

Leonhard Held, Stefanie von Felten

University of Zurich

SHARE-CTD 1st Schooling Event

27.-31. January 2025

Reisensburg, Germany

Complete lecture notes at

https://bookdown.org/charlotte_micheloud93/Clinical_Biostatistics

Lecture 1: Analysis of Continuous Outcomes

Analysis of Quantitative Outcomes

Comparison of Two Groups

t -Test

Adjusting for Baseline

Comparing Change from Baseline

Analysis of Covariance (ANCOVA)

Example: Digeridoo Study

- Puhan et al. (BMJ, 2005)
- Randomized controlled trial, simple randomization.
- Patients with moderate **obstructive sleep apnoea syndrome**
- **Treatment group**: 4 months Didgeridoo practice ($m = 14$)
- **Control group**: 4 months waiting list ($n = 11$)
- Primary endpoint: **Epworth** scale (0-24)
- Measurements are taken at the start of the study (**Baseline**) and after four months (**Follow-up**)

Abstract of Publication

Didgeridoo playing as alternative treatment for obstructive sleep apnoea syndrome: randomised controlled trial

Milo A Puhan, research fellow¹, **Alex Suarez**, didgeridoo instructor², **Christian Lo Cascio**, resident in internal medicine³, **Alfred Zahn**, sleep laboratory technician³, **Markus Heitz**, specialist in respiratory and sleep medicine⁴, **Otto Braendli**, specialist in respiratory and sleep medicine³

¹ Horten Centre, University of Zurich, 8091 Zurich, Switzerland, ² Asate Alex Suarez, 9630 Wattwil, Switzerland, ³ Zuercher Hoehenklinik Wald, CH-8639 Faltigberg-Wald, Switzerland, ⁴ Lungenpraxis Morgental, Zurich, Switzerland

Correspondence to: O Braendli otto.braendli@zhw.ch

Abstract

Objective To assess the effects of didgeridoo playing on daytime sleepiness and other outcomes related to sleep by reducing collapsibility of the upper airways in patients with moderate obstructive sleep apnoea syndrome and snoring.

Design Randomised controlled trial.

Setting Private practice of a didgeridoo instructor and a single centre for sleep medicine.

Participants 25 patients aged > 18 years with an apnoea-hypopnoea index between 15 and 30 and who complained about snoring.

Interventions Didgeridoo lessons and daily practice at home with standardised instruments for four months. Participants in the control group remained on the waiting list for lessons.

Main outcome measure Daytime sleepiness (Epworth scale from 0 (no daytime sleepiness) to 24), sleep quality (Pittsburgh quality of sleep index from 0 (excellent sleep quality) to 21), partner rating of sleep disturbance (visual analogue scale from 0 (not disturbed) to 10), apnoea-hypopnoea index, and health related quality of life (SF-36).

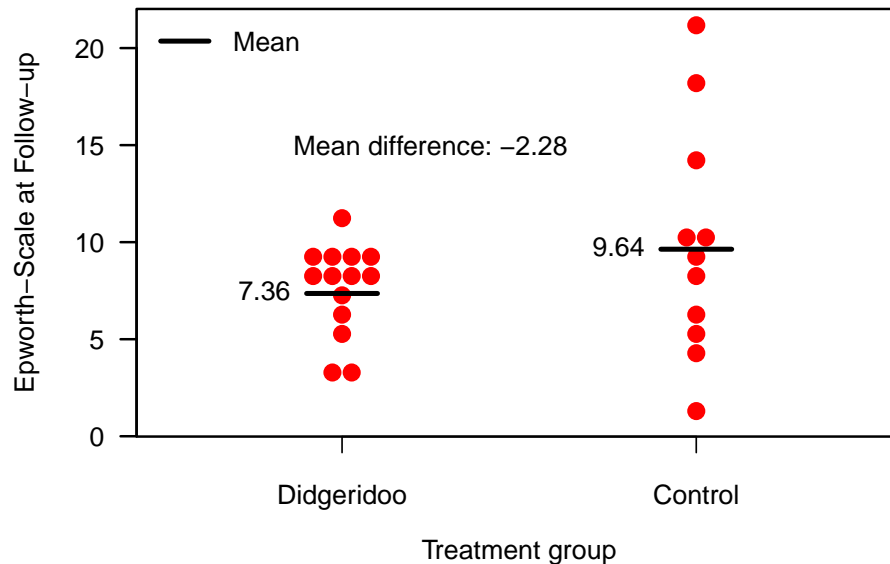
Results Participants in the didgeridoo group practised an average of 5.9 days a week (SD 0.86) for 25.3 minutes (SD 3.4). Compared with the control group in the didgeridoo group daytime sleepiness (difference -3.0, 95% confidence interval -5.7 to -0.3, $P = 0.03$) and apnoea-hypopnoea index (difference -6.2, -12.3 to -0.1, $P = 0.05$) improved significantly and partners reported less sleep disturbance (difference -2.8, -4.7 to -0.9, $P < 0.01$). There was no effect on the quality of sleep (difference -0.7, -2.1 to 0.6, $P = 0.27$). The combined analysis of sleep related outcomes showed a moderate to large effect of didgeridoo playing (difference between summary z scores -0.78 SD units, -1.27 to -0.28, $P < 0.01$). Changes in health related quality of life did not differ between groups.

Conclusion Regular didgeridoo playing is an effective treatment alternative well accepted by patients with moderate obstructive sleep apnoea syndrome.

Follow-up Measurements of Primary Endpoint

```
table(treatment)
```

```
## treatment
##      Control Didgeridoo
##          11         14
```



t-Test

Comparison of follow-up measurements

```
print(tTest1 <- t.test(f.up ~ treatment, var.equal=TRUE))

##
## Two Sample t-test
##
## data: f.up by treatment
## t = 1.3026, df = 23, p-value = 0.2056
## alternative hypothesis: true difference in means between group Control and
## 95 percent confidence interval:
## -1.340366 5.898807
## sample estimates:
## mean in group Control mean in group Didgeridoo
##          9.636364          7.357143

(DifferenceInMeans <- mean(tTest1$conf.int))

## [1] 2.279221
```

No evidence for a difference in follow-up means ($p = 0.21$)

Regression Analysis

Gives the same results

```
model1 <- lm(f.up ~ treatment)
tableRegression(model1, intercept=FALSE)
```

	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.28	from -5.90 to 1.34	0.21

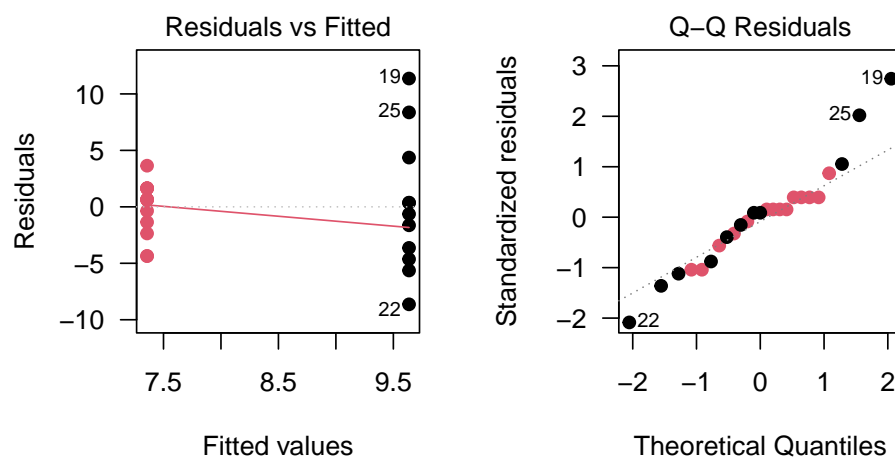
Advantages:

- Can be easily generalized
- Residuals can be checked

Regression Diagnostics

Indicate poor model fit (signs of variance heterogeneity)

```
par(mfrow=c(1,2), pty="s", las=1)
plot(model1, which=1, pch=19, col=treatment)
plot(model1, which=2, pch=19, col=treatment)
```



Assumptions of t -Test

Population model:

- Distribution assumption: **Normality**

Treatment: with mean μ_T and variance σ_T^2

Control: with mean μ_C and variance σ_C^2

- Random samples of size m and n
- Quantity of interest: **Mean difference** $\Delta = \mu_T - \mu_C$
- **Equal variances** assumption: $\sigma_T^2 = \sigma_C^2 = \sigma^2$

Test of Equality of Variance

Bartlett's test

```
print(bTest <- bartlett.test(f.up ~ treatment))  
  
##  
## Bartlett test of homogeneity of variances  
##  
## data: f.up by treatment  
## Bartlett's K-squared = 9.1295, df = 1, p-value = 0.002515
```

Strong evidence for variance heterogeneity ($p = 0.003$)

Confirms earlier findings based on regression diagnostics.

Welch's Test

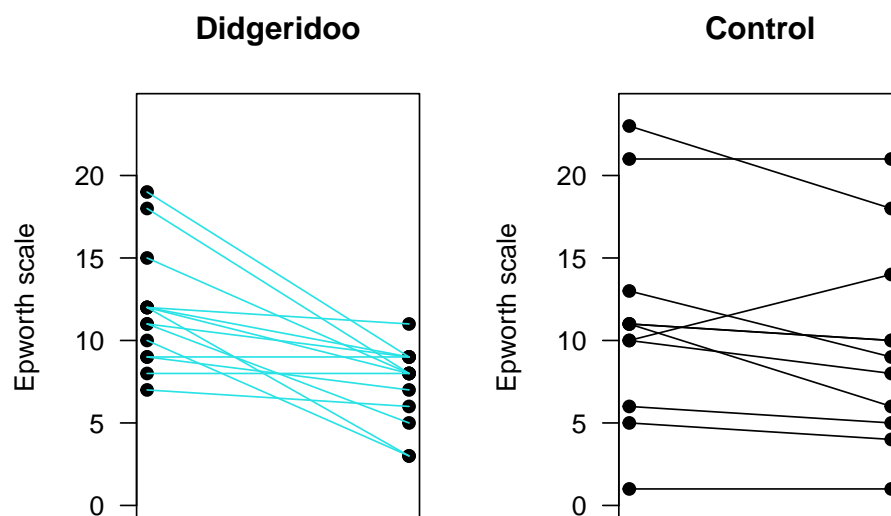
Does not assume equality of variances

```
(WelchTest1 <- t.test(f.up ~ treatment))

##
##  Welch Two Sample t-test
##
## data:  f.up by treatment
## t = 1.187, df = 12.381, p-value = 0.2575
## alternative hypothesis: true difference in means between group Control and
## 95 percent confidence interval:
##  -1.890289  6.448731
## sample estimates:
##      mean in group Control mean in group Didgeridoo
##           9.636364           7.357143
```

No evidence for a difference in follow-up means ($p = 0.26$)

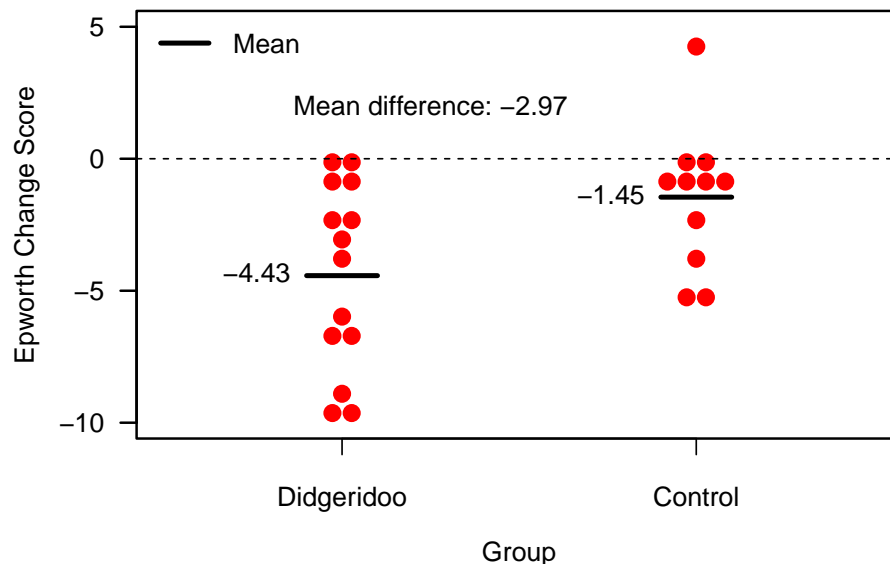
Analysis of Baseline and Follow-up Measurements



Change Scores

- Baseline values may be imbalanced between treatment groups just as any other prognostic factor.
- Analyse **change from baseline**:

$$\text{Change score} = \text{follow-up} - \text{baseline}$$



Comparing Change from Baseline

```
change.score <- f.up - baseline
print(tTest2 <- t.test(change.score ~ treatment, var.equal=TRUE))

##
## Two Sample t-test
##
## data: change.score by treatment
## t = 2.2748, df = 23, p-value = 0.03256
## alternative hypothesis: true difference in means between group Control and
## 95 percent confidence interval:
## 0.2695582 5.6784938
## sample estimates:
## mean in group Control mean in group Didgeridoo
## -1.454545 -4.428571

print(DifferenceInMeans <- mean(tTest2$conf.int))

## [1] 2.974026
```

Evidence for a difference in change from baseline ($p = 0.033$)

Reported Results

Table 2 Effects of intervention on sleep related outcomes

Outcome	Didgeridoo group	Control group	Raw difference* (95% CI)	Adjusted difference† (95% CI)
Epworth scale				
At 4 months	7.4 (2.3)	9.6 (6.0)		
Change from baseline	-4.4 (3.7)	-1.4 (2.6)	-3.0 (-5.7 to -0.3), P=0.03	-2.8 (-5.4 to -0.2), P=0.04
Pittsburgh quality of sleep index				
At 4 months	4.3 (2.1)	5.6 (2.7)		
Change from baseline	-0.9 (1.6)	-0.2 (1.7)	-0.7 (-2.1 to 0.6), P=0.27	-0.8 (-2.3 to 0.8), P=0.30
Partner rating of sleep disturbance				
At 4 months	2.3 (1.4)	4.8 (2.2)		
Change from baseline	-3.4 (2.4)	-0.6 (1.9)	-2.8 (-4.7 to -0.9), P<0.01	-2.7 (-4.2 to -1.2), P<0.01
Apnoea-hypopnoea index				
At 4 months	11.6 (8.1)	15.4 (9.8)		
Change from baseline	-10.7 (7.7)	-4.5 (6.9)	-6.2 (-12.3 to -0.1), P=0.05	-6.6 (-13.3 to -0.1), P=0.05

* Two sample t tests.

