

# **A Log Rank Test Statistic for Clustered or Paired Survival Data: Power and Sample Size Calculations**

Ronald E. Gangnon<sup>1</sup> and Michael R. Kosorok<sup>1,2</sup>

<sup>1</sup>Department of Biostatistics and Medical Informatics and <sup>2</sup>Department of Statistics  
University of Wisconsin    Madison, Wisconsin

# Problem Description

Common feature of ophthalmology clinical trials:

**Clustered or paired time-to-event data due to two eyes.**

Clustered data

Eyes within same subject receive same treatment.

Typical analysis: log rank test using time to first event.

Paired data

Eyes within same subject receive different treatments.

Typical analysis: log rank test ignoring the pairing.

Analyses using complete data would save time and resources.

Need to account for clustering in sample size calculations.

# Clustered Survival Data

Replicates within a cluster assigned to one of two treatments.  
For the  $k^{\text{th}}$  replicate of the  $j^{\text{th}}$  cluster ( $k = 1, 2, \dots, r_j; i = 1, 2, \dots, n$ ), observe

- $T_{jk}$  = failure (censoring) time
- $\delta_{jk}$  = censoring indicator
- $s_{jk}$  = treatment assignment

For the  $k^{\text{th}}$  replicate of the  $j^{\text{th}}$  cluster, define

- counting process  $N_{jk}(t) = I\{T_{jk} \geq t, \delta_{jk} = 1\}$ ,  $t \geq 0$ , and
- at-risk process  $Y_{jk}(t) = I\{T_{jk} \geq t\}$ ,  $t \geq 0$ .

For group  $i = 1, 2$ , define the counting and at-risk processes by

$$\bar{N}_i(t) = \sum_{j=1}^n \sum_{k=1}^{r_j} N_{jk}(t) I\{s_{jk} = i\} \quad \bar{Y}_i(t) = \sum_{j=1}^n \sum_{k=1}^{r_j} Y_{jk}(t) I\{s_{jk} = i\}$$

# Clustered Log Rank Test Statistic

The log rank statistic is defined to be

$$L = \int_0^{\infty} \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)} \left[ \frac{d\bar{N}_1(t)}{\bar{Y}_1(t)} - \frac{d\bar{N}_2(t)}{\bar{Y}_2(t)} \right].$$

Under some fairly weak conditions,

1. a consistent estimator of the variance of the log rank statistic is

$$Var(L) = \sum_{j=1}^n \left[ \sum_{k=1}^{r_j} \int_0^{\infty} \frac{\bar{Y}_{3-s_{jk}}(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)} \left\{ d\bar{N}_{jk}(t) - Y_{jk}(t)d\hat{\Lambda}(t) \right\} \right]^2.$$

2. the distribution of the log rank statistic is asymptotically normal.

# Calculation of Clustered Log Rank Test

The terms in the interior sums in the formula for  $\text{Var}(L)$ ,

$$\hat{S}_{jk} = \int_0^{\infty} \frac{\bar{Y}_{3-s_{jk}}(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)} \{d\bar{N}_{jk}(t) - Y_{jk}(t)d\hat{\Lambda}(t)\}$$

are the estimated score residuals from a Cox regression model with the regression coefficient set to zero.

Score residuals are available from Cox regression procedures in standard statistical packages such as S-Plus and SAS. An S-Plus function implementing the clustered log rank test is available upon request.

Note: For non-clustered data, the usual estimator of the variance for the ordinary log rank statistic is obtained by applying a martingale identity to the score residuals. In practice, for non-clustered data, the two estimators are very close.

# Intuition for Clustered Log Rank Test

Log rank test statistic equals the sum of the “true” score residuals.

$$S_{jk} = \int_0^{\infty} \frac{\bar{Y}_{3-s_{jk}}(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)} \{d\bar{N}_{jk}(t) - Y_{jk}(t)d\Lambda(t)\}$$

Under the null hypothesis of no treatment effect, the expected value of  $S_{jk}$  is 0, and score residuals from different subjects are assumed to be independent. Hence,

$$\text{Var}\left(\sum_{j,k} S_{jk}\right) = E\left(\sum_{j,k} S_{jk}\right)^2 = \sum_j E\left[\sum_k S_{jk}\right]^2.$$

A consistent estimator of the final term is obtained by replacing the score residuals with the estimated score residuals.

