident in 1994: Information about 2nd – 9th year
    - Incident in 1993: Information about 3rd – 10th year
    - …
    - Incident in 1988: Information about 8th – 15th year

37

37



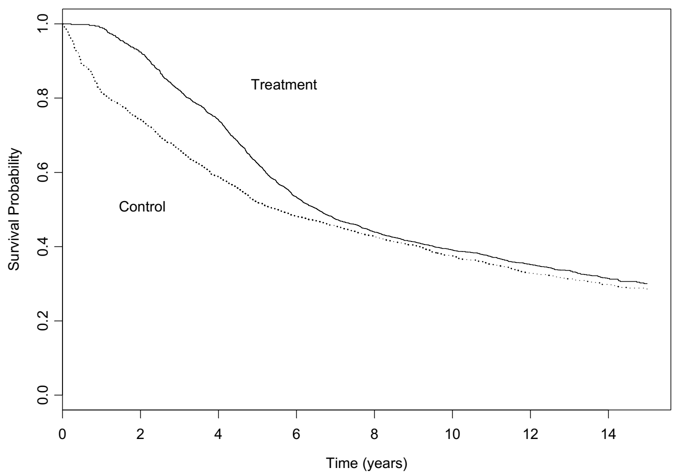
Hypothetical Example: Analysis

- Methods to account for censoring/truncation
- Descriptive statistics using Kaplan-Meier
- Options for inference
  - Parametric models
    - Weibull, lognormal, etc.
  - Semiparametric models
    - Proportional hazards, etc.
  - Nonparametric
    - Weighted rank tests: logrank, Wilcoxon, etc.
    - Comparison of Kaplan-Meier estimates

38

38

Hypothetical Example: KM Curves


Kaplan-Meier Curves for Simulated Data (n=5623)

Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

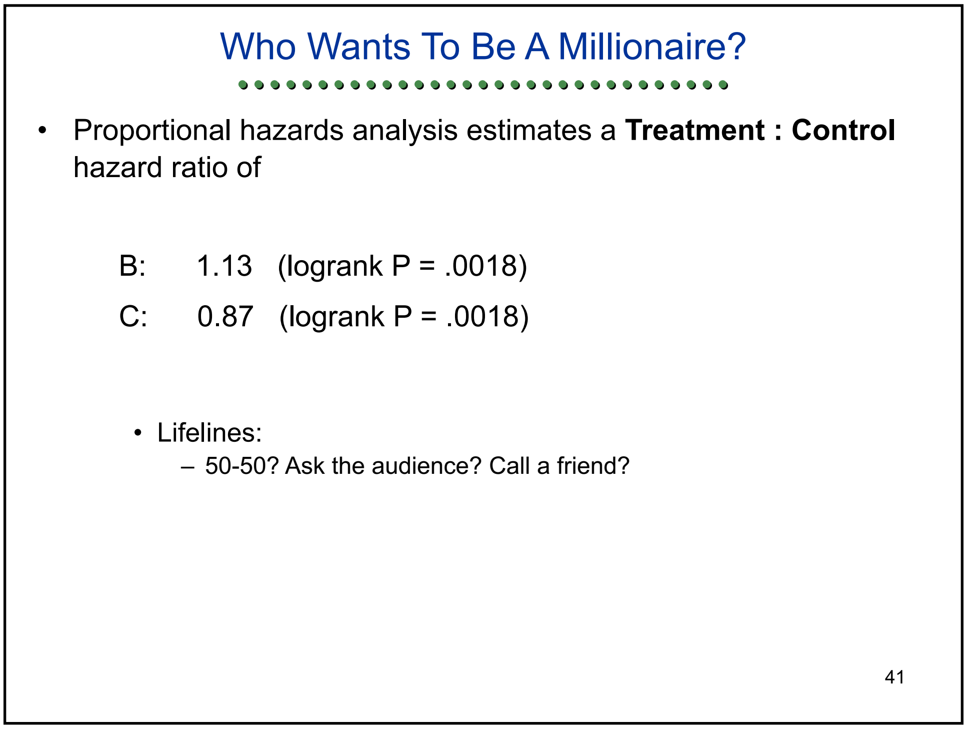
  A: 2.07 (logrank P = .0018)

  B: 1.13 (logrank P = .0018)

  C: 0.87 (logrank P = .0018)

  D: 0.48 (logrank P = .0018)

  - Lifelines:
    - 50-50? Ask the audience? Call a friend?



Who Wants To Be A Millionaire?