† Analysis of covariance with adjustment for severity of disease (apnoea-hypopnoea index and Epworth scale) and weight change during study period.

Change Score Analysis with a Regression Model

The change score analysis can also be done with regression:

```
model2 <- lm(f.up ~ treatment + offset(baseline))
tableRegression(model2, intercept=FALSE)
```

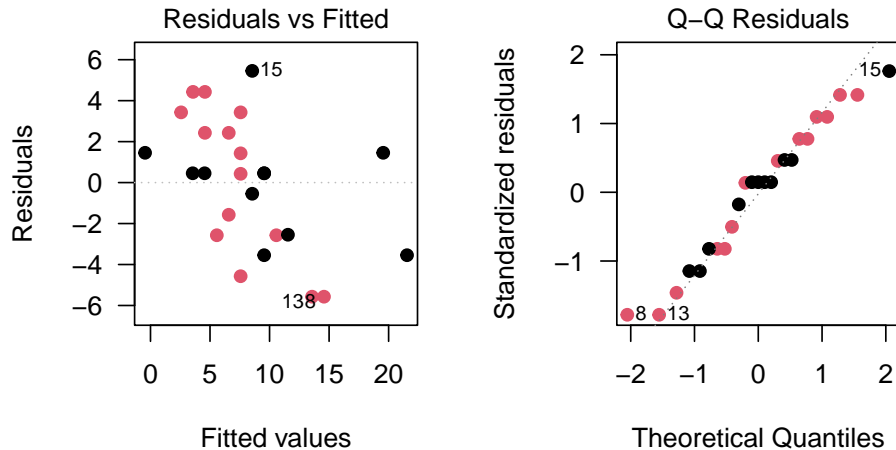
	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.97	from -5.68 to -0.27	0.033

The `offset(x)` command fixes the coefficient of `x` at 1.

Regression Diagnostics

Somewhat better model fit

```
par(mfrow=c(1,2), pty="s", las=1)
plot(model2, which=1, pch=19, col=treatment, add.smooth=FALSE)
plot(model2, which=2, pch=19, col=treatment)
```



Comparison of Effect Estimates

Some theory

- Outcome means

at Baseline in both groups: μ_B
at Follow-up in the control group: μ
at Follow-up in the treatment group: $\mu + \Delta$

- Mean difference Δ is of primary interest.
- Assume common variance σ^2 of all measurements
- Correlation ρ between baseline and follow-up measurements
- Assume there are n observations in each group.

Comparison of Effect Estimates

1. Difference of mean follow-up measurements $\hat{\Delta}_1$
 2. Difference of mean change scores $\hat{\Delta}_2$
- Both estimates are **unbiased** (assuming baseline balance).
 - For $\rho > 1/2$, $\hat{\Delta}_2$ will have **smaller variance** than $\hat{\Delta}_1$, so will produce narrower confidence intervals and more powerful tests:

$$\begin{aligned}\text{Var}(\hat{\Delta}_1) &= 2\sigma^2/n \\ \text{Var}(\hat{\Delta}_2) &= 4\sigma^2(1 - \rho)/n\end{aligned}$$

- Didgeridoo study: $\hat{\rho} = 0.72$

Analysis of Covariance (ANCOVA)

It is natural to extend

```
model2 <- lm(f.up ~ treatment + offset(baseline))
```

to the **ANCOVA** model:

```
model3 <- lm(f.up ~ treatment + baseline)
tableRegression(model3, intercept=FALSE)
```

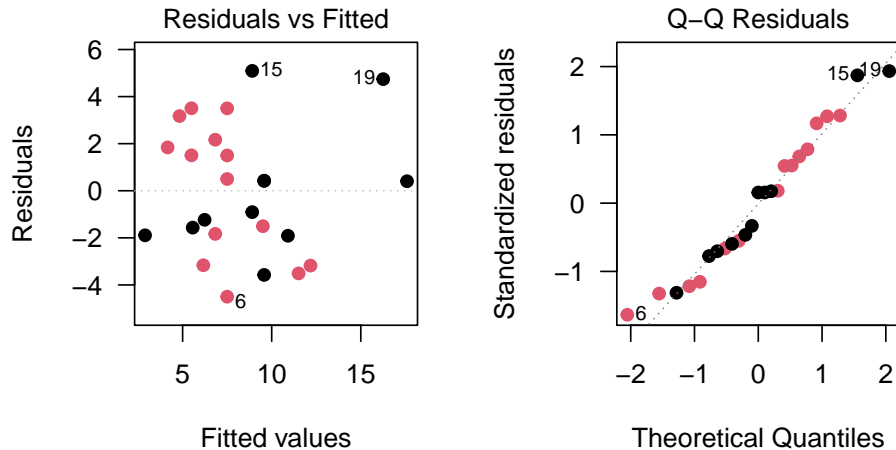
	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.74	from -5.14 to -0.35	0.027
baseline	0.67	from 0.42 to 0.92	< 0.0001

Now the coefficient of `baseline` is estimated from the data.

Regression Diagnostics

Good model fit

```
par(mfrow=c(1,2), pty="s", las=1)
plot(model3, which=1, pch=19, col=treatment, add.smooth=FALSE)
plot(model3, which=2, pch=19, col=treatment)
```



The Baseline Coefficient

- Denote the coefficient of the baseline variable as β .
- The ANCOVA model reduces
 - for $\beta = 0$ to the analysis of follow-up and
 - for $\beta = 1$ to the analysis of change scores.
- ANCOVA estimates β and the mean difference Δ jointly with **multiple regression**.
- The estimate $\hat{\beta}$ is usually close to the correlation ρ .

Didgeridoo Study: A Comparison

```
tableRegression(model1, intercept=FALSE)
```

	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.28	from -5.90 to 1.34	0.21

```
tableRegression(model2, intercept=FALSE)
```

	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.97	from -5.68 to -0.27	0.033

```
tableRegression(model3, intercept=FALSE)
```

	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.74	from -5.14 to -0.35	0.027
baseline	0.67	from 0.42 to 0.92	< 0.0001

Conditioning on Baseline Values

- Let \bar{b}_T and \bar{b}_C denote the **observed** mean baseline values in the current trial.
- Given \bar{b}_T and \bar{b}_C , $\hat{\Delta}_1$ and $\hat{\Delta}_2$ are both **biased** if
 1. there is **correlation** $\rho > 0$ between baseline and follow-up measurements
 2. and there is **baseline imbalance** ($\bar{b}_T \neq \bar{b}_C$):

$$E(\hat{\Delta}_1 | \bar{b}_T, \bar{b}_C) = \Delta + \underbrace{\rho \cdot (\bar{b}_T - \bar{b}_C)}_{\text{bias}}$$

$$E(\hat{\Delta}_2 | \bar{b}_T, \bar{b}_C) = \Delta + \underbrace{(\rho - 1) \cdot (\bar{b}_T - \bar{b}_C)}_{\text{bias}}$$

- Didgeridoo study has some imbalance: $\bar{b}_T = 11.1$, $\bar{b}_C = 11.8$

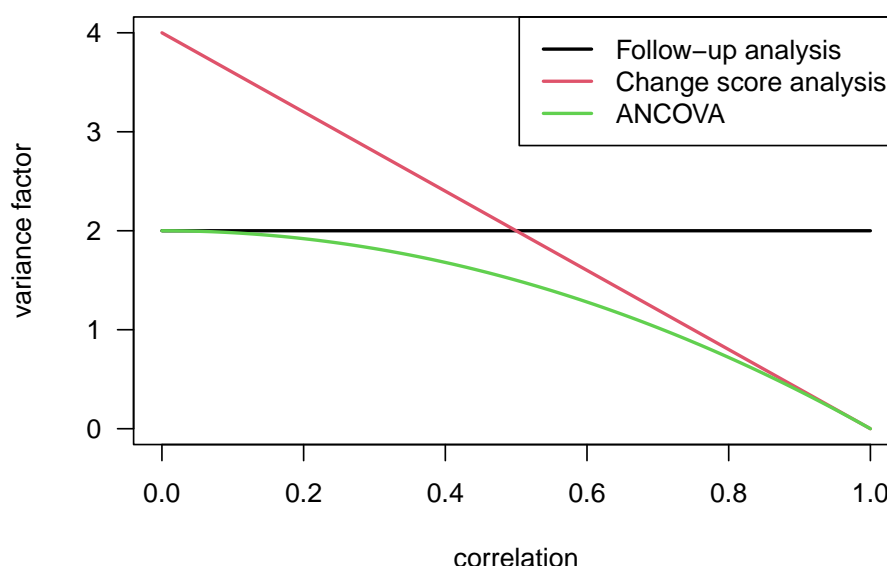
Comparison of Variance Factors

$$\text{Var}(\hat{\Delta}) = \text{variance factor} \cdot \sigma^2/n$$

$$\text{Var}(\hat{\Delta}_1) = 2 \cdot \sigma^2/n$$

$$\text{Var}(\hat{\Delta}_2) = 4(1 - \rho) \cdot \sigma^2/n$$

$$\text{Var}(\hat{\Delta}_3) = 2(1 - \rho^2) \cdot \sigma^2/n$$



Adjusting for Other Variables

```
model4 <- lm(f.up ~ treatment + baseline + weight.change + base.apnoea)
tableRegression(model4, intercept=FALSE)
```

	Coefficient	95%-confidence interval	p-value
treatmentDidgeridoo	-2.75	from -5.35 to -0.15	0.039
baseline	0.67	from 0.41 to 0.93	< 0.0001
weight.change	-0.17	from -0.92 to 0.57	0.63
base.apnoea	0.023	from -0.25 to 0.29	0.86

- ANCOVA allows a wide range of variables measured at baseline to be used to adjust the mean difference.
- The safest approach to selecting these variables is to decide this **before** the trial starts (in the study protocol).
- Prognostic variables used to stratify the allocation should **always** be included as covariates.

Reported Results

Table 2 Effects of intervention on sleep related outcomes

Outcome	Didgeridoo group	Control group	Raw difference* (95% CI)	Adjusted difference [†] (95% CI)
Epworth scale				
At 4 months	7.4 (2.3)	9.6 (6.0)		
Change from baseline	-4.4 (3.7)	-1.4 (2.6)	-3.0 (-5.7 to -0.3), P=0.03	-2.8 (-5.4 to -0.2), P=0.04
Pittsburgh quality of sleep index				
At 4 months	4.3 (2.1)	5.6 (2.7)		
Change from baseline	-0.9 (1.6)	-0.2 (1.7)	-0.7 (-2.1 to 0.6), P=0.27	-0.8 (-2.3 to 0.8), P=0.30
Partner rating of sleep disturbance				
At 4 months	2.3 (1.4)	4.8 (2.2)		
Change from baseline	-3.4 (2.4)	-0.6 (1.9)	-2.8 (-4.7 to -0.9), P<0.01	-2.7 (-4.2 to -1.2), P<0.01
Apnoea-hypopnoea index				
At 4 months	11.6 (8.1)	15.4 (9.8)		
Change from baseline	-10.7 (7.7)	-4.5 (6.9)	-6.2 (-12.3 to -0.1), P=0.05	-6.6 (-13.3 to -0.1), P=0.05

* Two sample t tests.

