Survival Analysis: Analysis of Right Censored Time to Event Data

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

May 1, 2004

Two Sample Inference

The Setting

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.

Survival Analysis Methods

Most commonly used methods

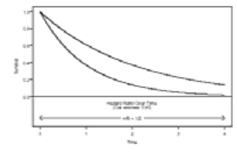
- Parametric
  - Accelerated failure time regression models
- Semiparametric
  - Proportional hazards regression models
- Nonparametric
  - Kaplan-Meier curves
  - Weighted logrank statistics

Weighted Logrank Statistics

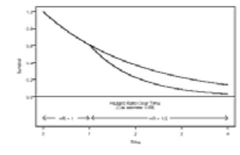
Generalization of statistics derived from the proportional hazards setting

- Particularly of interest in the setting of nonproportional hazards
  - Early, transient treatment effects
  - Late treatment effects occurring after some delay

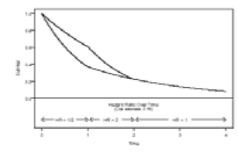
Constant, Late, Early Effects



(a)



(b)



(c)

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

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)

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

Partial Likelihood

$$\lambda_i(t) = \lambda_0(t)\exp\{X_i\beta\}$$

$$L(\beta) = \prod_{i=1}^{n}\left[\frac{\exp\{X_i\beta\}}{\sum_{j:T_j\ge T_i}\exp\{X_j\beta\}}\right]^{D_i}$$

$$\log L(\beta) = \sum_{i=1}^{n} D_i\left[X_i\beta - \log\sum_{j:T_j\ge T_i}\exp\{X_j\beta\}\right]$$

Partial Likelihood Based Score

$$U(\beta) = \frac{\partial}{\partial\beta}\log L(\beta) = \sum_{i=1}^{n} D_i\left[X_i - \frac{\sum_{j:T_j\ge T_i} X_j\exp\{X_j\beta\}}{\sum_{j:T_j\ge T_i}\exp\{X_j\beta\}}\right]$$

$$= \sum_t\left[d_{1t} - \frac{n_{1t}e^{\beta}}{n_{0t}+n_{1t}e^{\beta}}(d_{0t}+d_{1t})\right]$$

$$= \sum_t \frac{n_{0t}n_{1t}}{n_{0t}+n_{1t}}\left[\hat\lambda_{1t} - e^{\beta}\hat\lambda_{0t}\right]$$

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

Weighted Logrank Statistics

Choose additional weights to detect anticipated effects

$$W(\beta) = \sum_t w(t) \frac{n_{0t} n_{1t}}{n_{0t} + n_{1t}} \left[\hat{\lambda}_{1t} - e^{\beta} \hat{\lambda}_{0t} \right]$$

$G^{\rho\gamma}$ Family of weighted logrank statistics :

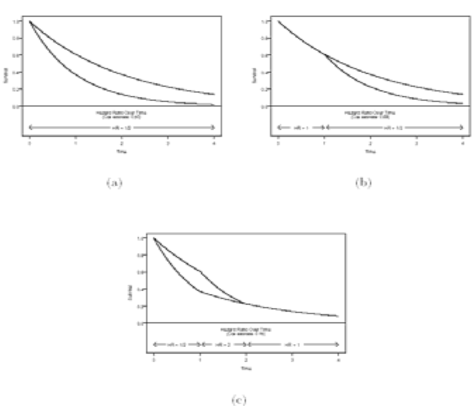
$$w(t) = \left[\hat{S}_{\bullet}(t) \right]^{\rho} \left[1 - \hat{S}_{\bullet}(t) \right]^{\gamma}$$

$G^{\rho\gamma}$ Family

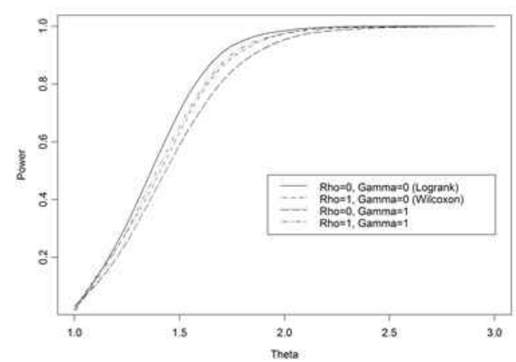
Fleming & Harrington:

- 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

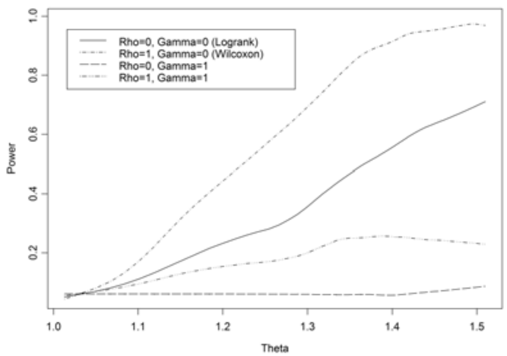
Constant, Late, Early Effects



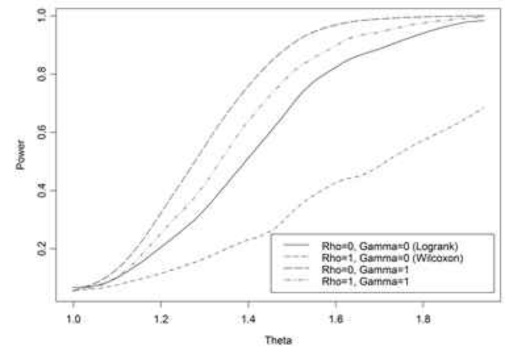
Constant (PH) Effects: Power



Early Effects: Power



Late Effects: Power



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

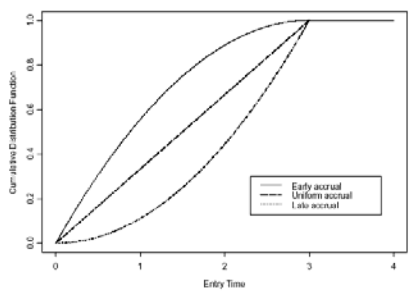
Other Factors Affecting Weights

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

Censoring Affected By Accrual

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



Inference for PH, Late Tx Effects

$G^{\rho,\gamma}$ 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

Effect of Censoring on Inference

The estimates of treatment benefit can vary even more markedly according to the censoring distribution

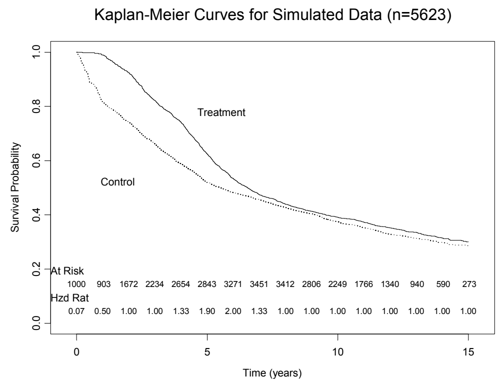
- With “crossing hazards”, changes in censoring can make any of the weighted logrank statistics qualitatively differ from each other
  - And it is possible for the conclusion drawn from the statistic to differ markedly from the conclusion suggested by the survival curves

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

Hypothetical Example: KM Curves



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

Transitivity

The weighting scheme used in the weighted logrank statistics also introduces intransitivity to studies

- 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%

Sequential Clinical Trials

Overview

Clinical Trials

Experimentation in human volunteers

- Efficacy: Can the treatment alter the disease process in a beneficial way?
  - Phase II (preliminary); Phase III
- Safety: Are there adverse effects that clearly outweigh any potential benefit?
  - Phase I; Phase II
- Effectiveness: Would adoption of the treatment as a standard affect morbidity / mortality in the population?
  - Phase III (therapy); Phase IV (prevention)

Collaboration of Multiple Disciplines

Discipline	Collaborators	Issues
Scientific	Epidemiologists Basic Scientists Clinical Scientists	Hypothesis generation Mechanisms Clinical benefit
Clinical	Experts in disease / treatment Experts in complications	Efficacy of treatment Adverse experiences
Ethical	Ethicists	Individual ethics Group ethics
Economic	Health services Sponsor management Sponsor marketers	Cost effectiveness Cost of trial / Profitability Marketing appeal
Governmental	Regulators	Safety Efficacy
Statistical	Biostatisticians	Estimates of treatment effect Precision of estimates
Operational	Study coordinators Data management	Collection of data / Study burden Data integrity