## Sample Size Formula (General Case)

$$K = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_1 p_2 \gamma^2} [1 - \rho + m \delta^2 \rho]$$

$K$  = number of events required

$\alpha$  = Type I error rate

$1 - \beta$  = power

$p_1$  = proportion of replicates assigned to group 1

$p_2 = 1 - p_1$

$\gamma$  = log hazard ratio (hypothesized treatment effect)

$m$  = number of replicates within each cluster

$\delta$  = difference in treatment assignment proportions within cluster

$\rho$  = correlation between  $M_{j1}$  and  $M_{j2}$

$$M_{jk} = \int_0^{\infty} d\bar{N}_{jk}(t) - Y_{jk}(t) d\Lambda(t)$$

## Sample Size Formula (Special Cases)

Clustered Survival Data (Treatments Assigned to Entire Cluster)

$$K = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_1 p_2 \gamma^2} [1 + (m-1)\rho]$$

Paired Survival Data

$$K = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_1 p_2 \gamma^2} [1 - \rho]$$

Reduces to Schoenfeld's formula if  $m = 1$  or  $\rho = 0$ .

$$K = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_1 p_2 \gamma^2}$$

# Comments on Sample Size Formulas

- Assumptions:
  - All clusters are the same size.
  - Observations within the same cluster are exchangeable.
  - Marginal proportional hazards model.
- Dependence within cluster measured by  $\rho$ .
  - $\rho$  depends on *both* the censoring and the failure time distribution.
  - Not a measure of true dependence of the failure times.
- $\rho$  is estimable using only masked data.
- In practice, adjust target number of events during trial, i.e., stop the trial when the observed number of events =  $K(\rho_{hat})$ .

## Simulations: Size and Power

Clustered data: 50, 100 or 200 subjects per group.

Use clustered log rank test (CLR) or time to 1<sup>st</sup> event (TFE).

Paired data: 50, 100 or 200 subjects.

Use clustered log rank test (CLR) or ignore pairing (OLR).

Two observations per subject.

Uniform recruitment over 1 year; 3 year study length.

Conditional PH model:

Gamma, log-normal, or inverse Gaussian frailty.

Baseline hazard (given frailty) equal to 1 (control) or HR (trt).

Marginal PH model:

Gamma frailty.

Marginal hazard equal to 1 (control) or HR (trt).

Treatment difference (HR) equal to 1, 0.75, and 0.5.

10,000 simulations. Tests performed at nominal 5% level.

# Simulation Results – Clustered Data, Null Hypothesis

		<u>Frailty distribution</u>							
		<u>Conditional models</u>				<u>Marginal model</u>			
<u>N</u>	<u>Frailty variance</u>	<u>Gamma</u>		<u>Lognormal</u>		<u>Inverse Gaussian</u>		<u>Gamma</u>	
		<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>
50	0.5	5.7	5.4	5.4	5.4	5.5	5.3	5.6	5.2
100	0.5	5.0	5.1	5.1	5.1	5.0	4.8	5.5	5.2
200	0.5	5.2	5.0	5.2	4.8	5.0	4.9	5.0	4.5
50	1	5.1	4.8	5.6	5.6	5.7	5.5	5.8	5.4
100	1	5.3	5.3	5.4	5.2	5.3	5.1	5.5	5.2
200	1	5.5	5.6	5.1	5.1	4.8	5.1	5.4	5.3
50	2	5.1	5.0	5.6	5.4	5.6	5.4	6.4	5.5
100	2	5.1	5.3	5.3	5.1	5.1	4.9	5.8	5.5
200	2	5.0	5.0	4.9	5.0	5.4	5.3	5.0	4.7