[†] Analysis of covariance with adjustment for severity of disease (apnoea-hypopnoea index and Epworth scale) and weight change during study

Least-Squares Means

- Problem: ANCOVA estimate is not equal to difference of **raw** mean change scores
- Solution: Compute adjusted **least-squares (LS) means** via fitted values in both groups:

```
library(lsmmeans)
## raw means
print(rawMeans <- ref.grid(model2))

## treatment baseline prediction SE df
## Control 11.5 10.03 0.978 23
## Didgeridoo 11.5 7.05 0.867 23

## adjusted LS means
print(adjMeans <- ref.grid(model3, "baseline"))

## treatment baseline prediction SE df
## Control 11.5 9.90 0.863 22
## Didgeridoo 11.5 7.15 0.764 22
```

- 11.5 is the mean baseline value in the dataset
- ANCOVA estimate is $7.15 - 9.90 = -2.74$

Example

Efficacy and safety of balovaptan for socialisation and communication difficulties in autistic adults in North America and Europe: a phase 3, randomised, placebo-controlled trial



Suma Jacob, Jeremy Veenstra-VanderWeele, Declan Murphy, James McCracken, Janice Smith, Kevin Sanders, Christoph Meyenberg, Thomas Wiese, Gurpreet Deol-Bhullar, Christoph Wandel, Elizabeth Ashford, Evdokia Anagnostou

Balovaptan group			Placebo group		Estimated treatment difference (95% CI)
n	Least-squares mean change from baseline (SE)		n	Least-squares mean change from baseline (SE)	
Vineland-II two-domain composite score					
Week 12	84	1.73 (1.52)	79	4.05 (1.59)	-2.32 (-5.64 to 1.01)
Week 24	79	2.91 (1.52)	71	4.75 (1.60)	-1.84 (-5.15 to 1.48)
Vineland-II adaptive behavior composite score					
Week 12	84	1.45 (1.17)	79	3.19 (1.22)	-1.74 (-4.29 to 0.81)
Week 24	79	2.39 (1.18)	71	2.90 (1.24)	-0.52 (-3.11 to 2.07)
Vineland-II communication domain score					
Week 12	84	1.21 (2.08)	79	3.06 (2.17)	-1.85 (-6.38 to 2.68)
Week 24	79	1.89 (2.34)	71	4.09 (2.46)	-2.20 (-7.31 to 2.92)
Vineland-II socialisation domain score					
Week 12	84	2.27 (1.60)	79	5.05 (1.67)	-2.77 (-6.26 to 0.72)
Week 24	79	4.00 (1.58)	71	5.43 (1.66)	-1.43 (-4.88 to 2.01)
Vineland-II daily living skills domain score					
Week 12	84	1.79 (1.20)	79	1.60 (1.25)	0.19 (-2.43 to 2.81)
Week 24	79	2.56 (1.32)	71	-0.18 (1.40)	2.74 (-0.16 to 5.63)

Table 2: Vineland-II scores for the futility analysis population at weeks 12 and 24

Table 2: Vineland-II scores for the futility analysis population at weeks 12 and 24

Microlearnings

Join Course **Share-CTD Schooling Event January 2025** on klickerUZH



<https://pwa.klicker.uzh.ch/course/1f51197f-d4b4-407c-b05b-a94d327e4db0/join?pin=856832075>

Leonhard Held, Stefanie von Felten

University of Zurich

SHARE-CTD 1st Schooling Event

27.-31. January 2025

Reisensburg, Germany

Complete lecture notes at

https://bookdown.org/charlotte_micheloud93/Clinical_Biostatistics

Lecture 2: Subgroup Analysis

Subgroup Analysis

Comparing Subgroups

Selecting Subgroups

Qualitative and Quantitative Interaction

Subgroup Analysis

- Possible subgroups:
 - Male/Female
 - Children/Adults
 - etc.
- Question: Is the treatment effect different for different types of patients?
- Problems:
 - Analyses of subgroups will then have **less power** to detect a significant difference of the treatment effect between subgroups.
 - If subgroups are defined by many variables, the risk of a **false positive** (significant) finding increases.
 - Choice of subgroups difficult: *a priori* versus *post hoc*.

Example: The Neonatal Hypocalcemia Trial

- From Matthews (2006), Section 9.2
- A placebo-controlled trial of vitamin D supplementation of expectant mothers for the prevention of neonatal hypocalcemia.
- Primary endpoint: Serum calcium level of the babies at 1 week of age
- Subgroups: bottle-fed versus breast-fed babies

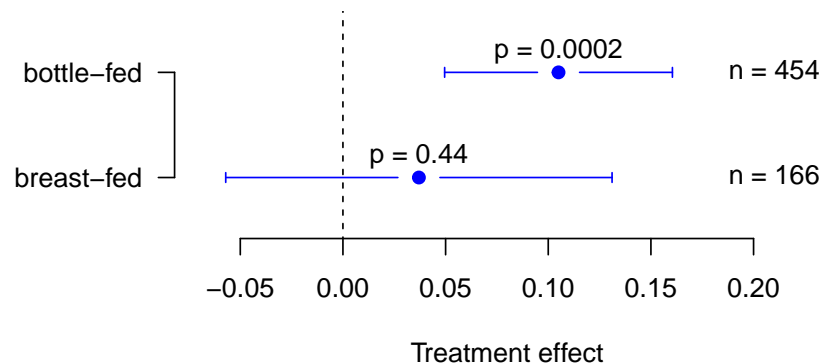
Compare Effect Sizes not P-Values

Statistics Notes

Interaction 2: compare effect sizes not P values

John N S Matthews, Douglas G Altman

- It is tempting to assess the existence of a **treatment interaction** based on the p -values in each subgroup.



- But p -values can differ, even if effect sizes are the same!

University of Zurich, Department of Biostatistics

Page 5

Example: The Neonatal Hypocalcemia Trial Data

Summary-level data:

	Breast-fed		Bottle-fed	
	Supplement	Placebo	Supplement	Placebo
n	64	102	169	285
Mean	2.445	2.408	2.300	2.195
Var	0.0853	0.0987	0.0752	0.1018
se	0.0365	0.0311	0.0211	0.0189

```
## Breast-fed
theta.breast <- mean[1] - mean[2]
se.breast <- sqrt(se[1]^2 + se[2]^2)
## Bottle-fed
theta.bottle <- mean[3] - mean[4]
se.bottle <- sqrt(se[3]^2 + se[4]^2)
```

Treatment Effects in Subgroups

- Breast-fed babies:

```
printWaldCI(theta.breast, se.breast)
##          Effect 95% Confidence Interval P-value
## [1,] 0.037   from -0.057 to 0.131      0.44
```

- Bottle-fed babies:

```
printWaldCI(theta.bottle, se.bottle)
##          Effect 95% Confidence Interval P-value
## [1,] 0.105   from 0.049 to 0.161      0.0002
```

The Correct Analysis of Subgroups

Test for interaction

- Suppose we have two subgroups defined by males (M) and females (F). To investigate if the treatment effect θ differs between subgroups we consider the **subgroup difference**

$$\Delta\hat{\theta} = \hat{\theta}^M - \hat{\theta}^F$$

where $\hat{\theta}^M$ and $\hat{\theta}^F$ are the estimated treatment effects for males and females, respectively.

- Based on the standard error

$$\text{se}(\Delta\hat{\theta}) = \sqrt{\text{se}^2(\hat{\theta}^M) + \text{se}^2(\hat{\theta}^F)},$$

tests and confidence intervals can be constructed.

Example Revisited

The test for interaction can be performed with the function `printWaldCI()` from `biostatUZH`.

```
## Compare effect sizes
theta.diff <- theta.breast - theta.bottle
se.diff <- sqrt(se.breast^2 + se.bottle^2)
printWaldCI(theta.diff, se.diff)

##          Effect 95% Confidence Interval P-value
## [1,] -0.068 from -0.177 to 0.041      0.22
```

- There is **no evidence** for a different treatment effect in the two subgroups ($p = 0.22$).
- Perhaps the study sample size and specifically the breast-fed group was too small to provide such evidence.

Methods of Selecting Subgroups

- The risk of **false positive findings** increases with increasing number of subgroup comparisons.
- For example, if we compare treatment effects in 20 different subgroups, then we would expect one of the analyses to yield a significant result (at $\alpha = 5\%$), even if the treatment effect is the same in all subgroups (under the null hypothesis H_0).
- If the subgroups have been selected **before** the data were collected based on biological or clinical reasoning, then positive (significant) findings are more trustworthy than if subgroups have been defined *post hoc*.

Qualitative and Quantitative Interactions

- The **test for interaction** has null hypothesis

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_I$$

of equal treatment effects θ_i in $i = 1, \dots, I$ subgroups.

- This is known as the **Q-test** in meta-analysis.
- A useful distinction between **quantitative** and **qualitative** interactions (subgroup differences) has been made.
- For a **quantitative interaction**, the size of the treatment effect θ_i varies between subgroups, but is always **in the same direction**.
- For a **qualitative interaction**, also the **direction of the treatment effect varies** between subgroups.

The Method of Gail and Simon

- It has been argued that quantitative interactions are not surprising, whereas qualitative interactions are of greater clinical importance.
- A modified test for interaction can be constructed, with null hypothesis that there is no **qualitative** interaction (Gail and Simon, 1985), i.e.

$$H_0 : \theta_i \geq 0 \text{ for all } i \text{ or } \theta_i \leq 0 \text{ for all } i.$$

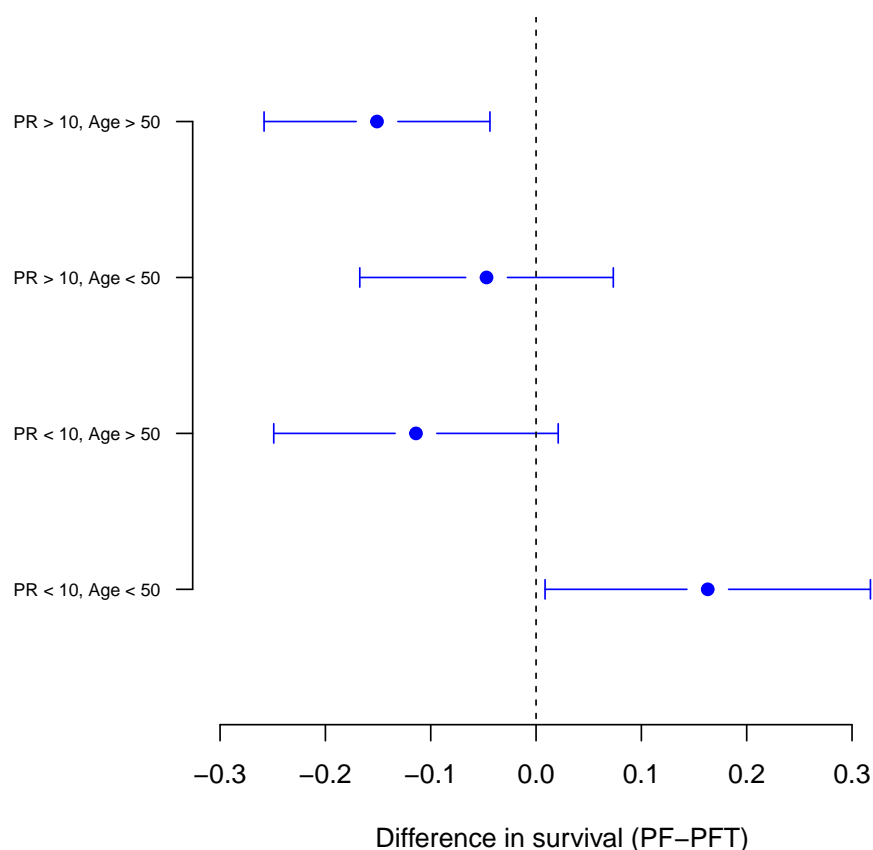
Example: Treatment of Breast Cancer Trial

From Matthews (2006), Section 9.2

- The outcome of the trial was disease-free survival at 3 years, comparing chemotherapy (PF) only versus chemotherapy plus tamoxifen (PFT).
- Risk difference θ is positive if disease free survival is more likely under PF.
- Four subgroups based on the progesterone receptor status (PR) and age are considered:

	PR < 10 fmol		PR ≥ 10 fmol	
	Age < 50	Age ≥ 50	Age < 50	Age ≥ 50
Risk difference $\hat{\theta}$	0.163	-0.114	-0.047	-0.151
se($\hat{\theta}$)	0.0788	0.0689	0.0614	0.0547
$\hat{\theta}/\text{se}(\hat{\theta})$	2.07	-1.65	-0.77	-2.76
p-value	0.039	0.098	0.44	0.006

Treatment Effects with Confidence Intervals



The Method of Gail-Simon

- Compute the test statistic $Z_i = \hat{\theta}_i / \text{se}(\hat{\theta}_i)$ in each subgroup i and the overall test statistic $T = \min\{Q^+, Q^-\}$, where

$$Q^+ = \sum_{i: \hat{\theta}_i \geq 0} Z_i^2 \quad \text{and} \quad Q^- = \sum_{i: \hat{\theta}_i < 0} Z_i^2.$$

- For the breast cancer trial we obtain $T = 2.07^2 = 4.28$.
- The reference distribution for T is non-standard, so we use statistical software to calculate a p -value.

Example: Treatment of Breast Cancer Trial

Tests for Interaction in R

```
rd <- c(0.163, -0.114, -0.047, -0.151)
rd.se <- c(0.0788, 0.0689, 0.0614, 0.0547)
## Q-test for quantitative interaction
library(meta)
mymeta <- metagen(TE=rd, seTE=rd.se)
print(formatPval(mymeta$pval.Q))

## [1] "0.01"

## Test for qualitative interaction
library(biostatUZH)
pGS <- gailSimon(thetahat = rd, se = rd.se)
print(formatPval(pGS["p-value"]))

## p-value
## "0.088"
```

- Moderate evidence for a quantitative interaction ($p = 0.01$)
- Weak evidence for a qualitative interaction ($p = 0.088$)

Clinical Biostatistics

Leonhard Held, Stefanie von Felten

University of Zurich

SHARE-CTD 1st Schooling Event

27.-31. January 2025

Reisensburg, Germany

Complete lecture notes at

https://bookdown.org/charlotte_micheloud93/Clinical_Biostatistics

Lecture 3: Sequential Methods

Monitoring Accumulating Data

Data Monitoring Committees

Group Sequential Methods

Stopping Rules

Monitoring Accumulating Data

- Adequate evidence to settle which treatment is superior may have accumulated long before a clinical trial runs to its planned conclusion.
- The **ethical issue** emerges, that patients may be receiving a treatment, that could have been known to be inferior **at the time of treatment**.
- Application of **repeated significance tests** (RST) at several **interim analyses**, where a decision will be made whether or not to stop the trial.
- Neyman-Pearson hypothesis test suitable
- However, **statistical issues** emerge to ensure that the Type I error rate α is maintained.