Statistical Planning

Ensure that the trial will satisfy the various collaborators as much as possible

- Discriminate between relevant scientific hypotheses
  - Scientific and statistical credibility
- Protect economic interests of sponsor
  - Efficient designs; Economically important estimates
- Protect interests of patients on trial
  - Stop if unsafe or unethical and when credible decision can be made
- Promote rapid discovery of new beneficial treatments

Address Variability

At the end of the study

- Estimate of the treatment effect
  - Single best estimate
  - Precision of estimates
- Decision for or against hypotheses
  - Binary decision
  - Quantification of strength of evidence

Statistical Design: Sampling Plan

Ethical and efficiency concerns are addressed through sampling which might allow early stopping

- During the conduct of the study, data are analyzed at periodic intervals and reviewed by the DMC
- Using interim estimates of treatment effect
  - Decide whether to continue the trial
  - If continuing, decide on any modifications to sampling scheme

Sampling Plan

- Perform analyses at sample sizes $N_1 \ldots N_J$
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T(X_1 \ldots X_{N_J})$
  - Stop if $T < a_j$ (extremely low)
  - Stop if $b_j < T < c_j$ (approximate equivalence)
  - Stop if $T > d_j$ (extremely high)
  - Otherwise continue (with possible adaptive modification of analysis schedule, sample size, etc.)

Sampling Plan

- Issues when using a sequential sampling plan
  - Design stage
    - Boundaries to satisfy desired operating characteristics
      - » E.g., type I error, power, sample size requirements
  - Monitoring stage
    - Flexible implementation of the stopping rule to account for assumptions made at design stage
      - » E.g., sample size adjustment to account for observed variance
  - Analysis stage
    - Providing inference based on true sampling distribution of test statistics

Sampling Plan: Examples

Alternative plans for a sepsis trial comparing 28 day mortality rates with 90% power to detect a 7% improvement using N=1700

- Fixed sample study:
  - Gather data on 1700 patients and analyze data
- Group sequential study (OBF efficacy, P=0.8 futility):
  - Perform analysis after 425 patients
  - If test statistic very low or very high, stop
  - If test statistic intermediate, accrue another 425
  - Repeat, as necessary, until maximum of 1700 patients

Sampling Plan: Examples

Advantage of stopping rule:

- Fixed sample: 4.18% improvement is significant
  - Harmful: Power= 0.001; Average N= 1700
  - No effect: Power= 0.025; Average N= 1700
  - Low effect: Power= 0.500; Average N= 1700
  - Beneficial: Power= 0.975; Average N= 1700
- Grp sequential: 4.24% improvement is significant
  - Harmful: Power= 0.001; Average N= 785
  - No effect: Power= 0.025; Average N= 987
  - Low effect: Power= 0.477; Average N= 1330
  - Beneficial: Power= 0.966; Average N= 1104

Major Issue: Frequentist Inference

Often, the criteria for judging statistical evidence in clinical trial results are based on frequentist criteria

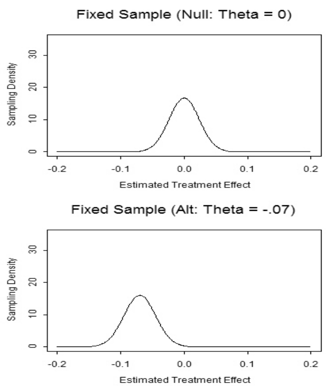
- Experimentwise error probabilities
  - Type I, II errors, power
- Optimality of point estimates
  - Bias, mean squared error
- Computation of precision
  - Confidence intervals

Major Issue: Frequentist Inference

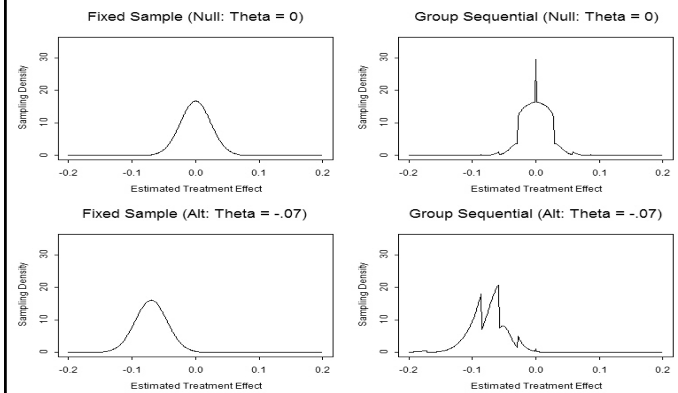
Frequentist operating characteristics are based on the sampling distribution