# Simulation Results – Paired Data, Null Hypothesis

		<u>Frailty distribution</u>						<u>Marginal model</u>	
		<u>Conditional models</u>							
<u>N</u>	<u>Frailty variance</u>	<u>Gamma</u>		<u>Lognormal</u>		<u>Inverse Gaussian</u>		<u>Gamma</u>	
		<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>
50	0.5	5.1	1.9	5.4	2.7	5.2	2.5	5.6	1.8
100	0.5	5.2	1.9	5.2	2.8	5.4	2.7	5.4	1.5
200	0.5	5.4	2.1	5.4	2.7	5.4	2.7	5.1	1.3
50	1	4.5	0.6	5.5	2.1	5.2	2.0	4.9	0.4
100	1	4.8	0.7	4.7	1.7	5.1	2.0	4.9	0.3
200	1	4.8	0.7	5.4	1.9	5.2	1.8	5.0	0.3
50	2	4.7	0.2	4.8	1.2	5.0	1.2	4.5	0.0
100	2	4.7	0.2	5.2	1.3	4.7	1.2	4.4	0.0
200	2	5.0	0.2	5.3	1.5	5.2	1.3	5.0	0.0

## Simulation Results – Clustered Data, HR = 0.5

		<u>Frailty distribution</u>							
		<u>Conditional models</u>				<u>Marginal model</u>			
<u>N</u>	<u>Frailty variance</u>	<u>Gamma</u>		<u>Lognormal</u>		<u>Inverse Gaussian</u>		<u>Gamma</u>	
		<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>
50	0.5	81.5	65.0	88.8	74.4	88.5	73.9	97.1	82.3
100	0.5	98.3	91.6	99.4	96.2	99.4	95.9	100.0	98.4
200	0.5	100.0	99.7	100.0	100.0	100.0	99.9	100.0	100.0
50	1	60.7	46.9	78.3	64.0	77.6	63.9	95.0	78.7
100	1	88.2	75.2	97.2	90.6	96.8	90.4	99.6	90.6
200	1	99.3	96.2	100.0	99.6	100.0	99.6	100.0	100.0
50	2	35.1	27.3	66.0	54.8	62.8	52.8	92.2	77.3
100	2	62.4	49.2	91.6	82.8	89.9	81.8	99.9	96.9
200	2	88.5	78.1	99.7	98.5	99.5	98.1	100.0	100.0

## Simulation Results – Paired Data, HR = 0.5

		<u>Frailty distribution</u>							
		<u>Conditional models</u>				<u>Marginal model</u>			
<u>N</u>	<u>Frailty variance</u>	<u>Gamma</u>		<u>Lognormal</u>		<u>Inverse Gaussian</u>		<u>Gamma</u>	
		<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>
50	0.5	75.2	65.0	79.3	72.7	78.3	71.4	96.5	92.0
100	0.5	97.0	93.7	95.9	90.8	97.3	95.6	100.0	99.8
200	0.5	100.0	99.9	100.0	100.0	100.0	99.9	100.0	100.0
50	1	67.4	45.0	73.7	61.7	72.2	60.4	99.2	95.1
100	1	93.1	81.2	95.7	91.7	95.5	90.6	100.0	100.0
200	1	99.9	99.2	100.0	99.5	99.9	99.8	100.0	100.0
50	2	53.5	21.2	67.1	50.2	63.7	45.5	100.0	97.4
100	2	84.2	52.3	92.6	84.2	91.3	81.7	100.0	100.0
200	2	99.0	90.7	99.8	99.4	99.6	98.9	100.0	100.0

## Simulation Results – Clustered Data, HR = 0.75

		<u>Frailty distribution</u>							
		<u>Conditional models</u>				<u>Marginal model</u>			
<u>N</u>	<u>Frailty variance</u>	<u>Gamma</u>		<u>Lognormal</u>		<u>Inverse Gaussian</u>		<u>Gamma</u>	
		<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>	<u>CLR</u>	<u>TFE</u>
50	0.5	22.9	16.9	28.2	19.5	26.7	18.9	39.1	23.1
100	0.5	39.9	28.1	48.3	34.2	48.9	35.2	65.3	41.9
200	0.5	67.0	49.8	77.2	59.8	77.4	59.1	90.8	68.6
50	1	15.6	12.4	22.3	16.9	21.7	16.7	35.1	21.5
100	1	26.1	19.6	38.2	28.8	36.0	27.3	59.2	38.2
200	1	45.7	34.4	64.4	49.2	63.2	49.6	86.5	64.9
50	2	9.9	8.8	17.7	14.4	17.2	14.3	32.8	22.2
100	2	15.1	12.6	29.1	23.4	27.3	22.6	53.7	37.0
200	2	26.7	20.7	51.3	40.7	48.6	40.3	81.6	63.4