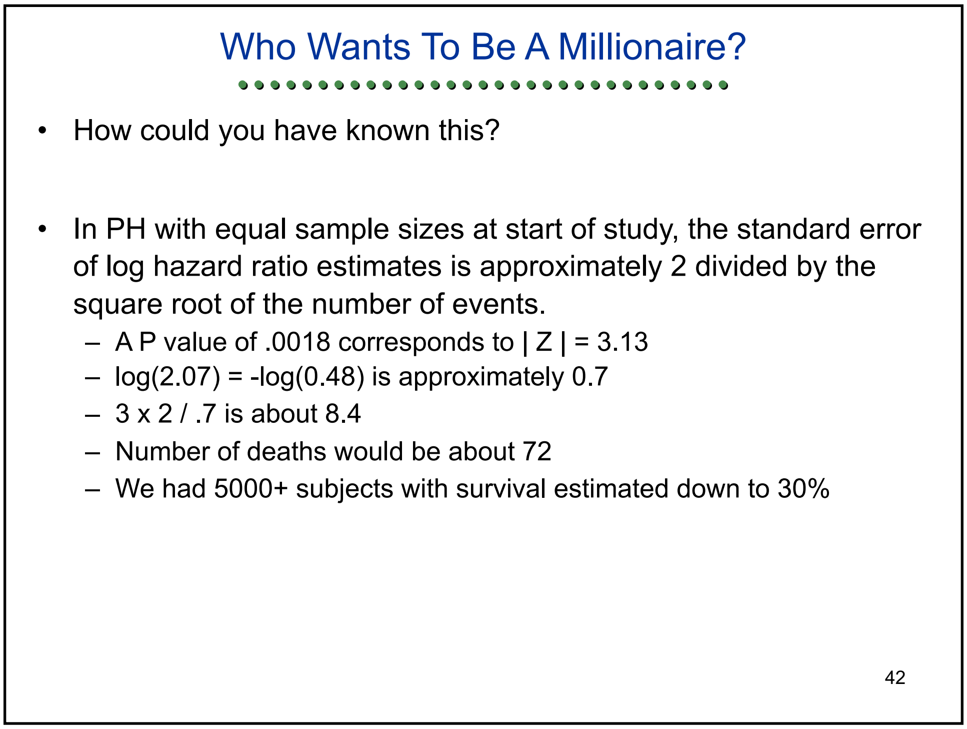
- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

B:	1.13	(logrank P = .0018)
C:	0.87	(logrank P = .0018)

- Lifelines:
  - 50-50? Ask the audience? Call a friend?

41

41



Who Wants To Be A Millionaire?

- How could you have known this?

- In PH with equal sample sizes at start of study, the standard error of log hazard ratio estimates is approximately 2 divided by the square root of the number of events.
  - A P value of .0018 corresponds to | Z | = 3.13
  - log(2.07) = -log(0.48) is approximately 0.7
  - 3 x 2 / .7 is about 8.4
  - Number of deaths would be about 72
  - We had 5000+ subjects with survival estimated down to 30%

42

42

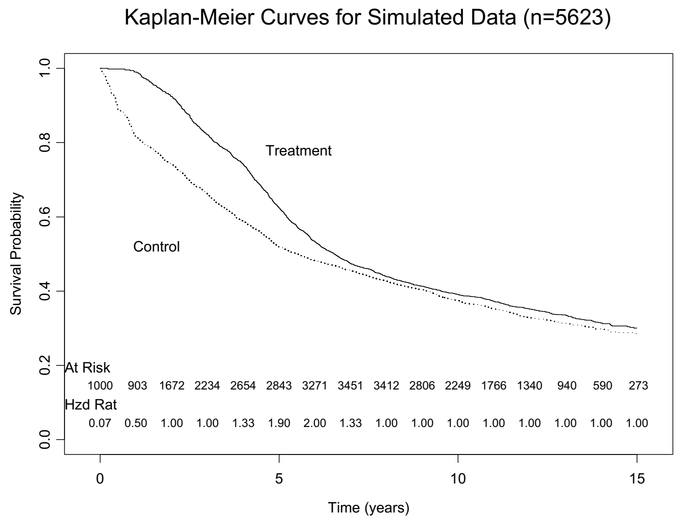
Who Wants To Be A Millionaire?

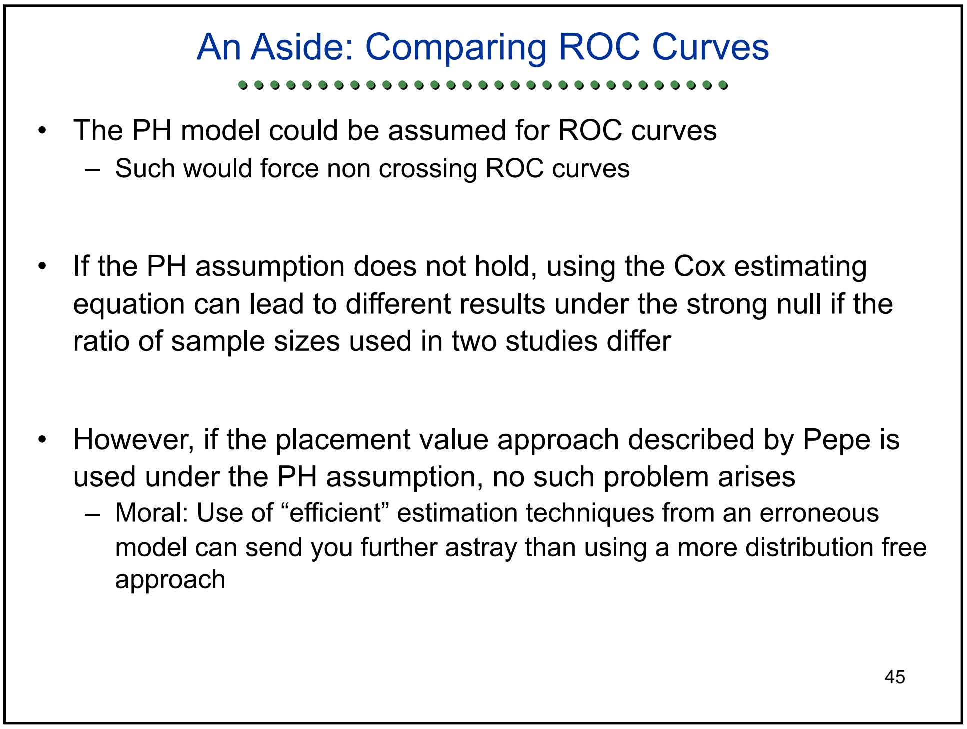
- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  B: 1.13 (logrank P = .0018)