- Stopping rules do affect the sampling distribution of the usual statistics
  - MLEs are not normally distributed
  - Z scores are not standard normal under the null
    - (1.96 is irrelevant)
  - The null distribution of fixed sample P values is not uniform
    - (They are not true P values)

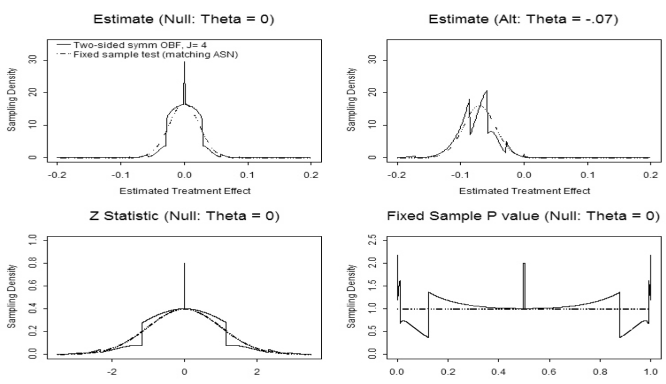
Sampling Distribution of Estimates



Sampling Distribution of Estimates



Sampling with Stopping Rules



Operating Characteristics

For any stopping rule, however, we can compute the correct sampling distribution with specialized software

- From the computed sampling distributions we then compute
  - Bias adjusted estimates
  - Correct (adjusted) confidence intervals
  - Correct (adjusted) P values
- Candidate designs can then be compared with respect to their operating characteristics

Stopping Criteria: Boundary Scales

Various test statistics are transformations

- A stopping rule for one test statistic is easily transformed to a rule for another statistic
  - "Group sequential stopping rules"
    - Sum of observations
    - Point estimate of treatment effect
    - Normalized (Z) statistic
    - Fixed sample P value
    - Error spending function
  - Conditional probability
  - Predictive probability
  - Bayesian posterior probability

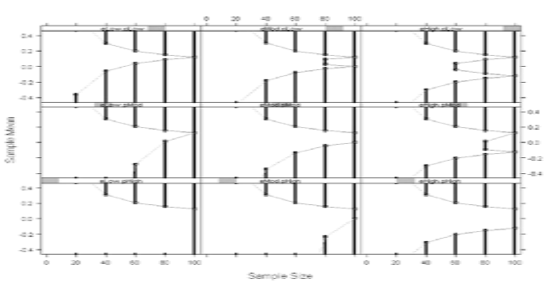
Unified Family: MLE Scale

Boundary shape function unifies families of stopping rules

- Wang & Tsiatis (1987) based families
  - O'Brien & Fleming (1979); Pocock (1977)
  - Also used by Emerson & Fleming (1989); Pampallona & Tsiatis (1994)
- Triangular test (Whitehead, 1983)
- Seq cond probability ratio test (Xiong & Tan, 1994)
- Conditional or predictive power
- Peto-Haybittle (using Burington & Emerson, 2003)

Conditions for Early Stopping

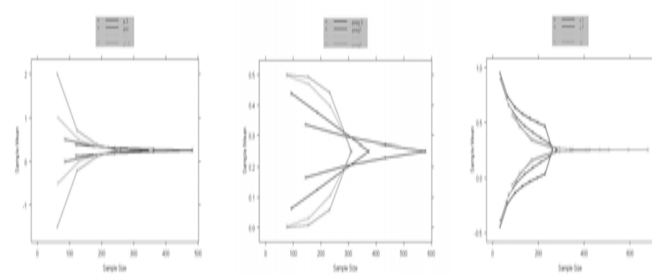
- Down columns: Early vs no early stopping
- Across rows: One-sided vs two-sided



Boundary Shape Functions

A wide variety of boundary shapes possible

- All of the rules depicted have the same type I error and power to detect the alternative



Evaluation of Designs

Process of choosing a trial design

- Define candidate design
  - Usually constrain two operating characteristics
    - Type I error, power at design alternative
    - Type I error, maximal sample size
- Evaluate other operating characteristics
  - Different criteria of interest to different investigators
- Modify design
- Iterate