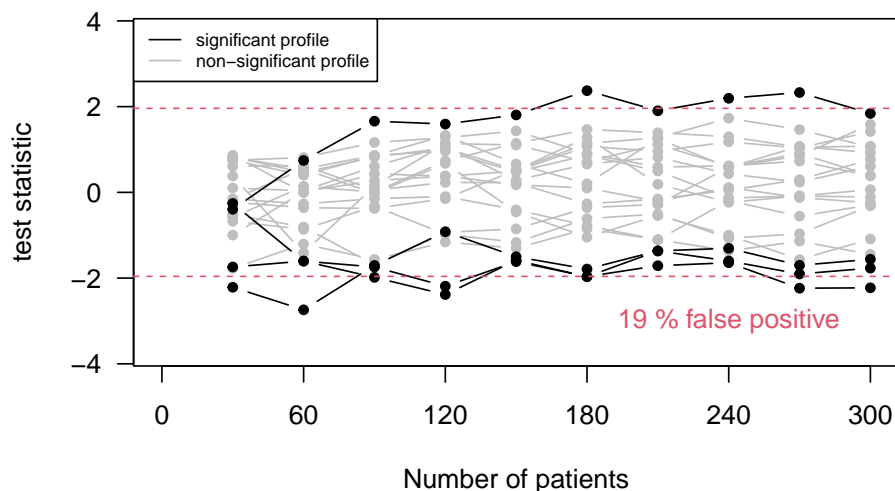
Data Monitoring Committees

- The decision to terminate a trial can be based on:
 - **Efficacy** of new treatment
 - Worryingly high incidence of **side effects**
 - Evidence that the new treatment is **less efficacious** than the existing treatment.
 - **Futility**, i.e., there is little chance of showing that the new treatment is better.
- A **data (and safety) monitoring committee** (D[S]MC) (with clinicians, statisticians, ...) periodically reviews the evidence currently available from the trial.
- This is done at a relatively **small number** of times and may require **unblinded** study information.
- Extensive use of DMCs has led to the widespread use of **group-sequential** methods
- Note: “group” no longer refers to treatment group, but to **successive groups of patients** used at each interim analysis.

Group Sequential Methods: A Simulation Study

Standard significance threshold

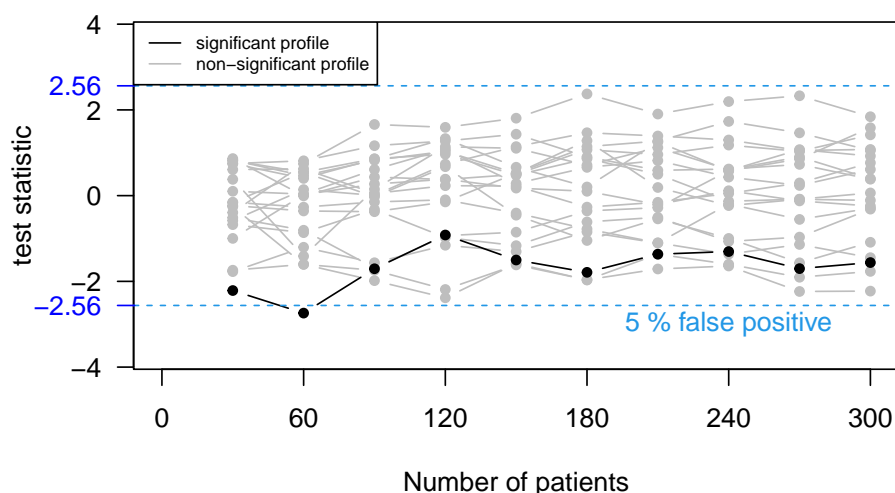
- **Accumulating data** is generated without treatment effect.
 - Each simulation is analysed in 10 groups of 30 patients.
 - Only 20 profiles out of 10000 simulations are shown below.
- 4 profiles are significant at standard significance threshold 1.96, false positive rate over all 10000 simulations is 19%.



Group Sequential Methods: A Simulation Study

Adjusted significance threshold

- 1 profile is significant at **adjusted** significance threshold 2.56, false positive rate over all 10000 simulations is 5%.

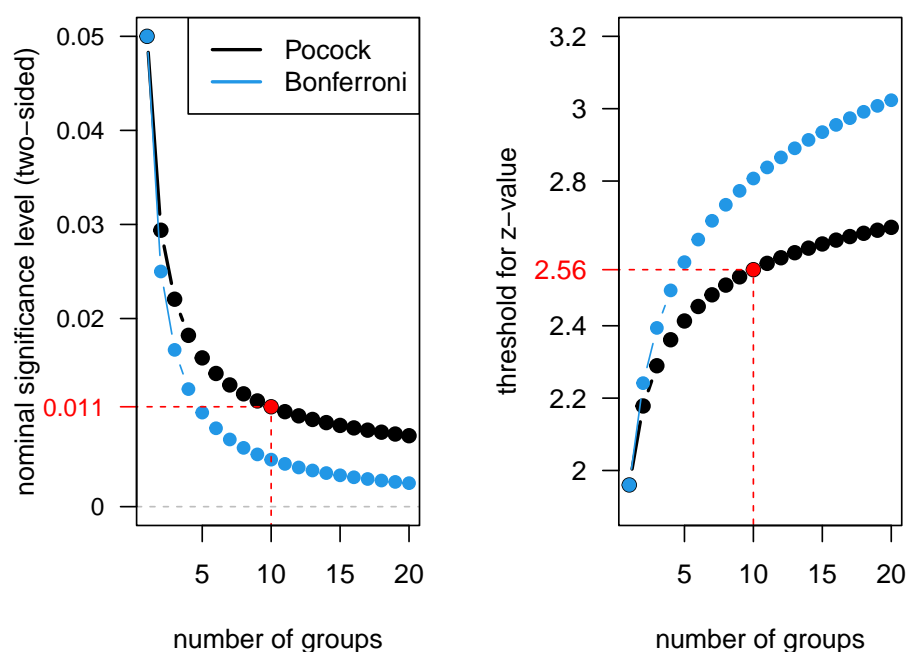


Group Sequential Methods

- Data analysis in **interim analyses** after every successive group of $2n$ patients e.g. $2n = 20$ or 30 .
- Fix a maximum number of N groups.
- A trial is stopped at interim
 - if the (two-sided) P -value is smaller than a pre-specified **nominal significance level** $\tilde{\alpha}$,
 - or if N groups of patients have been recruited.
- The nominal significance level $\tilde{\alpha}$ depends on the Type I error rate α and the number of groups N .
- Standard adjustments for multiple testing are too conservative, since tests are based on accumulating data with a specific dependence structure.

Group Sequential Methods

Pocock nominal significance level



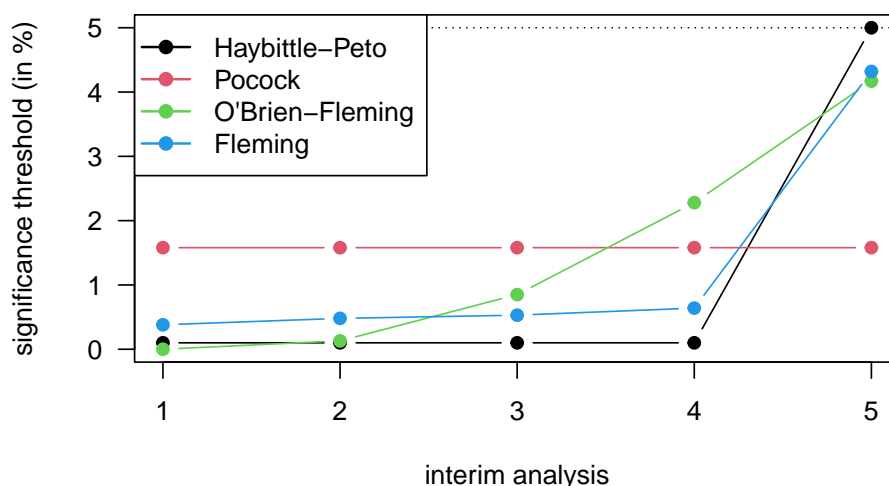
Stopping Rules

- The previously introduced method, in which each test uses the same nominal significance level $\tilde{\alpha}$ is known as **Pocock stopping rule**.
- It has two disadvantages:
 - It is not too difficult for a trial to be halted early, which is considered as undesirable and unconvincing by some authors.
 - Suppose the trial terminates at the final analysis with $\tilde{\alpha} < p < \alpha$. Many clinicians find it difficult to accept that the result of this trial is non-significant!
- Better use a **stopping rule** a_1, a_2, \dots, a_N where a_j is the **significance threshold** at the j -th interim with
 - a_1 **very small** and
 - a_j gradually increasing to a_N **close to** α .

Various Stopping Rules

All the following stopping rules control the overall Type I error rate $\alpha = 5\%$ with $N = 5$ maximal interim analyses, the **Haybittle-Peto** approach only approximately.

Method	Interim analysis				
	1	2	3	4	5
Haybittle-Peto (1971)	0.001	0.001	0.001	0.001	0.05
Pocock (1977)	0.0158	0.0158	0.0158	0.0158	0.0158
O'Brien-Fleming (1979)	5×10^{-6}	0.0013	0.0085	0.0228	0.0417
Fleming <i>et al.</i> (1984)	0.0038	0.0048	0.0053	0.0064	0.0432



Implementation in R

R package `gsDesign`

- `gsDesign` derives group sequential clinical trial designs and describes their properties
- argument `alpha`: **one-sided** significance level
- argument `test.type`: types 1–6 available, but we only use `test.type=1` (one-sided), i.e., we have no interest in stopping early for futility (lower bound)
- we only use `gsDesign` for relatively simple designs, but many more things are possible

Alternative: R package `rpact`

Implementation in R

Pocock stopping rule with 3 groups

```
library(gsDesign)
gsDesign(k=3, sfu="Pocock", test.type=1)

## One-sided group sequential design with
## 90 % power and 2.5 % Type I Error.
##           Sample
##           Size
## Analysis Ratio* Z   Nominal p   Spend
##           1  0.384 2.29      0.011 0.0110
##           2  0.767 2.29      0.011 0.0079
##           3  1.151 2.29      0.011 0.0060
##           Total                                0.0250
##
## ++ alpha spending:
## Pocock boundary.
## * Sample size ratio compared to fixed design with no interim
##
## Boundary crossing probabilities and expected sample size
## assume any cross stops the trial
##
## Upper boundary (power or Type I Error)
##           Analysis
##           Theta   1   2   3 Total   E{N}
##           0.0000 0.011 0.0079 0.0060 0.025 1.1391
##           3.2415 0.389 0.3421 0.1689 0.900 0.7210
```

Implementation in R

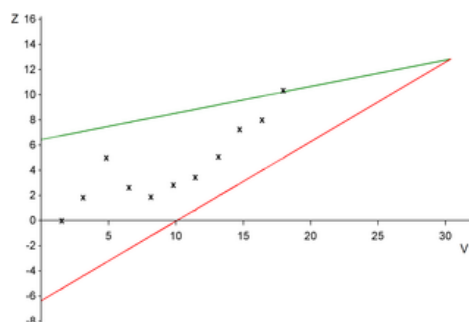
O'Brien-Fleming stopping rule with 3 groups

```
library(gsDesign)
gsDesign(k=3, sfu="OF", test.type=1)

## One-sided group sequential design with
## 90 % power and 2.5 % Type I Error.
##      Sample
##      Size
## Analysis Ratio* Z   Nominal p   Spend
##      1  0.339 3.47   0.0003 0.0003
##      2  0.677 2.45   0.0071 0.0069
##      3  1.016 2.00   0.0225 0.0178
##      Total                                0.0250
##
## ++ alpha spending:
## O'Brien-Fleming boundary.
## * Sample size ratio compared to fixed design with no interim
##
## Boundary crossing probabilities and expected sample size
## assume any cross stops the trial
##
## Upper boundary (power or Type I Error)
##      Analysis
##      Theta      1      2      3 Total   E{N}
##      0.0000 0.0003 0.0069 0.0178 0.025 1.0136
##      3.2415 0.0565 0.5288 0.3147 0.900 0.7987
```

Other Forms of Stopping Rules

- A more flexible approach based on the Lan-DeMets **alpha spending function** does not require the maximum number of interim analyses to be specified in advance.
- Another popular approach to analyse accumulating data is Whitehead's **triangular test** based on the score statistic Z and the Fisher information V :