- The weighting using the risk sets made no scientific sense
  - Statistical precision to estimate a meaningless quantity is meaningless
- This happened due to left entry.
  - In a RCT, we would have monotonically decreasing risk sets

Hypothetical Example: KM Curves





An Aside: Comparing ROC Curves

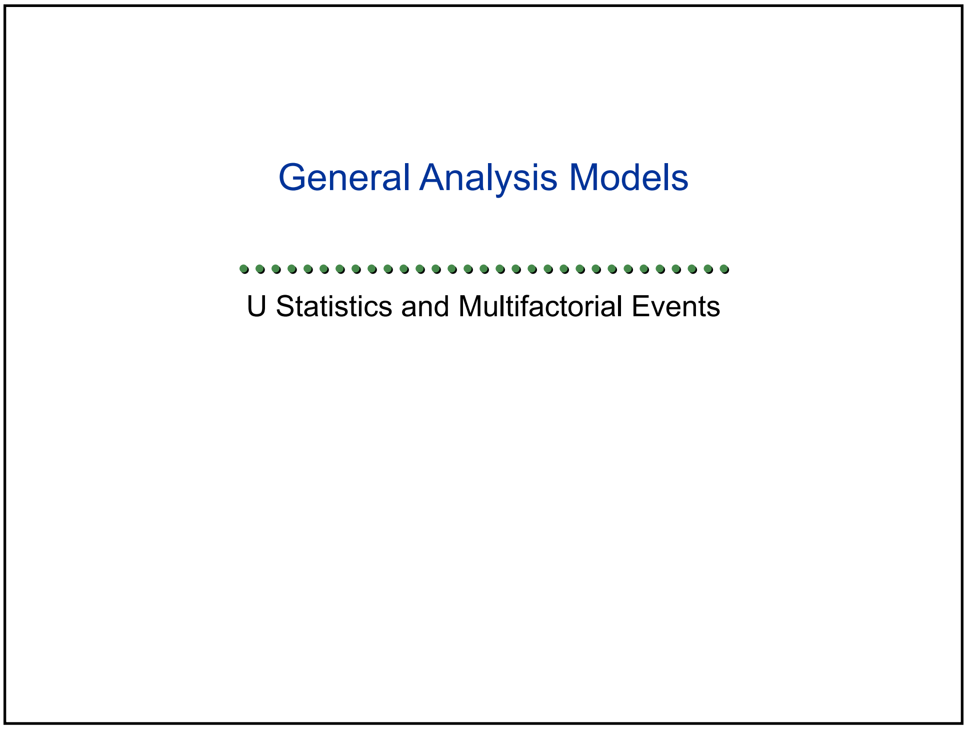
- The PH model could be assumed for ROC curves
  - Such would force non crossing ROC curves

- If the PH assumption does not hold, using the Cox estimating equation can lead to different results under the strong null if the ratio of sample sizes used in two studies differ

- However, if the placement value approach described by Pepe is used under the PH assumption, no such problem arises
  - Moral: Use of “efficient” estimation techniques from an erroneous model can send you further astray than using a more distribution free approach