## Simulation Results – Paired Data, HR = 0.75

		<u>Frailty distribution</u>							
		<u>Conditional models</u>				<u>Marginal model</u>			
<u>N</u>	<u>Frailty variance</u>	<u>Gamma</u>		<u>Lognormal</u>		<u>Inverse Gaussian</u>		<u>Gamma</u>	
		<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>	<u>CLR</u>	<u>OLR</u>
50	0.5	21.0	12.2	22.2	15.7	21.4	14.8	37.3	22.5
100	0.5	37.7	24.1	38.7	29.2	38.6	29.4	64.4	46.9
200	0.5	64.0	48.9	66.1	56.4	66.5	56.8	91.5	81.8
50	1	17.1	5.9	20.5	11.6	19.3	10.7	50.2	18.6
100	1	31.3	12.7	34.9	22.3	34.0	21.7	80.8	46.6
200	1	55.8	29.6	61.3	46.7	59.3	44.3	98.1	86.2
50	2	13.6	1.7	17.5	7.7	16.0	7.3	75.5	13.5
100	2	23.9	4.2	31.1	16.0	29.6	15.3	97.0	46.7
200	2	44.1	10.6	55.2	35.1	52.8	33.3	100.0	93.2

# Comments on Simulation Results

## Clustered data

- Either CLR or TFE will maintain correct size.
- Substantial improvements in power using CLR over TFE.

## Paired data

- CLR maintains correct size.
- OLR dramatically overstates size.
- Substantial improvements in power using CLR over OLR.

## Overall comments

- Empirical power agrees with theoretical calculations.
- Inverse Gaussian and lognormal frailty models very similar.
- Martingale correlations range between 0.23 and 0.79.

## **Example:**

# **Early Treatment Diabetic Retinopathy Study**

- 3711 patients with nonproliferative or early proliferative diabetic retinopathy
- Enrollment lasted from April 1980 to July 1985
- Final follow-up visit in June 1989
  
- Multifactorial treatment design with several different endpoints.
- For illustrative purposes, we consider only a single question.
  
- One eye per patient randomized to early photocoagulation
- Fellow eye assigned to deferral until development of high risk proliferative diabetic retinopathy
  
- Survival endpoint:
  - Time to severe visual loss or vitrectomy

# ETDRS Design Considerations

- Treatment effect:
  - 5-year rate in eyes assigned to deferral = 10%
  - 5-year rate in eyes assigned to early photocoagulation = 6%
  - Hazard ratio = 0.587 ( $\gamma = -0.533$ ).
- Two-sided Type I error rate ( $\alpha$ ) = 1%
- Power ( $1-\beta$ ) = 98%
- Paired assignment of treatments ( $p_1 = p_2 = 0.5$ ).
- Required number of events (ignoring pairing) = 303.  
Assumes fellow eyes are uncorrelated.
- Correlation-adjusted required number of events = 303 (1- $\rho$ ).

# Timing of Analysis in Event-Driven ETDRS Design

Analysis Date	N	$\rho_{\text{hat}}$	$K(\rho_{\text{hat}})$	Stop N>K( $\rho_{\text{hat}}$ )?	Stop N>K(0)?
April 9, 1985	125	0.401	182	No	No
Oct. 9, 1985	165	0.359	195	No	No
April 9, 1986	202	0.337	201	Yes	No
Oct. 9, 1986	240	0.318	207		No
April 9, 1987	276	0.314	208		No
Oct. 9, 1987	318	0.316	208		Yes
April 9, 1988	352	0.331	203		
Oct. 9, 1988	378	0.330	203		
April 9, 1989	388	0.325	205		

N = observed number of events as of analysis date

$\rho_{\text{hat}}$  = estimated martingale correlation as of analysis date

$K(\rho_{\text{hat}})$  = correlation-adjusted required number of events

## Analyses of ETDRS Data

As of April 9, 1986 (final analysis using correlation adjustment),

Cluster log rank:  $Z = -2.64$ , p-value = 0.0084.

Ordinary log rank:  $Z = -2.15$ , p-value = 0.0318.

As of October 9, 1987 (final analysis using number of events),

Cluster log rank:  $Z = -4.30$ , p-value = 0.000017.

Ordinary log rank:  $Z = -3.56$ , p-value = 0.00037.