Operating Characteristics

Generally the same for all stopping rule s

- Frequentist power curve
  - Type I error (null) and power (design alternative)
- Sample size requirements
  - Maximum, average, median, other quantiles
  - Stopping probabilities
- Inference at study termination (at each boundary)
  - Frequentist inference
  - Bayesian inference under spectrum of priors
- Futility measures
  - Conditional power, predictive power

Sequential Clinical Trials

Time Varying Treatment Effects

Time Invariant Treatment Effects

The design, monitoring, and analysis of sequential trials is fairly well established for treatment effects that do not vary over time

- Means
- Proportions
- Odds
- Proportional hazards

Nonproportional Hazards

With nonproportional hazards, new issues must be addressed

- Choice of summary measure
  - Handling any dependence on the censoring distribution
- Definition of alternative
- Computation of operating characteristics
- Flexible implementation

Censoring Distribution

A summary measure that depends on the censoring distribution is the biggest problem

- In a survival study, we typically have a different censoring distribution at successive analyses
- Hence, different summary measures are being tested at different analyses

Weighted Logrank Statistics

This is particularly true with weighted logrank statistics

- At the final analysis, weights will be applied over a wider range of time than is possible at earlier analyses
- At the earlier analyses, early results are weighted more heavily than they will be later

Example

A 7 year trial is planned using a weighted logrank statistic to place weight late

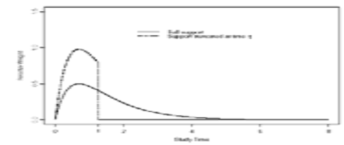
- Plan:
  - 1/28, 2/28, 3/28, …, 7/28 weight over the 7 years
- An interim analysis conducted after 3 years
  - 1/6, 2/6, 3/6 over the first three years
    - (later years have no data, hence no weights)

One Proposed Solution

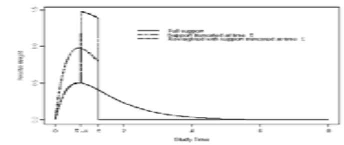
Apply weights due to be used late in study to the most longterm experience

- In the example, we would apply weights
  - 1/28, 2/28, 25/28
- Tends to (appropriately) inflate variability of statistic at interim analyses
- Intuitively reasonable in that the results for the longest observations should be more indicative of the future
  - Similar to imputing future observations

Reassigning weights



(a) Relative weight by time



(b) Reweighted relative weight by time

Analysis Time (yrs)	Proportionate Information	Null Hypothesis ($S_0 = S_1$) $G^{1,1}$ (s.e.)	Null Hypothesis ($S_0 = S_1$) Reweighted $G^{1,1}$ (s.e.)	Alternative (Figure 5.1) $G^{1,1}$ (s.e.)	Alternative (Figure 5.1) Reweighted $G^{1,1}$ (s.e.)
		Non-staggered Entry			
0.50	0.12	-0.005 (0.048)	-0.016 (0.125)	0.005 (0.048)	0.012 (0.125)
1.00	0.42	-0.007 (0.096)	0.020 (0.223)	-0.064 (0.089)	-0.221 (0.196)
1.50	0.67	-0.010 (0.129)	-0.024 (0.239)	-0.209 (0.113)	-0.350 (0.197)
2.00	0.84	-0.010 (0.146)	0.005 (0.207)	-0.297 (0.126)	-0.373 (0.175)
2.50	0.93	-0.003 (0.154)	0.006 (0.176)	-0.346 (0.133)	-0.383 (0.152)
3.00	0.98	0.003 (0.158)	0.010 (0.162)	-0.375 (0.136)	-0.398 (0.141)
3.50	0.99	0.006 (0.159)	0.011 (0.159)	-0.387 (0.137)	-0.396 (0.138)
4.00	1.00	0.007 (0.159)	0.009 (0.159)	-0.389 (0.137)	-0.394 (0.138)
		Entry Times Distributed $Unif(0,5)$			
1.37	0.12	-0.003 (0.078)	0.003 (0.139)	-0.022 (0.068)	-0.050 (0.127)
2.38	0.42	-0.002 (0.108)	-0.014 (0.135)	-0.084 (0.095)	-0.114 (0.121)
3.10	0.67	-0.010 (0.121)	-0.013 (0.132)	-0.143 (0.105)	-0.179 (0.118)
3.57	0.84	-0.012 (0.126)	-0.017 (0.132)	-0.180 (0.110)	-0.203 (0.118)
3.79	0.93	-0.013 (0.128)	-0.017 (0.134)	-0.196 (0.112)	-0.215 (0.118)
3.88	0.98	-0.013 (0.129)	-0.015 (0.134)	-0.203 (0.113)	-0.222 (0.118)
3.97	0.99	-0.011 (0.130)	-0.014 (0.134)	-0.208 (0.113)	-0.226 (0.118)
4.00	1.00	-0.012 (0.130)	-0.015 (0.134)	-0.212 (0.113)	-0.231 (0.118)