45

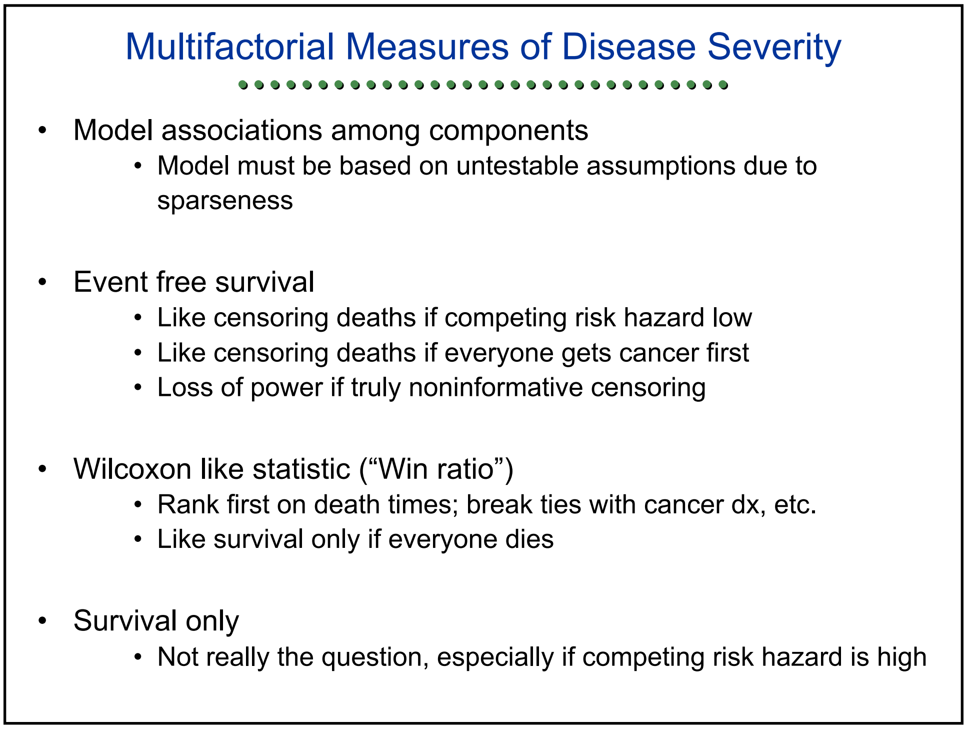
45



General Analysis Models

U Statistics and Multifactorial Events

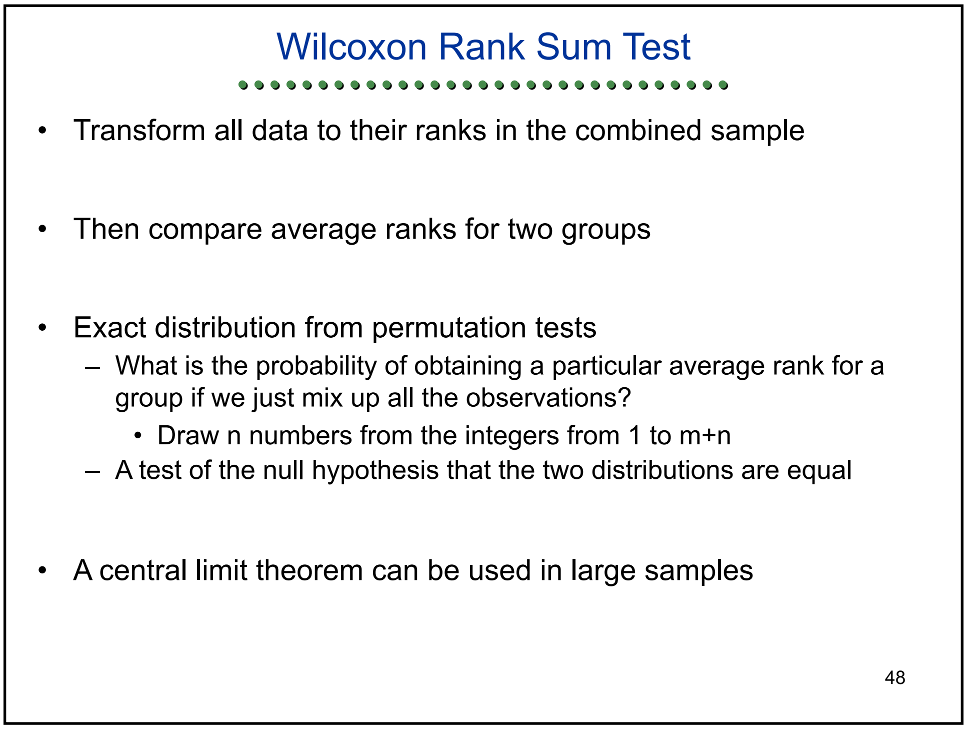
46



Multifactorial Measures of Disease Severity

- Model associations among components
  - Model must be based on untestable assumptions due to sparseness
- Event free survival
  - Like censoring deaths if competing risk hazard low
  - Like censoring deaths if everyone gets cancer first
  - Loss of power if truly noninformative censoring
- Wilcoxon like statistic (“Win ratio”)
  - Rank first on death times; break ties with cancer dx, etc.
  - Like survival only if everyone dies
- Survival only
  - Not really the question, especially if competing risk hazard is high

47

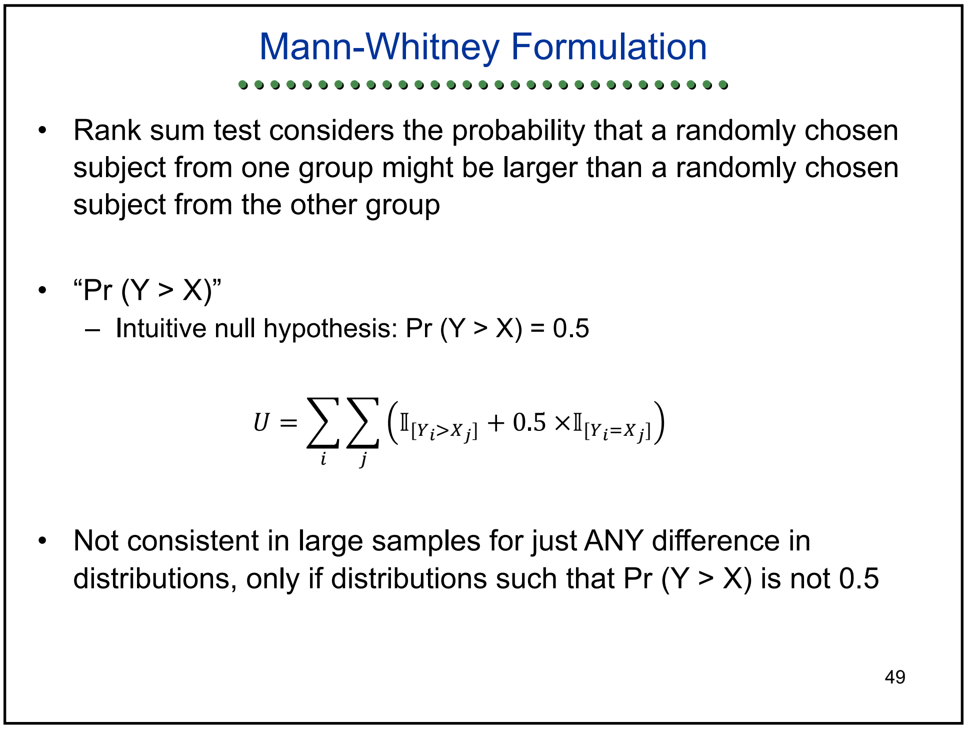


Wilcoxon Rank Sum Test

- Transform all data to their ranks in the combined sample
- Then compare average ranks for two groups
- Exact distribution from permutation tests
  - What is the probability of obtaining a particular average rank for a group if we just mix up all the observations?
    - Draw n numbers from the integers from 1 to m+n
  - A test of the null hypothesis that the two distributions are equal
- A central limit theorem can be used in large samples