Crossing of green/red line → positive/negative trial conclusion

Problems of Stopping at Interim

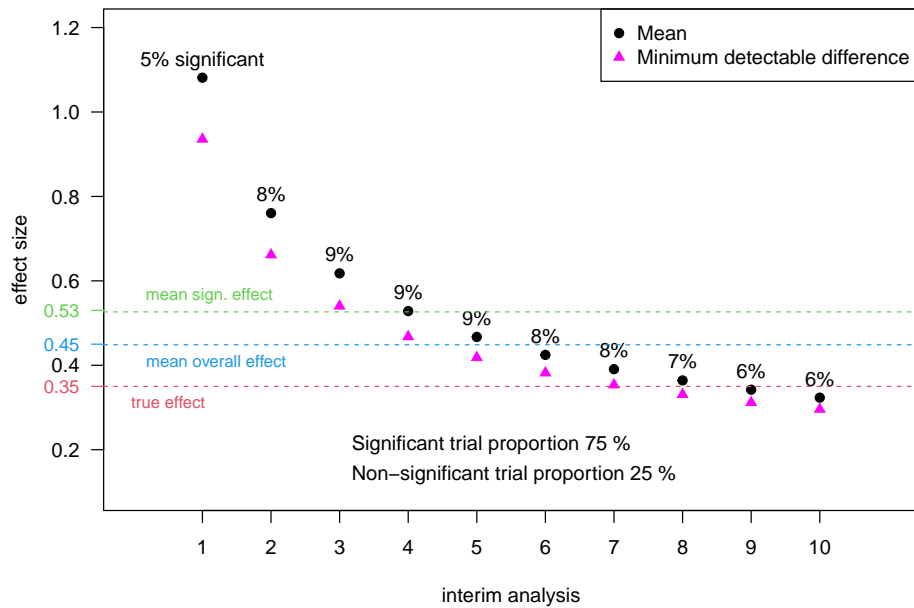
- If a trial terminates early, then there are problems with obtaining unbiased treatment effects due to the sequential nature of the trial.
- If a traditional analysis is performed in a trial that stops at interim because treatment is found to be significantly better than control, then
 - the treatment effect estimate will be **too large**,
 - the confidence interval will be **too narrow**,
 - the P value will be **too small**.
- <https://vimeo.com/231014768>
- Advanced methods for attempting to correct this bias are available, but rarely used.

Simulation

- Two equally sized treatment arms, continuous primary outcome
- Difference in means θ as effect size
- $N = 10$ interim analyses, Pocock bound $z_P = 2.56$ (nominal significance level $\tilde{\alpha} = 0.011$)
- $n = 15$ patients per group and treatment arm
- True treatment effect $\theta = 0.35$, standard deviation $\sigma = 1$
- Note: Significance at the k -th interim analysis implies **minimum detectable difference** for treatment effect estimate

$$\hat{\theta} \geq z_P \sqrt{2/(k \cdot n)}$$

Simulation



- Mean effect size of all significant trials is $0.53 > 0.35$
- Mean effect size of all non-significant trials is $0.23 < 0.35$
- Mean effect size of all trials is $0.45 > 0.35$: **stopping bias!**

Leonhard Held, Stefanie von Felten

University of Zurich

SHARE-CTD 1st Schooling Event

27.-31. January 2025

Reisensburg, Germany

Complete lecture notes at

https://bookdown.org/charlotte_micheloud93/Clinical_Biostatistics

Lecture 4: Analysis of Binary Outcomes

Comparison of Two Proportions

Statistical Tests

χ^2 -Test

Fisher's Exact Test

Effect Measures and Confidence Intervals

Absolute Risk Reduction

Number Needed to Treat

Relative Risk

Relative Risk Reduction

Odds Ratio

Adjusting for Baseline Observations

Logistic Regression

Example: APSAC Study

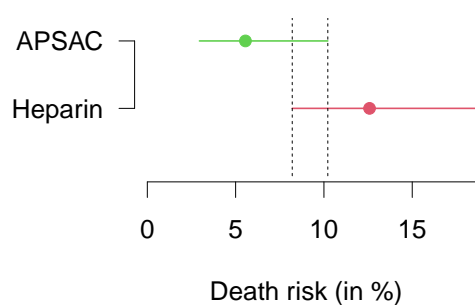
Meinerz *et al.* , Am J Card, 1988

- RCT to compare a new thrombolytikum (APSAC) with standard treatment (Heparin) in patients with acute cardiac infarction.
- Outcome: Mortality within 28 days of hospital stay.
- Results:

Therapy	Dead	Survived	Total	Percentage
APSAC	9	153	162	5.6%
Heparin	19	132	151	12.6%

Example: APSAC Study

	dead	alive	total	Percent dead	Standard error	95% Wilson-CI
APSAC	9	153	162	5.6%	1.8%	3.0 to 10.2%
Heparin	19	132	151	12.6%	2.7%	8.2 to 18.8%



- The 95% Wilson confidence intervals overlap.
- However, this does not necessarily indicate a non-significant treatment effect.

χ^2 -Test

Expected number of cases and test statistics

```
## observed
APSAC.observed

##          dead alive
## APSAC      9   153
## Heparin    19   132

## expected
round(APSAC.expected <- chisq.test(APSAC.observed)$expected, 2)

##          dead  alive
## APSAC    14.49 147.51
## Heparin  13.51 137.49

## test statistic
print(sum((APSAC.observed-APSAC.expected)^2/APSAC.expected))

## [1] 4.738067
```

χ^2 -Test

P-values

```
chisq.test(APSAC.observed, correct=FALSE)

##
## Pearson's Chi-squared test
##
## data: APSAC.observed
## X-squared = 4.7381, df = 1, p-value = 0.0295

chisq.test(APSAC.observed)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: APSAC.observed
## X-squared = 3.9146, df = 1, p-value = 0.04787
```

Fisher's Test

```
fisher.test(APSAC.observed)

##
##  Fisher's Exact Test for Count Data
##
## data:  APSAC.observed
## p-value = 0.04588
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1576402 0.9892510
## sample estimates:
## odds ratio
##  0.409812
```

Function `fisher.test()` also provides an estimate (“conditional MLE”) of the **odds ratio** (with 95% CI).

Effect Measures

- Notation: π_0 and π_1 are the “true” risks of death in the control and intervention group, resp. Assume $\pi_0 \geq \pi_1$ holds.
- The following quantities are used to compare π_0 and π_1 :

$$\text{The absolute risk reduction ARR} = \pi_0 - \pi_1$$

$$\text{The number needed to treat NNT} = 1/\text{ARR}$$

$$\text{The relative risk RR} = \pi_1/\pi_0$$

$$\text{The relative risk reduction RRR} = \frac{\text{ARR}}{\pi_0} = 1 - \text{RR}$$

$$\text{The odds ratio OR} = \frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)}$$

- **No difference** between groups (i.e. $\pi_0 = \pi_1$) corresponds to $\text{ARR} = \text{RRR} = 0$ and $\text{RR} = \text{OR} = 1$.

Absolute Risk Reduction

- Also called **risk difference** RD or **probability difference**
- Estimated **absolute risk reduction**:

$$\widehat{ARR} = 12.6\% - 5.6\% = 7\%$$

with standard error

$$se(\widehat{ARR}) = \sqrt{\frac{\hat{\pi}_0(1 - \hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1}} = 3.2\%$$

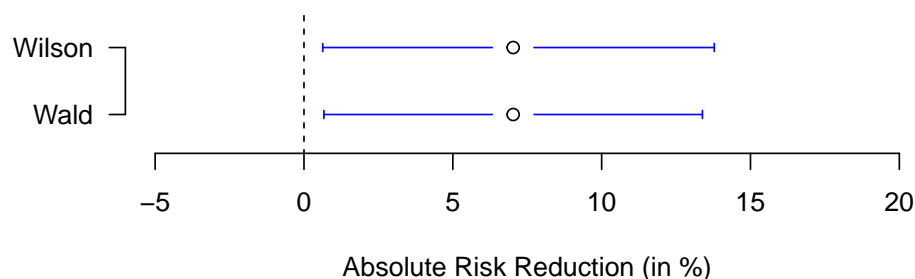
→ Additive Wald-CI for ARR

- An improved **Wilson-CI** for ARR can be calculated using the **“square-and-add”** approach.

CI for ARR

```
library(biostatUZH)
x <- c(19, 9)
n <- c(151, 162)
print(confIntRiskDiff(x, n))

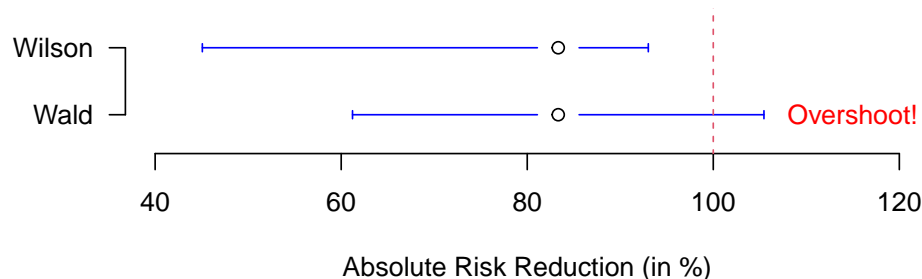
## $rd
##           [,1]
## [1,] 0.07027226
##
## $CIs
##      type      lower      upper
## 1  Wald 0.006691788 0.1338527
## 2 Wilson 0.006303897 0.1378381
```



CI for ARR for Artificial Data

```
x.art <- c(11,1)
n.art <- c(12,12)
print(confIntRiskDiff(x.art, n.art))

## $rd
##           [,1]
## [1,] 0.8333333
##
## $CIs
##      type      lower      upper
## 1  Wald 0.6121830 1.054484
## 2 Wilson 0.4507227 0.930162
```



Number Needed to Treat

- Suppose we have n patients in each treatment group. The expected number of deaths in the control and intervention group are:

$$N_0 = n\pi_0 \text{ and } N_1 = n\pi_1.$$

- Suppose we want the **expected difference**

$$N_0 - N_1 = n(\pi_0 - \pi_1)$$

to be one patient. The required sample size is

$$n = 1/(\pi_0 - \pi_1) = 1/\text{ARR}.$$

- This is the **number needed to treat**, the required number of patients to be treated with the intervention rather than control to avoid one death.
- Depending on the direction of the effect, this is also called **number needed to benefit** or **number needed to harm**.

Number Needed to Treat

Estimate and confidence interval

- Estimated NNT: $\widehat{NNT} = 1/\widehat{ARR} = 1/0.07 = 14.2$.
- Interpretation: To avoid **one** death, we need to treat 14.2 patients with APSAC rather than Heparin.
- A **confidence interval** for NNT can be obtained by inverting the limits of the CI for ARR:

```
(ci.arr <- confIntRiskDiff(x, n)$CIs[2,])

##      type      lower      upper
## 2 Wilson 0.006303897 0.1378381

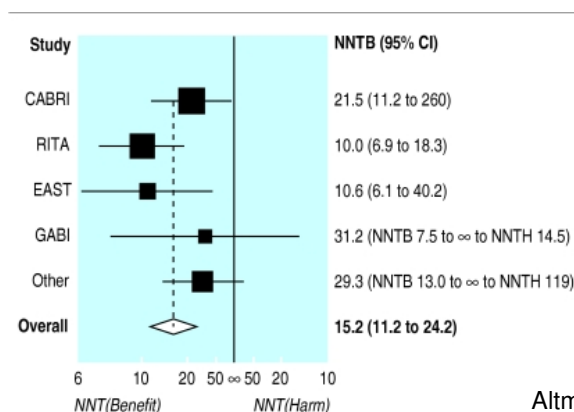
ci.ntt <- 1/ci.arr[c(3,2)]
colnames(ci.ntt) <- c("lower", "upper")
print(round(ci.ntt, 1))

##      lower upper
## 2       7.3 158.6
```

- CI for NNT: $1/0.138$ to $1/0.006 = 7.3$ to 158.6 .
- Note: CI for NNT is not well-defined if CI for ARR contains 0.

Confidence Intervals for NNT

- CI for NNT is not well-defined if CI for ARR contains 0.
- This problem can be circumvented by plotting NNT on the ARR scale.
- Example: Forest plot of data from randomised trials comparing bypass surgery with coronary angioplasty in relation to angina in one year:



Altman (1998, BMJ, 1309–12)

Relative Risk

- The estimated **death risk** is

$$x_1/n_1 = 9/162 = 5.6\% \text{ for APSAC,}$$
$$x_0/n_0 = 19/151 = 12.6\% \text{ for Heparin.}$$

- The estimated **relative risk** is therefore

$$\widehat{RR} = \frac{5.6\%}{12.6\%} = 0.44.$$

- Using the standard error of the **log** relative risk

$$se(\log(\widehat{RR})) = \sqrt{\frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_0} - \frac{1}{n_0}}$$

we can calculate a **multiplicative Wald-CI** for RR.

```
x <- c(9, 19)
n <- c(162, 151)
(ci.rr <- confIntRiskRatio(x, n))

##      lower Risk Ratio      upper
## 0.2061781 0.4415205 0.9454948
```

Relative Risk Reduction

- Estimated RRR:

$$\widehat{RRR} = \frac{\widehat{ARR}}{\pi_0} = \frac{0.07}{0.126} = 56\%$$
$$\text{or: } \widehat{RRR} = 1 - \widehat{RR} = 1 - 0.44 = 56\%$$

- A **confidence interval** for RRR can be obtained based on the limits L_{RR} and U_{RR} of the confidence interval for RR:

$$(1 - U_{RR}) \text{ to } (1 - L_{RR}) = (1 - 0.95) \text{ to } (1 - 0.21)$$
$$= 0.05 \text{ to } 0.79$$

- Interpretation: The risk of death with APSAC has been reduced by 56% (95% CI: 5% to 79%) compared to Heparin.