Final data as of June 1989,

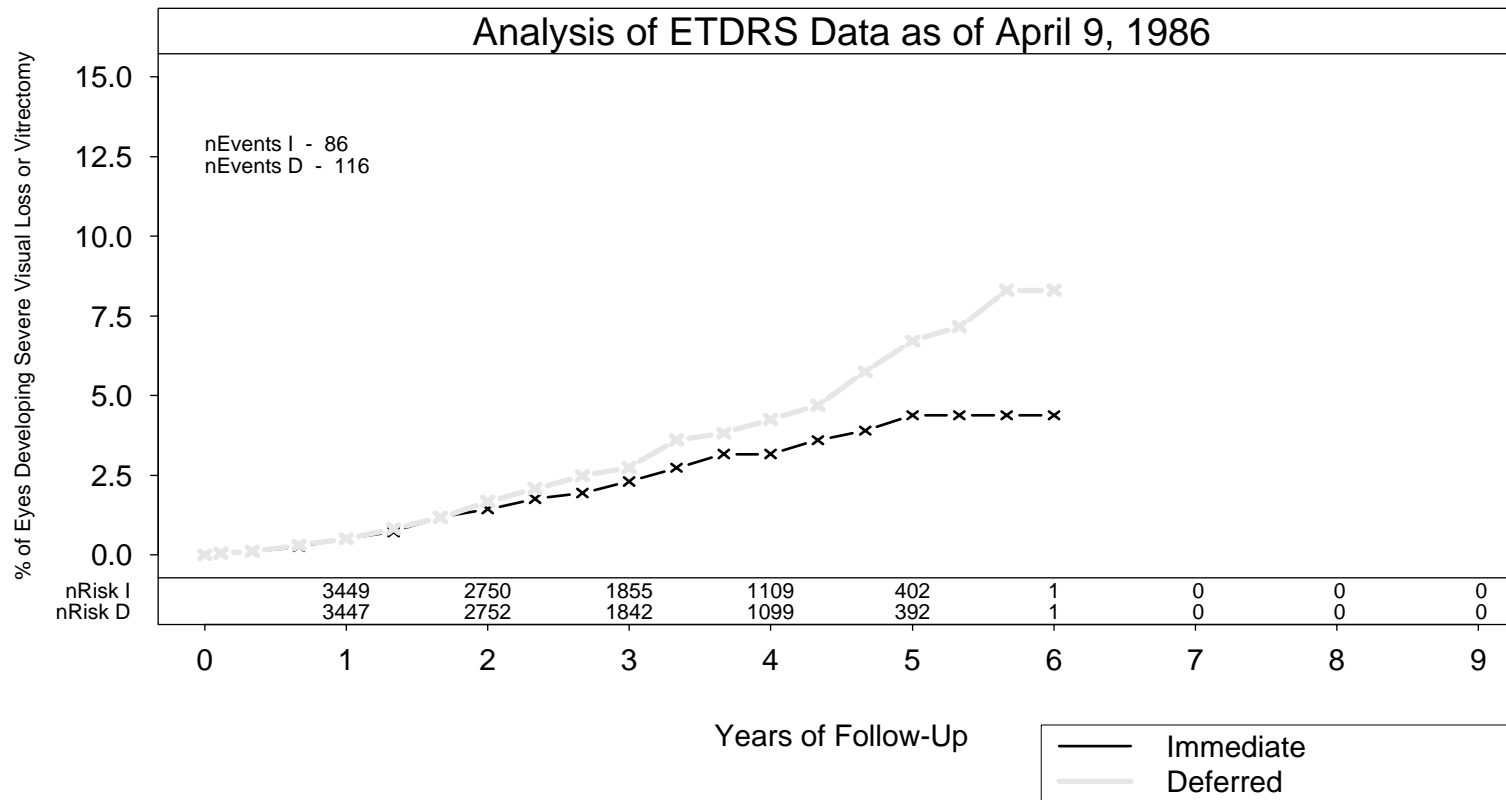
Cluster log rank:  $Z = -4.89$ , p-value = 0.0000016.

Ordinary log rank:  $Z = -3.98$ , p-value = 0.000067.