48

48



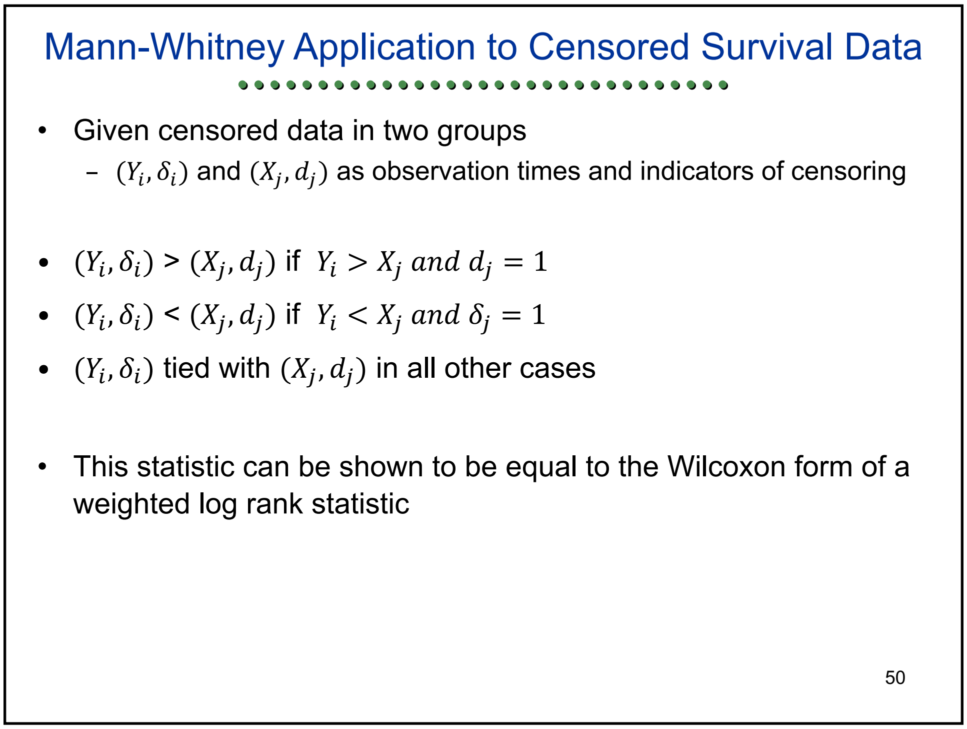
Mann-Whitney Formulation

- Rank sum test considers the probability that a randomly chosen subject from one group might be larger than a randomly chosen subject from the other group

- “Pr (Y > X)”
  - Intuitive null hypothesis: Pr (Y > X) = 0.5

$$U = \sum_i \sum_j \left(\mathbb{I}_{[Y_i > X_j]} + 0.5 \times \mathbb{I}_{[Y_i = X_j]} \right)$$

- Not consistent in large samples for just ANY difference in distributions, only if distributions such that Pr (Y > X) is not 0.5



Mann-Whitney Application to Censored Survival Data

- Given censored data in two groups
  - (Y_i, δ_i) and (X_j, d_j) as observation times and indicators of censoring

- (Y_i, δ_i) > (X_j, d_j) if $Y_i > X_j \; and \; d_j = 1$
- (Y_i, δ_i) < (X_j, d_j) if $Y_i < X_j \; and \; \delta_j = 1$
- (Y_i, δ_i) tied with (X_j, d_j) in all other cases

- This statistic can be shown to be equal to the Wilcoxon form of a weighted log rank statistic

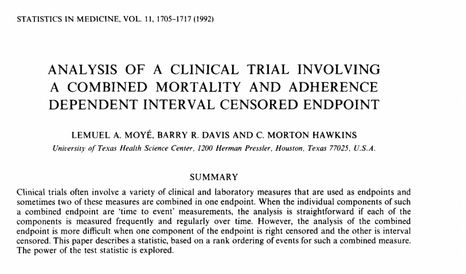
Mann-Whitney Application to Censored Survival Data

- Given censored data in two groups
  - (Y_i, δ_i) and (X_j, d_j) as observation times and indicators of censoring
- (Y_i, δ_i) > (X_j, d_j) if $Y_i > X_j \; and \; d_j = 1$
- (Y_i, δ_i) < (X_j, d_j) if $Y_i < X_j \; and \; \delta_j = 1$
- (Y_i, δ_i) tied with (X_j, d_j) in all other cases
- Then compute U statistic using this definition for ordering
- This statistic can be shown to be equal to the Wilcoxon form of a weighted log rank statistic

51

51

Extensions to Multiple Endpoints



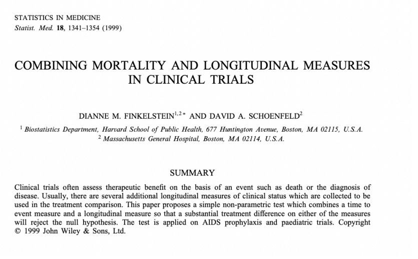
STATISTICS IN MEDICINE, VOL. 11, 1705–1717 (1992)

ANALYSIS OF A CLINICAL TRIAL INVOLVING
A COMBINED MORTALITY AND ADHERENCE
DEPENDENT INTERVAL CENSORED ENDPOINT

LEMUEL A. MOYÉ, BARRY R. DAVIS AND C. MORTON HAWKINS
University of Texas Health Science Center, 1200 Herman Pressler, Houston, Texas 77025, U.S.A.