Absolute and Relative Effect Measures

<https://vimeo.com/231013132>

- Example: RCT with

$$\text{death risk} = \begin{cases} 0.3\% & \text{in the placebo group} \\ 0.1\% & \text{in the treatment group} \end{cases}$$

- Then we have $\text{ARR} = 0.2\% = 0.002$ and $\text{NNT} = 500$, so there is a **very small absolute effect** of treatment.
- However, we have $\text{RR} = 1/3$ and $\text{RRR} = 2/3 = 67\%$, so there is a **large relative effect** of treatment.
- We cannot transform absolute to relative effect measures (and vice versa) without knowledge of the underlying risks.

Odds Ratio

Therapy	Dead		Total
	Yes	No	
APSAC	$a = 9$	$b = 153$	162
Heparin	$c = 19$	$d = 132$	151
			$n = 313$

- The odds of death for APSAC are $a/b = 9/153$ and $c/d = 19/132$ for Heparin.
- The estimated **odds ratio** is therefore

$$\widehat{\text{OR}} = \frac{a/b}{c/d} = \frac{9/153}{19/132} = \frac{9 \cdot 132}{153 \cdot 19} = 0.41.$$

- The formulation $\widehat{\text{OR}} = (a \cdot d)/(b \cdot c)$ motivates the alternative name **cross-product ratio**.

Odds Ratio

Standard error and confidence interval

As for RR we calculate the standard error of the odds ratio on the log scale:

$$se(\log(\widehat{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

to compute a **multiplicative Wald-CI** for OR:

```
x <- c(9, 19)
n <- c(162, 151)
(ci.r <- confIntOddsRatio(x, n))

##      lower Odds Ratio      upper
## 0.1788117 0.4086687 0.9339999
```

Odds Ratio

Confidence intervals and sample size

Confidence intervals for odds ratios may differ even if group sample sizes remain the same:

```
x <- c(25, 25)
n <- c(50, 50)
(ci.or1 <- confIntOddsRatio(x, n))

##      lower Odds Ratio      upper
## 0.4565826 1.0000000 2.1901841

x <- c(1, 1)
n <- c(50, 50)
(ci.or2 <- confIntOddsRatio(x, n))

##      lower Odds Ratio      upper
## 0.06081319 1.00000000 16.44380070
```

What matters are the **number of events!**

Sensitivity Analysis

P-Values for ARR, RR and OR

- P -values can be obtained for each effect estimate (ARR, RR or OR) with the corresponding standard error (log scale for RR and OR)
- The one for OR is called Asymptotic P -value in `twoby2()`.

Method	p -value
ARR	$p = 0.03$
RR	$p = 0.035$
OR	$p = 0.034$
χ^2 -test	$p = 0.03$
χ^2 -test with continuity correction	$p = 0.046$
Fisher's exact test	$p = 0.048$

Adjusting for Baseline Observations

- For binary outcomes, adjusting for baseline observations is usually done using **logistic regression** and will produce **adjusted odds ratios**.
- Alternatively, the **Mantel-Haenszel (MH) method** can be used, which gives a **weighted average** of strata-specific odds ratios.
- Adjustments for continuous variables using MH is only possible after suitable categorization.

Example: PUVA Trial

- PUVA (drug followed by UVA exposure) versus TL-01 lamp therapy for treatment of psoriasis
 - Primary outcome: Was patient clear of psoriasis at or before the end of the treatment?
 - Treatment allocation used RPBs stratified according to whether **predominant plaque size** was large or small.
- Balanced distribution in treatment arms (29:22 vs. 28:21)

```
print(puva)

##   plaqueSize treatment cleared total
## 1      Small    TL-01      23     29
## 3      Small     PUVA      25     28
## 2      Large    TL-01       9     22
## 4      Large     PUVA      16     21
```

Unadjusted Analysis

Logistic regression

```
m1 <- glm(cbind(cleared, total-cleared) ~ treatment,
          data=puva, family=binomial)
## tableRegression gives profile confidence intervals
tableRegression(m1)
```

	Odds Ratio	95%-confidence interval	p-value
treatmentTL-01	0.33	from 0.12 to 0.82	0.021

Strata-Specific Estimates

```
m2Small <- glm(cbind(cleared, total-cleared) ~ treatment,
               subset=(plaqueSize=="Small"), data=puva, family=binomial)
tableRegression(m2Small)
```

	Odds Ratio	95%-confidence interval	p-value
treatmentTL-01	0.46	from 0.09 to 1.96	0.31

```
m2Large <- glm(cbind(cleared, total-cleared) ~ treatment,
               subset=(plaqueSize=="Large"), data=puva, family=binomial)
tableRegression(m2Large)
```

	Odds Ratio	95%-confidence interval	p-value
treatmentTL-01	0.22	from 0.05 to 0.77	0.023

Adjusted Analysis with Logistic Regression

```
m3 <- glm(cbind(cleared, total-cleared) ~ treatment + plaqueSize,
          data=puva, family=binomial)
tableRegression(m3)
```

	Odds Ratio	95%-confidence interval	p-value
treatmentTL-01	0.30	from 0.10 to 0.78	0.017
plaqueSizeLarge	0.24	from 0.09 to 0.61	0.004

- The adjusted treatment effect is $OR = 0.30$ with 95% CI from 0.10 to 0.78.
- The model also quantifies the effect of the variable used for adjustment, here plaque size, and gives a better model fit:

```
anova(m1, m3)
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(cleared, total - cleared) ~ treatment
## Model 2: cbind(cleared, total - cleared) ~ treatment + plaqueSize
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         2     9.5077
## 2         1     0.5426 1    8.9651 0.002752 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Leonhard Held, Stefanie von Felten

University of Zurich

SHARE-CTD 1st Schooling Event

27.-31. January 2025

Reisensburg, Germany

Complete lecture notes at

https://bookdown.org/charlotte_micheloud93/Clinical_Biostatistics

Lecture 5: Survival Analysis

Analysis of Survival Outcomes

Life Table Method

Kaplan-Meier Estimate of the Survival Function

Median Survival Time

Comparison of Survival Curves

Log-Rank Test

Hazard Rate

Hazard Ratio

The Cox Model

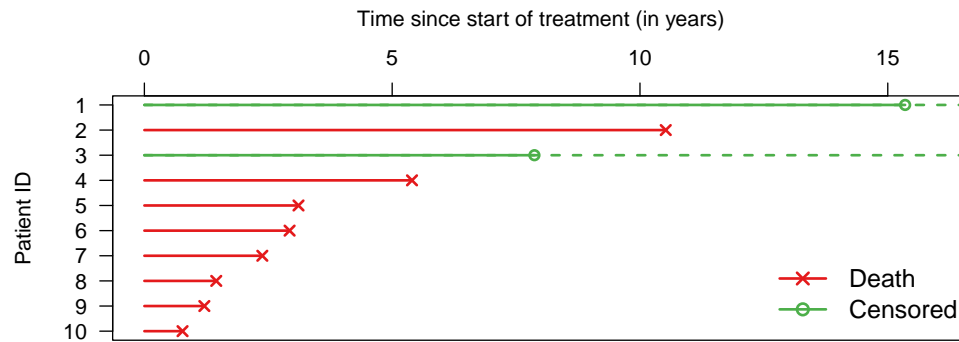
Survival Outcomes

- A common form of data in clinical trials is survival (time to event) data.
- Survival outcomes can often be considered as **continuous**, but are usually quite skewed.
- The issue of **censoring** requires special techniques for statistical analysis.

Example: Chemotherapy-Study

- Randomised study to compare two treatments of **head and neck cancer** on 224 patients
- Combination of Radiotherapy and Chemotherapy (RT+CT) ($n = 112$) versus Radiotherapy alone (RT) ($n = 112$)
- Outcome: Survival time (in years) from start of treatment

Data from Selected Patients



OS = „Overall survival“ = survival time (in years), +: censored

LK = Lymph nodes

PS = Performance status

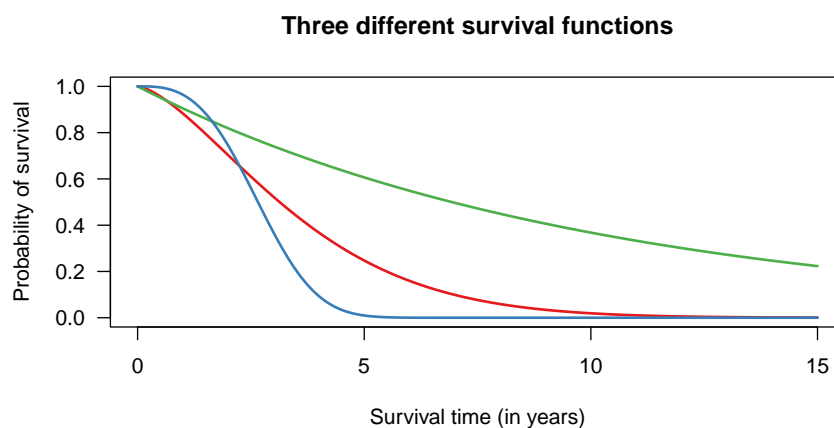
ID	Sex	Age	Treatment	LK	Stage	PS	OS
1	Female	57.4	RT+CT	2c	4	1	15.3+
2	Male	66.9	RT+CT	0	3	1	10.5
3	Male	68.3	RT+CT	0	3	1	7.9+
4	Male	51.6	RT	0	2	0	5.4
5	Male	38.8	RT	2c	4	0	3.1

Survival Function and Curve

- The **survival function** $S(t)$ gives the probability of survival at least up to time t :

$$S(t) = \Pr(T > t)$$

- The **survival curve** is a plot of $S(t)$ versus t .



Life Table Method

- The **life table** method is a classical approach to estimate the survival function $S(t)$.
- It is based on a partition of the time axis into time intervals $i = 1, \dots, I$ of typically equal length.
- The **risk of death** r_i in interval i is estimated as

$$\hat{r}_i = \frac{\text{\# patients which died in interval } i}{\underbrace{\text{\# patients at risk in interval } i}_{\text{risk set}}} = \frac{d_i}{n_i}$$

- By convention, observations censored in interval i contribute half to the number of patients at risk.
- Note that the **size of the risk set** n_i is monotonically decreasing as a function of time.

Life Table Method

- The **survival function** is then estimated as

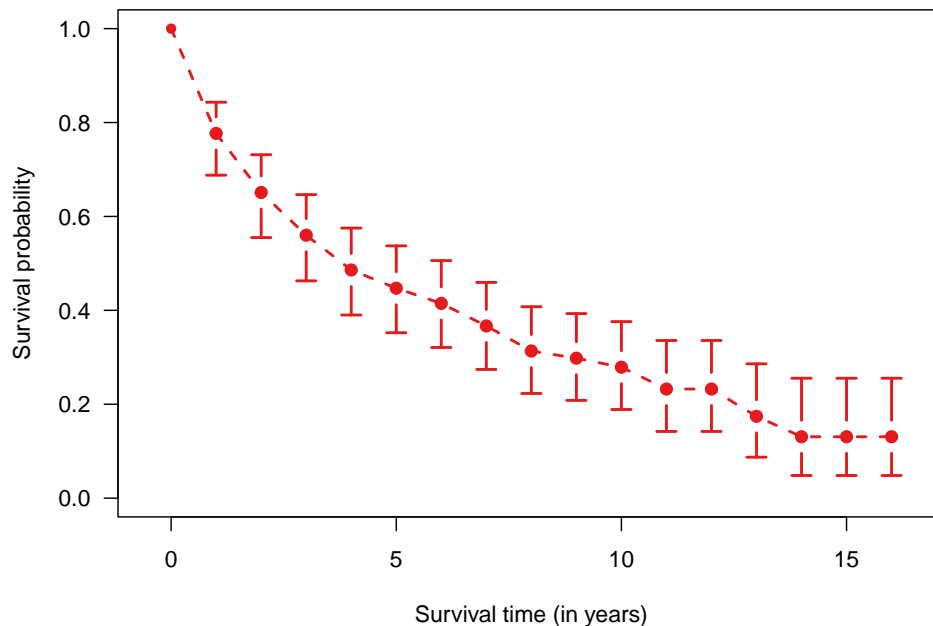
$$\hat{S}(i) = \prod_{t=1}^i (1 - \hat{r}_t) = (1 - \hat{r}_1) \times \dots \times (1 - \hat{r}_i)$$

- Wald confidence intervals can be calculated for $S(i)$, $\log \hat{S}(i)$ or $\log\{-\log \hat{S}(i)\}$ with appropriate back-transformation.
- The corresponding standard errors can be found in textbooks, e.g. Collett (2014), Section 2.2.

Life Table Method in Chemotherapy-Study

RT+CT group

With 95% confidence intervals for every year.



Kaplan-Meier Estimate of the Survival Function

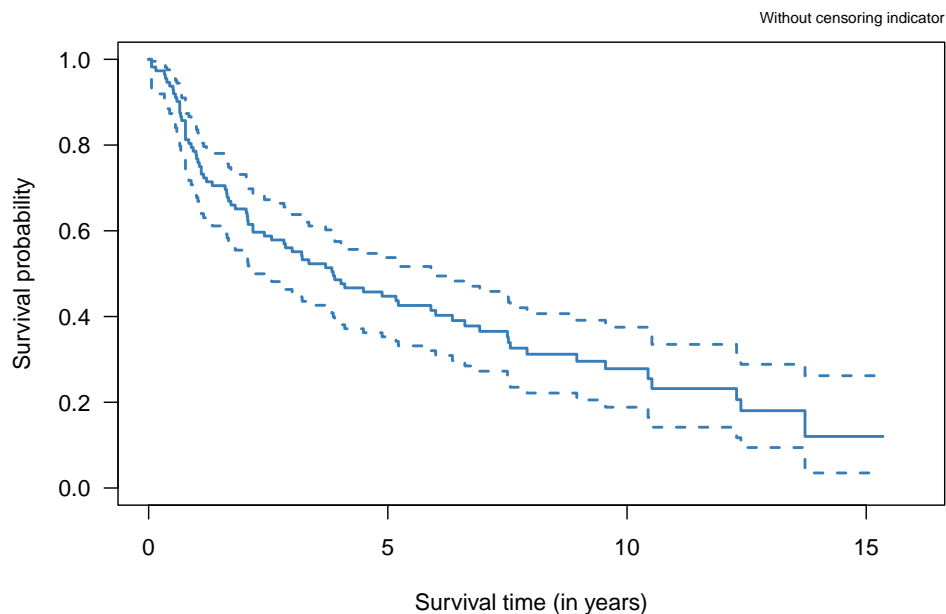
- In many studies the **exact follow-up time** for each individual is known.
- Aggregation in arbitrary time intervals can then be avoided.
- The **Kaplan-Meier estimate** of the survival function uses intervals, which contain only one event, so $d_i \in \{0, 1\}$ (if there are no ties).
- Application of the life table method to these intervals yields the Kaplan-Meier estimate, a step function with jumps of relative size
$$1 / \text{number of subjects at risk}$$

at the time of each event.

Kaplan-Meier Estimate

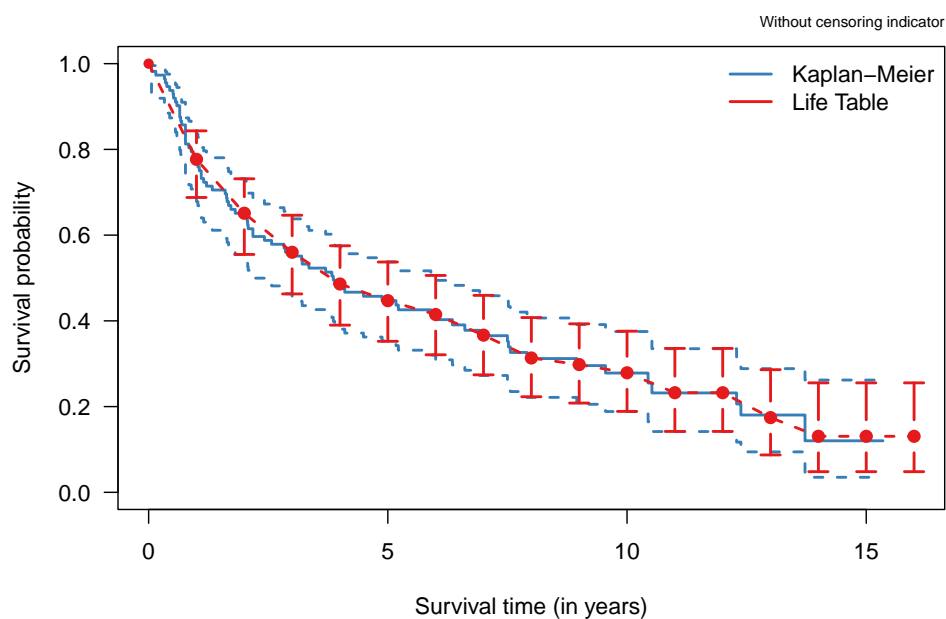
RT+CT group