Inferential Methods

Analysis at the end of the trial must take into account the sampling plan

- Methods for confidence intervals involve defining an “ordering of the sample space”
  - Must decide how to order results obtained at different stopping times
- Previously described methods
  - Analysis time or stagewise ordering
  - MLE ordering
  - Z statistic ordering

Optimality Criteria

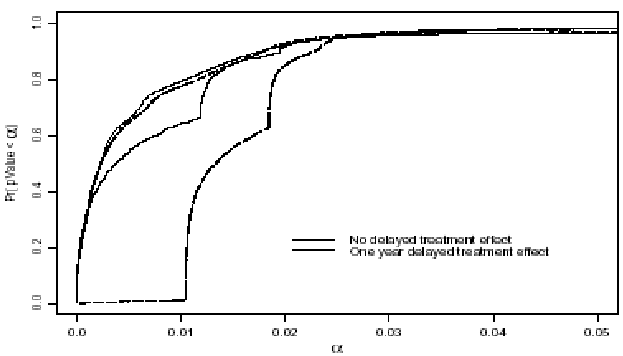
There is no single best ordering

- Whitehead and Jennison & Turnbull prefer the analysis time ordering
- In the presence of time invariant treatment effects, it does not usually make too much of a difference

Optimality Criteria

However, the analysis time ordering corresponds to the error spending function

- You can never get a P value less than the error spent
- This means that with late onset treatment effects, you can not achieve as low P values as might otherwise be indicated
  - Great impact on “pivotal trials”



(a) Pocock

Power to Obtain Low P values

Delay in Treatment Effect (yrs)	Ordering	α .000625	.001	.01	.025	.000625	.001	.01	.025
		*DSN*1 (Pocock)				*DSN*2 (O'Brien-Fleming)			
0	Z-statistic	0.284	0.330	0.766	0.918	0.384	0.431	0.872	0.954
	Analysis Time	0.266	0.311	0.621	0.914	0.266	0.311	0.850	0.952
1	Z-statistic	0.112	0.119	0.152	0.160	0.185	0.202	0.314	0.327
	Analysis Time	0.001	0.001	0.012	0.160	0.001	0.001	0.311	0.327
2	Z-statistic	0.010	0.010	0.017	0.029	0.034	0.034	0.042	0.047
	Analysis Time	0.000	0.000	0.007	0.029	0.000	0.000	0.009	0.047
		*DSN*3				*DSN*4			
0	Z-statistic	0.390	0.440	0.878	0.958	0.496	0.553	0.888	0.964
	Analysis Time	0.266	0.311	0.852	0.956	0.266	0.552	0.853	0.961
1	Z-statistic	0.249	0.287	0.695	0.792	0.272	0.316	0.822	0.947
	Analysis Time	0.000	0.000	0.221	0.492	0.001	0.240	0.606	0.945
2	Z-statistic	0.054	0.054	0.061	0.066	0.352	0.387	0.585	0.600
	Analysis Time	0.000	0.000	0.009	0.066	0.000	0.000	0.010	0.600

Final Comments

We have found that our first attempts at improving the scientific use of the weighted logrank statistics has worked well

- Greatly improved consistent estimation
- Minimal loss of power

Final Comments

Much more work is needed when using sequential methods with time varying treatment effects

- We are exploring the use of Bayesian random walk processes to model the types of alternatives that might be addressed
- However, this is truly an insoluble problem:
  - There is nothing in the data that can guarantee what future data might look like

Final Comments

In any case, however, the issue of paramount importance is that decisions about the summary measure be driven by the scientifically important effects

- Censored survival data requires a bit of extra care
- But the scientific issues are the same