SUMMARY

Clinical trials often involve a variety of clinical and laboratory measures that are used as endpoints and sometimes two of these measures are combined in one endpoint. When the individual components of such a combined endpoint are 'time to event' measurements, the analysis is straightforward if each of the components is measured frequently and regularly over time. However, the analysis of the combined endpoint is more difficult when one component of the endpoint is right censored and the other is interval censored. This paper describes a statistic, based on a rank ordering of events for such a combined measure. The power of the test statistic is explored.



STATISTICS IN MEDICINE
Statist. Med. **18**, 1341–1354 (1999)

COMBINING MORTALITY AND LONGITUDINAL MEASURES
IN CLINICAL TRIALS

DIANNE M. FINKELSTEIN[1,2]* AND DAVID A. SCHOENFELD[2]
[1] *Biostatistics Department, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.*
[2] *Massachusetts General Hospital, Boston, MA 02114, U.S.A.*

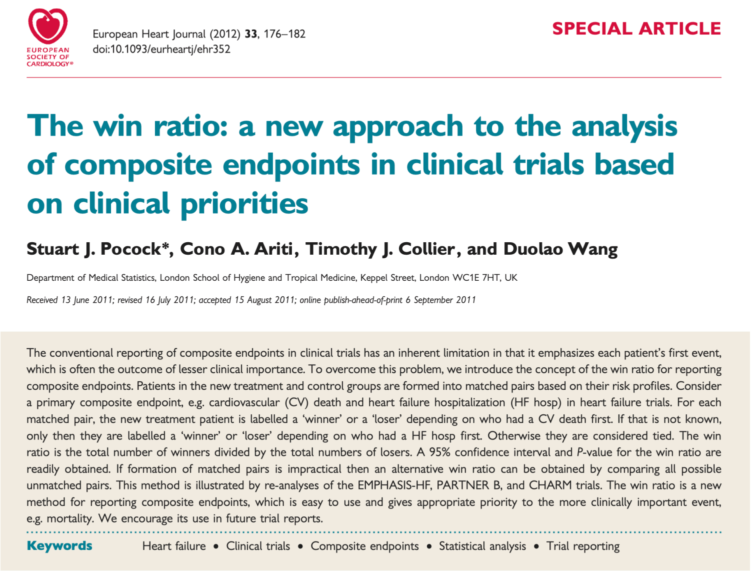
SUMMARY

Clinical trials often assess therapeutic benefit on the basis of an event such as death or the diagnosis of disease. Usually, there are several additional longitudinal measures of clinical status which are collected to be used in the treatment comparison. This paper proposes a simple non-parametric test which combines a time to event measure and a longitudinal measure so that a substantial treatment difference on either of the measures will reject the null hypothesis. The test is applied on AIDS prophylaxis and paediatric trials. Copyright © 1999 John Wiley & Sons, Ltd.

52

52

Extensions to Covariate Adjustment



European Heart Journal (2012) **33**, 176–182
doi:10.1093/eurheartj/ehr352

SPECIAL ARTICLE

The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities

Stuart J. Pocock*, Cono A. Ariti, Timothy J. Collier, and Duolao Wang

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

Received 13 June 2011; revised 16 July 2011; accepted 15 August 2011; online publish-ahead-of-print 6 September 2011

The conventional reporting of composite endpoints in clinical trials has an inherent limitation in that it emphasizes each patient's first event, which is often the outcome of lesser clinical importance. To overcome this problem, we introduce the concept of the win ratio for reporting composite endpoints. Patients in the new treatment and control groups are formed into matched pairs based on their risk profiles. Consider a primary composite endpoint, e.g. cardiovascular (CV) death and heart failure hospitalization (HF hosp) in heart failure trials. For each matched pair, the new treatment patient is labelled a 'winner' or a 'loser' depending on who had a CV death first. If that is not known, only then they are labelled a 'winner' or 'loser' depending on who had a HF hosp first. Otherwise they are considered tied. The win ratio is the total number of winners divided by the total numbers of losers. A 95% confidence interval and *P*-value for the win ratio are readily obtained. If formation of matched pairs is impractical then an alternative win ratio can be obtained by comparing all possible unmatched pairs. This method is illustrated by re-analyses of the EMPHASIS-HF, PARTNER B, and CHARM trials. The win ratio is a new method for reporting composite endpoints, which is easy to use and gives appropriate priority to the more clinically important event, e.g. mortality. We encourage its use in future trial reports.