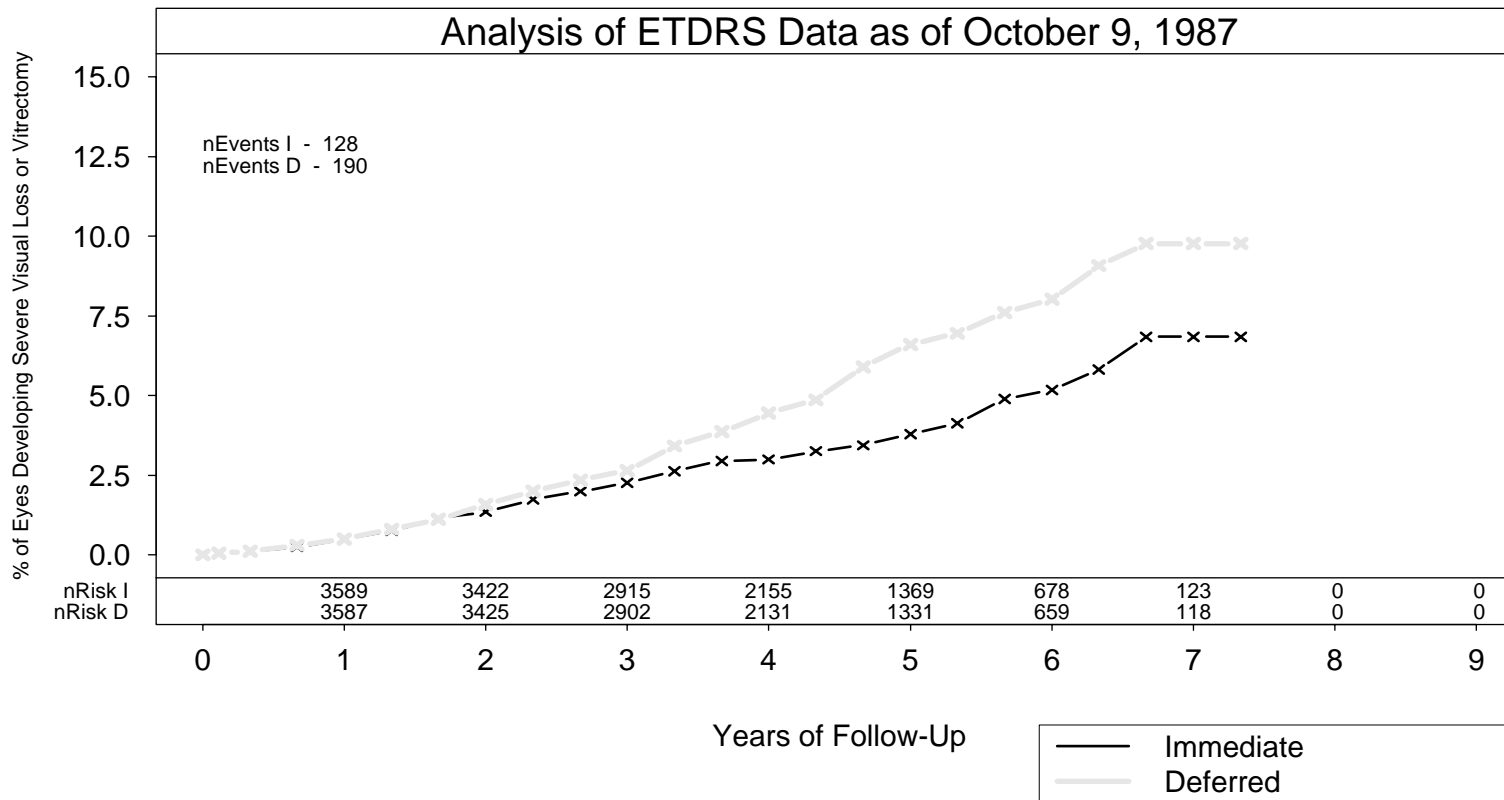
Accounting for correlation of fellow eyes:

- *reduces study length by 1.5 years and*
- *requires 1/3 fewer events (202 v. 303).*