```
par(mar = c(4.5, 4, 1.7, 1), las = 1)
sakk.surv <- survfit(Surv(os, death) ~ 1, data = sakk.rtct, conf.type = "log-log")
plot(sakk.surv, conf.int = T, mark.time = FALSE, lwd = lwd.lines, col = col.blue, log = F,
     xlab = "Survival time (in years)", ylab = "Survival probability", xlim = c(0, 1max))
mtext("Without censoring indicator", side = 3, line = 0.5, adj = 1, cex = 0.7)
```



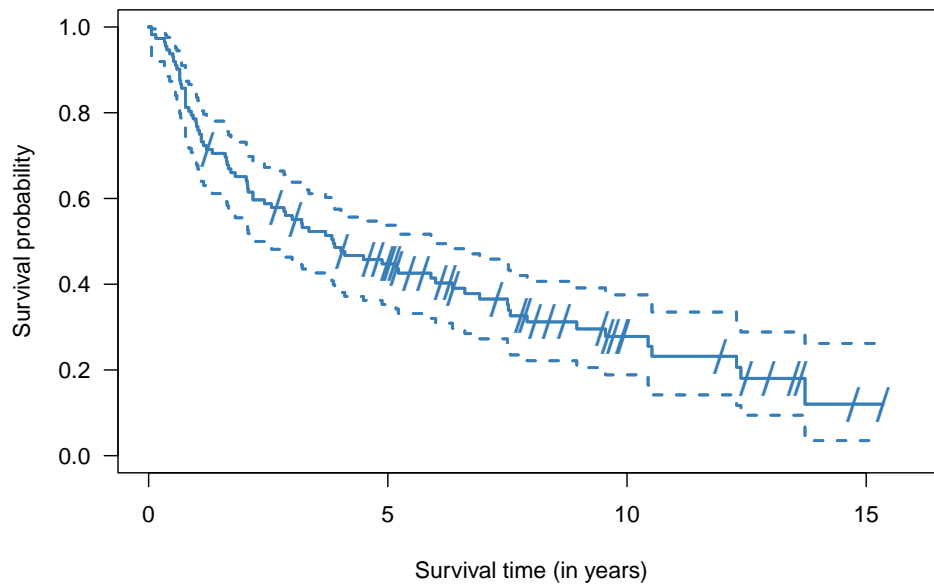
Comparison with Life Table Method

RT+CT group



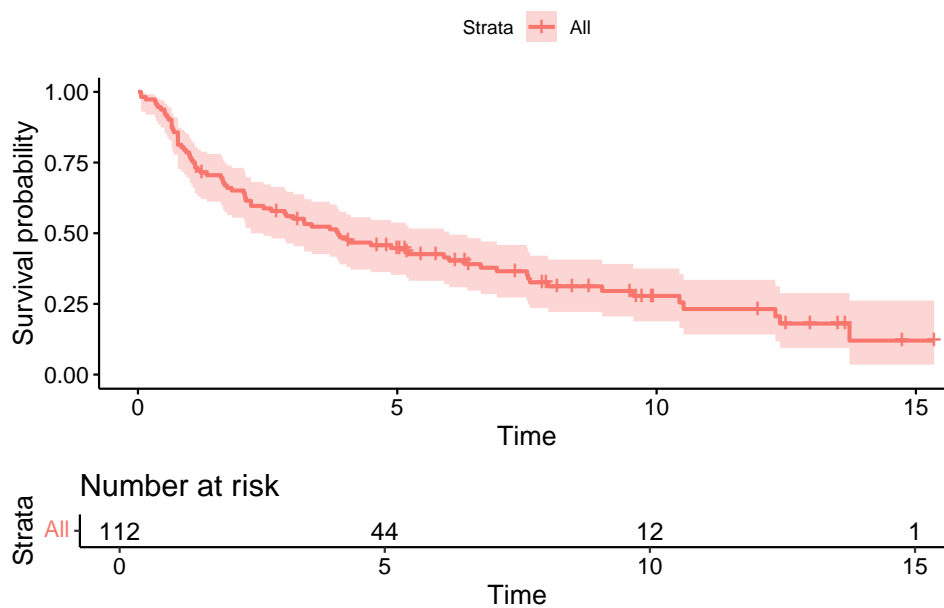
Kaplan-Meier Estimate with Censoring Indicator

RT+CT group



Kaplan-Meier Estimate with Risk Table

```
library(survminer)
ggsurvplot(sakk.surv, risk.table = TRUE)
```

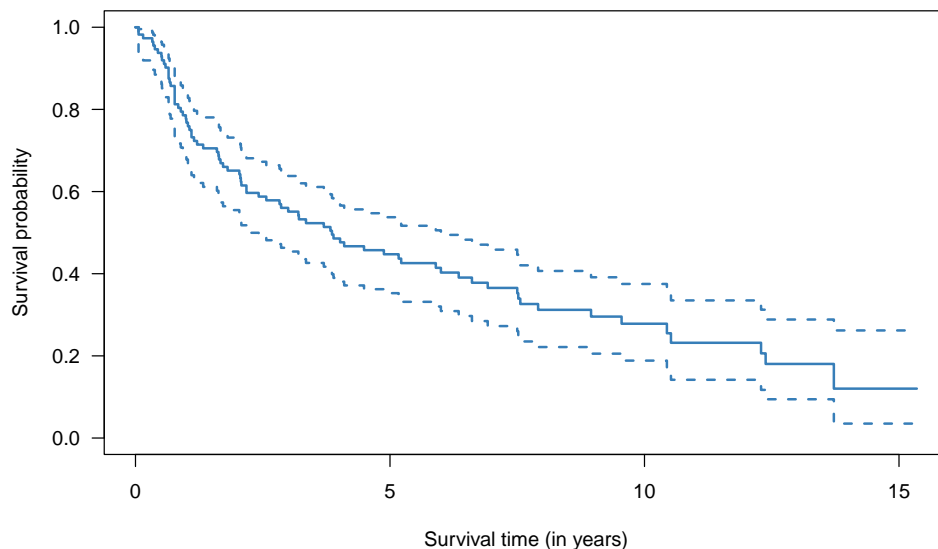


The Median Survival Time

- The **median survival time** t_{Med} (and other quantiles) can be easily read off the Kaplan-Meier estimate.
- Also a confidence interval can be derived in this way.
- The “square-and-add” method can be used to compute a confidence interval for the **difference in median survival time** between two groups.

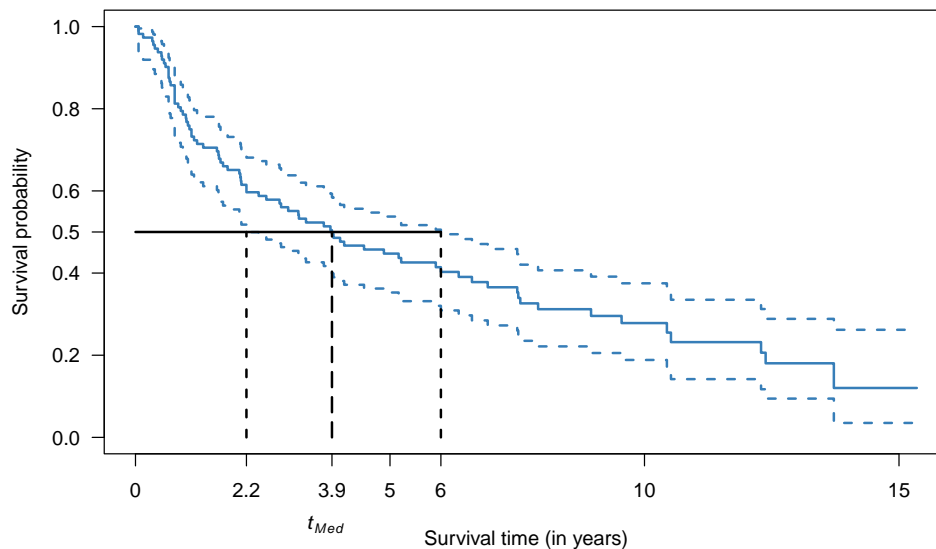
Graphical Illustration

RT+CT group



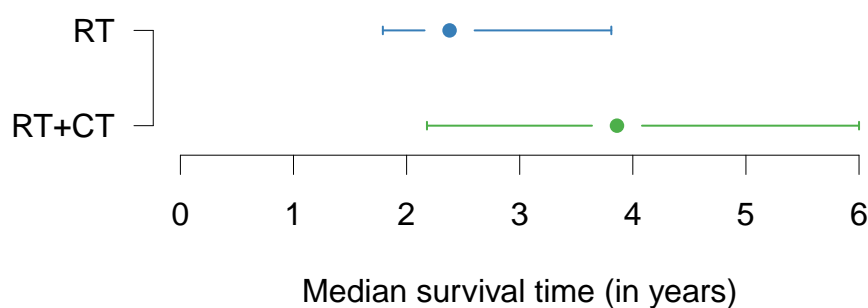
Graphical Illustration

RT+CT group



Comparison of Median Survival Time

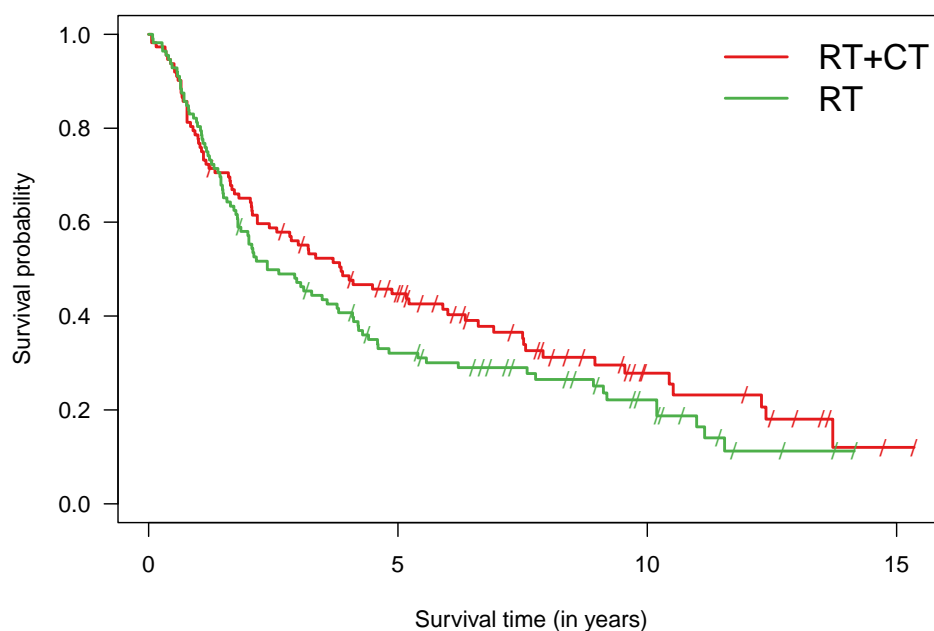
	n	Number of Deaths	t_{Med} (in years)	95%-CI (in years)
RT+CT	112	79	3.9	from 2.2 to 6.0
RT	112	88	2.4	from 1.8 to 3.8