Keywords Heart failure • Clinical trials • Composite endpoints • Statistical analysis • Trial reporting

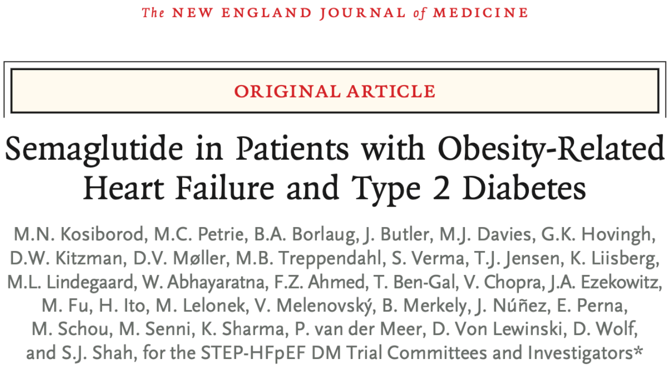
Basic Idea of Extensions: Tie Breakers

- Given possibly censored or longitudinal data vectors in two groups: $\overrightarrow{\mathbb{Y}_i}, \overrightarrow{\mathbb{X}_j}$ where components might be
  - (Y_{ik}, δ_{ik}) and (X_{jk}, d_{jk}) as observation times, indicators of censoring
  - $(Y_{ik}(t), t_{ik})$ and $(X_{jk}(s), s_{jk})$ as longitudinal processes measured up to specified times
- For every pair to be compared, evaluate first on the component highest in the hierarchy
- In the event of ties, go to the next component in the hierarchy, etc.
- Can either do all pairs, or first stratify subjects according to prognostic variables
  - Similar to van Elteren statistic

Basic Idea of Extensions: Tie Breakers

- Given any pair of observation, one from each group, we order
  - Censored survival times
    - $(Y_i, \delta_i) > (X_j, d_j)$ if $Y_i > X_j \; and \; d_j = 1$
    - $(Y_i, \delta_i) < (X_j, d_j)$ if $Y_i < X_j \; and \; \delta_j = 1$
    - (Y_i, δ_i) tied with (X_j, d_j) in all other cases
  - Longitudinal processes
    - $(Y_{ik}(t), t_{ik}) > (X_{jk}(s), s_{jk})$ by judging process up to min(t_{ik}, s_{jk})
    - $(Y_{ik}(t), t_{ik}) < (X_{jk}(s), s_{jk})$ by judging process up to min(t_{ik}, s_{jk})
    - tied in all other cases

Example: STEP-HFpEF (semaglutide



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Semaglutide in Patients with Obesity-Related Heart Failure and Type 2 Diabetes

M.N. Kosiborod, M.C. Petrie, B.A. Borlaug, J. Butler, M.J. Davies, G.K. Hovingh, D.W. Kitzman, D.V. Møller, M.B. Treppendahl, S. Verma, T.J. Jensen, K. Liisberg, M.L. Lindegaard, W. Abhayaratna, F.Z. Ahmed, T. Ben-Gal, V. Chopra, J.A. Ezekowitz, M. Fu, H. Ito, M. Lelonek, V. Melenovský, B. Merkely, J. Núñez, E. Perna, M. Schou, M. Senni, K. Sharma, P. van der Meer, D. Von Lewinski, D. Wolf, and S.J. Shah, for the STEP-HFpEF DM Trial Committees and Investigators*