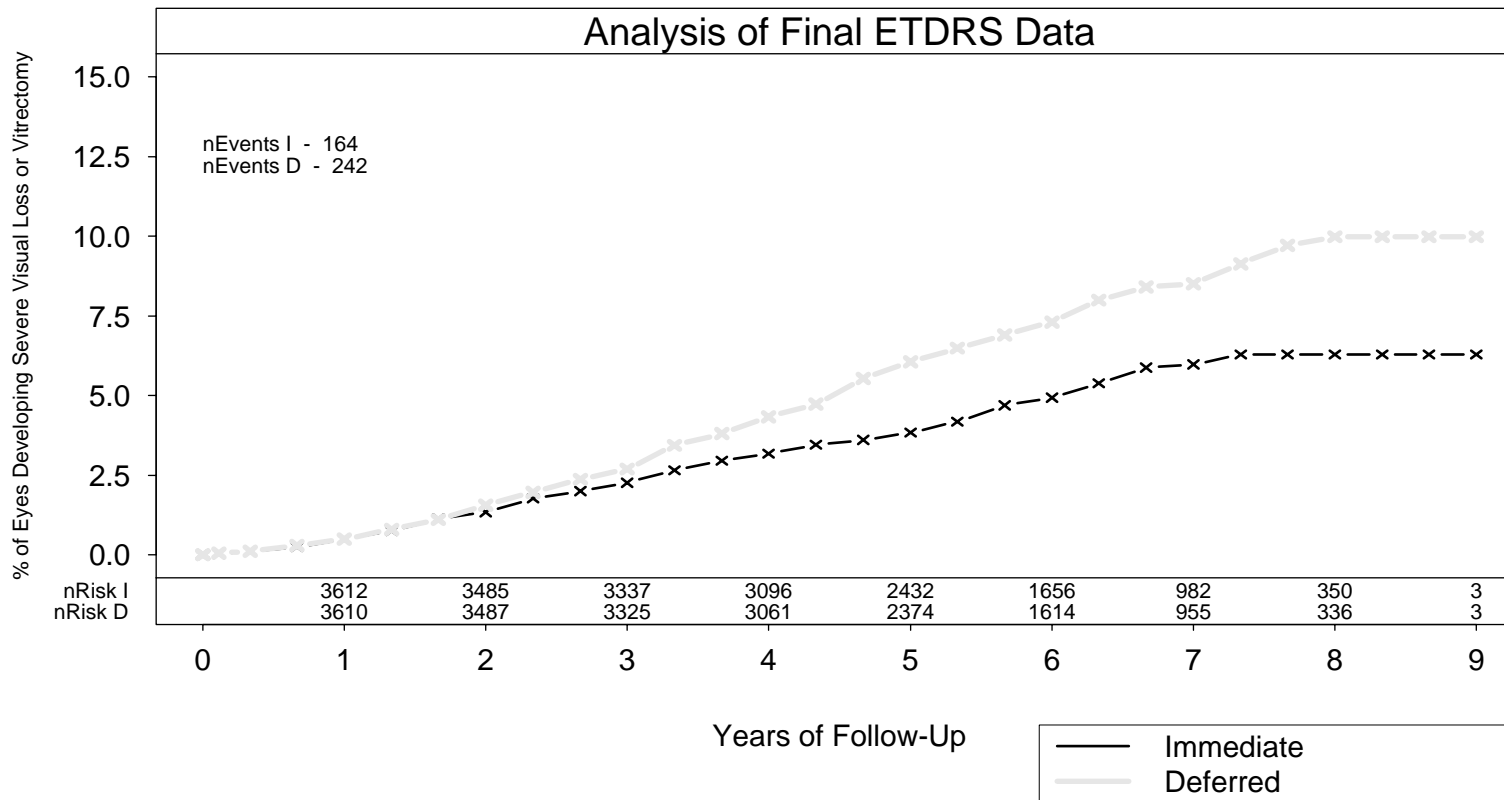
# Severe Visual Loss or Vitrectomy over Time



# Severe Visual Loss or Vitrectomy over Time



# Severe Visual Loss or Vitrectomy over Time



## Extensions & Future Work

- Stratification  
Calculate score residuals separately for each stratum.  
Observations within clusters can belong to different strata.
- Weights  
Allows for an arbitrary predictable weight function such as Fleming-Harrington  $G^{\rho,\gamma}$  family.  
Replace of score residuals with weighted score residuals.
- Sequential monitoring  
Independent increments structure (?).

### Contact Information

e-mail: [ronald@biostat.wisc.edu](mailto:ronald@biostat.wisc.edu)

Web: <http://www.biostat.wisc.edu/~ronald>