Example: STEP-HFpEF (semaglutide

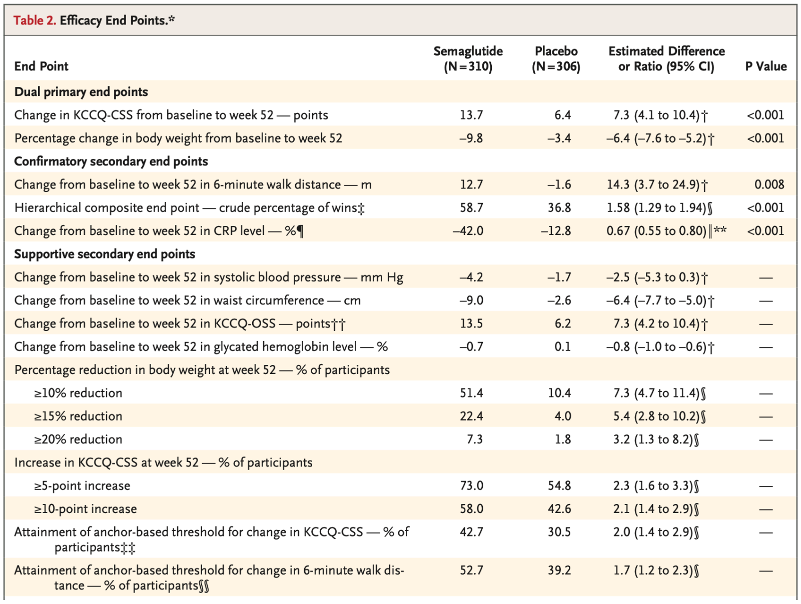


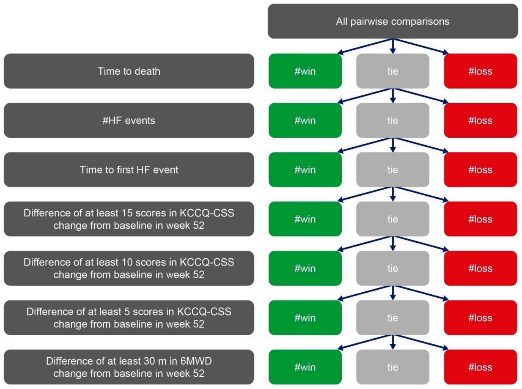
Table 2. **Efficacy End Points.***

End Point	Semaglutide (N=310)	Placebo (N=306)	Estimated Difference or Ratio (95% CI)	P Value
Dual primary end points				
Change in KCCQ-CSS from baseline to week 52 — points	13.7	6.4	7.3 (4.1 to 10.4)†	<0.001
Percentage change in body weight from baseline to week 52	−9.8	−3.4	−6.4 (−7.6 to −5.2)†	<0.001
Confirmatory secondary end points				
Change from baseline to week 52 in 6-minute walk distance — m	12.7	−1.6	14.3 (3.7 to 24.9)†	0.008
Hierarchical composite end point — crude percentage of wins‡	58.7	36.8	1.58 (1.29 to 1.94)§	<0.001
Change from baseline to week 52 in CRP level — %¶	−42.0	−12.8	0.67 (0.55 to 0.80)‖**	<0.001
Supportive secondary end points				
Change from baseline to week 52 in systolic blood pressure — mm Hg	−4.2	−1.7	−2.5 (−5.3 to 0.3)†	—
Change from baseline to week 52 in waist circumference — cm	−9.0	−2.6	−6.4 (−7.7 to −5.0)†	—
Change from baseline to week 52 in KCCQ-OSS — points††	13.5	6.2	7.3 (4.2 to 10.4)†	—
Change from baseline to week 52 in glycated hemoglobin level — %	−0.7	0.1	−0.8 (−1.0 to −0.6)†	—
Percentage reduction in body weight at week 52 — % of participants				
≥10% reduction	51.4	10.4	7.3 (4.7 to 11.4)§	—
≥15% reduction	22.4	4.0	5.4 (2.8 to 10.2)§	—
≥20% reduction	7.3	1.8	3.2 (1.3 to 8.2)§	—
Increase in KCCQ-CSS at week 52 — % of participants				
≥5-point increase	73.0	54.8	2.3 (1.6 to 3.3)§	—
≥10-point increase	58.0	42.6	2.1 (1.4 to 2.9)§	—
Attainment of anchor-based threshold for change in KCCQ-CSS — % of participants‡‡	42.7	30.5	2.0 (1.4 to 2.9)§	—
Attainment of anchor-based threshold for change in 6-minute walk distance — % of participants§§	52.7	39.2	1.7 (1.2 to 2.3)§	—

57

Example: STEP-HFpEF (semaglutide)

Supplemental Figure 1. Hierarchy of testing for the hierarchical composite endpoint


All pairwise comparisons
Time to death
#win
tie
#loss
#HF events
#win
tie
#loss
Time to first HF event
#win
tie
#loss
Difference of at least 15 scores in KCCQ-CSS change from baseline in week 52
#win
tie
#loss
Difference of at least 10 scores in KCCQ-CSS change from baseline in week 52
#win
tie
#loss
Difference of at least 5 scores in KCCQ-CSS change from baseline in week 52
#win
tie
#loss
Difference of at least 30 m in 6MWD change from baseline in week 52
#win
tie
#loss

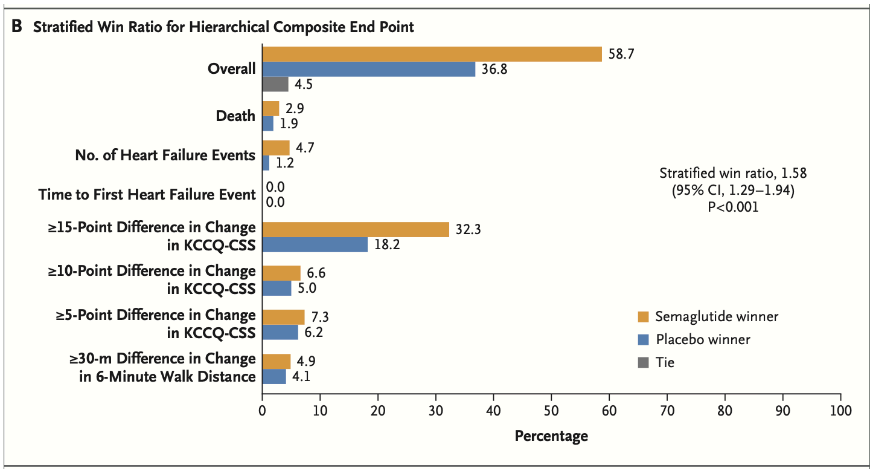
Analysis of the hierarchical composite endpoint will be based on direct comparisons of each participant randomized to semaglutide versus each participant randomized to placebo within each stratum. For each of these participant pairs, a 'treatment winner' based on similar observation time will be declared based on the endpoint hierarchy (as shown in the figure).

Abbreviations: 6MWD, 6-minute walk distance; HF, heart failure; KCCQ-CSS, Kansas City Cardiomyopathy Questionnaire clinical summary score.

58

58

Example: STEP-HFpEF (semaglutide



59

59

Additional Comments

- The Wilcoxon rank sum test can be shown to be “intransitive”
  - It is possible to simultaneously decide that
    - Group A tends to be higher than Group B
    - Group B tends to be higher than Group C
    - Group C tends to be higher than Group A
  - Arises because Pr (Y > X) is intransitive
- By adding in a great many other variables into the hierarchy and perhaps having different ”censoring” or “sampling” distribution for each component, the generalizability across studies is even more difficult
- It is not at all clear to me how one would judge clinical importance

60

60