```
confIntSquareAdd(theta1=3.9, lower1=2.2, upper1=6.0,
                  theta2=2.4, lower2=1.8, upper2=3.8)
```

```
## $difference
## [1] 1.5
##
## $CI
##      lower      upper
## 1 -0.7022716 3.684033
```

Comparison of Survival Curves



Comparison of Survival Curves: Log-Rank-Test

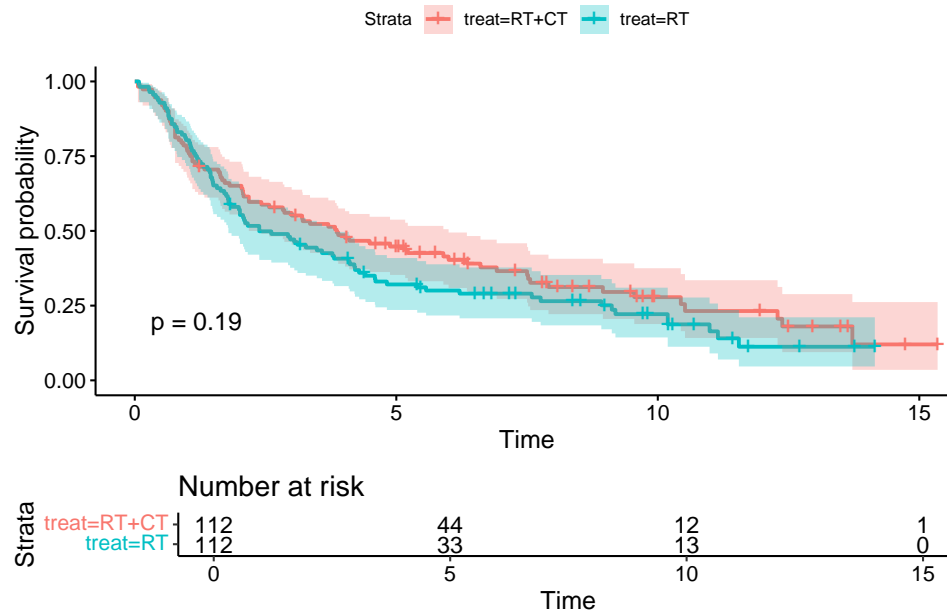
- The **Log-Rank Test** quantifies the evidence against the null hypothesis of equal survival curves $S_A(t) = S_B(t)$ for all times t with a p -value.
- The method compares the **observed number** of events with the corresponding **expected number** of events (under the null hypothesis H_0) in each group.

```
(survdif(Surv(time, death) ~ treat, data = sak))

## Call:
## survdif(formula = Surv(time, death) ~ treat, data = sak)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## treat=RT+CT 112         79      87.5      0.822      1.74
## treat=RT    112         88      79.5      0.904      1.74
##
## Chisq= 1.7  on 1 degrees of freedom, p= 0.2
```

Comparison of Survival Curves: Log-Rank Test

```
sakk.surv1 <- survfit(Surv(time, death) ~ treat, data = sakk,  
                      conf.type = "log-log")  
ggsurvplot(sakk.surv1, risk.table=TRUE, conf.int = TRUE, pval=TRUE)
```



The Hazard Rate

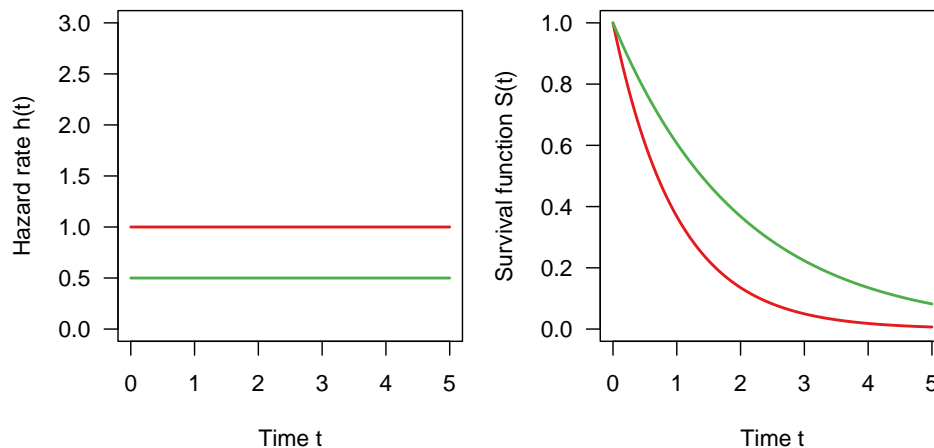
- We can also use the **mortality rate** to describe the risk of death in a certain interval:

$$\frac{\text{Death risk in interval}}{\text{Length of interval}}$$

- The mortality rate is conditional on having survived up to the interval of interest.
- The **hazard rate** $h(t)$ at time t is obtained by making the length of the interval very small.
- The hazard rate is also referred to hazard function, intensity rate and instantaneous death rate.

Hazard and Survival Function

Constant hazard rate



$$S(t) = \exp \left(- \int_0^t h(u) du \right)$$

Hazard Ratio

Definition and interpretation

Let $h_A(t)$ and $h_B(t)$ denote the hazard rates in groups A and B. The **proportional hazards assumption** implies that the **hazard ratio** is the same at all times t :

$$HR = h_A(t)/h_B(t).$$

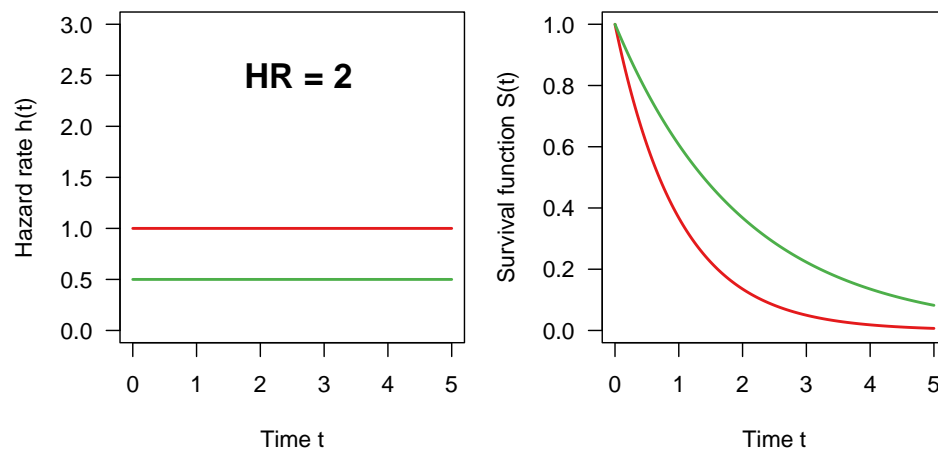
Interpretation of the hazard ratio:

- As (instantaneous) **relative risk** at any given time t .
- As **power transformation** to relate the two survival functions:
At any time t , the survival function $S_A(t)$ in group A is

$$S_A(t) = S_B(t)^{HR}.$$

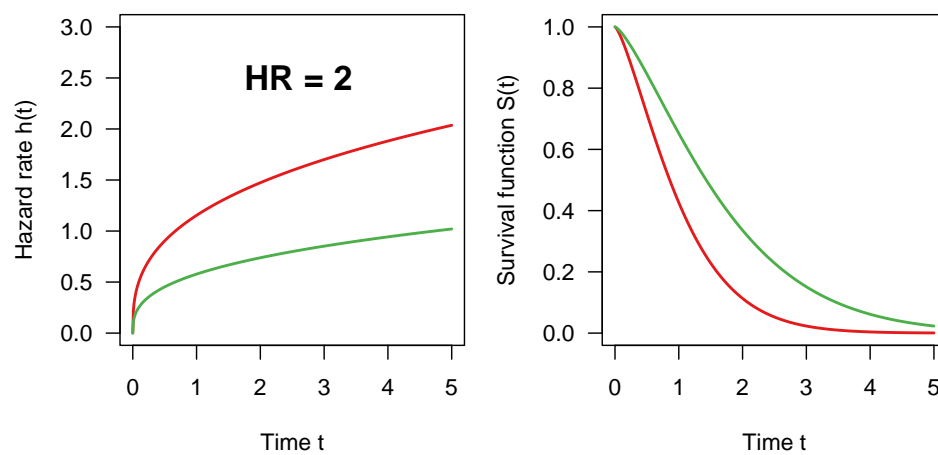
Hazard Ratio

Example 1



Hazard Ratio

Example 2



Calculation of Hazard Ratio from Log-Rank Test

- The hazard ratio can be calculated from the output of the log-rank test.
- Only the observed and expected number of cases are required.

```
lrTest <- survdiff(Surv(time, death) ~ treat, data = sak)
observed <- lrTest$obs
expected <- lrTest$exp
ratio <- observed/expected
HR <- ratio[2]/ratio[1]
SElogHR <- sqrt(sum(1/expected))

printWaldCI(log(HR), SElogHR, FUN = exp)

##      Effect 95% Confidence Interval P-value
## [1,] 1.225   from 0.904 to 1.660      0.19
```

The Cox Model

- Consider two individuals i and j with
 - hazard rates $h_i(t)$ and $h_j(t)$
 - and covariates \mathbf{x}_i and \mathbf{x}_j (treatment, gender, age etc.)
- The **Cox proportional hazards model** assumes that the hazard ratio does not depend on time:

$$\frac{h_i(t)}{h_j(t)} = \exp((\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\beta})$$

- A **likelihood approach** is used to estimate the **log hazard ratios** $\boldsymbol{\beta}$.
- Each term in the likelihood is the conditional probability that the event occurred in the actual case rather than in some other member of the **risk set**.

Estimation of Hazard Ratio

- To estimate the treatment effect we use **Cox-Regression**:

```
model.cox0 <- coxph(Surv(os, death) ~ trt, data = dat.tmp)
tableRegression(model.cox0)
```

	Hazard Ratio	95%-confidence interval	p-value
trt (RT)	1.23	from 0.90 to 1.66	0.19

- The comparison RT vs. RT+CT gives $HR = 1.23$ with relatively large confidence interval.
- no evidence for a treatment effect ($p = 0.19$).

Adjusting in Survival Analysis

- There may be differences between treatment groups in important baseline characteristics.
- Distribution of Gender in the two treatment groups:

Gender	Treatment	
	RT + CT	RT
F	11	23
M	101	89

- It is recommended to adjust the treatment effect for potentially unbalanced baseline values.
- Cox regression

Adjusting for Baseline Values with Cox-Regression

```
model.cox1 <- coxph(Surv(os, death) ~ trt + sex, data = dat.tmp)
tableRegression(model.cox1)
```

	Hazard Ratio	95%-confidence interval	<i>p</i> -value
trt (RT)	1.26	from 0.92 to 1.71	0.14
sex (F)	0.81	from 0.52 to 1.26	0.34

Interpretation:

- The adjusted death risk is increased by **26%** for RT relative to RT + CT ($p = 0.14$).
- Women have a **19%** reduced death risk ($p = 0.34$) compared to men.

Fall Semester 2024

University of Zurich

Leonhard Held

Lecture 6: Some Special Designs

Cluster-Randomized Trials

Equivalence and Non-Inferiority Trials

Cluster-Randomized Trials

- Up to now: Randomization of patients.
- **Cluster-randomized trials** allocate **groups of patients** to the same treatment
- Example: Intervention study in primary care
 - Does additional training of nurses and GPs in a general practice improve the care of patients with newly diagnosed type II *diabetes mellitus*?
 - 41 practices were randomized to the *status quo* or to receive additional training for their staff.

Analysis of Cluster-Randomized Trials

- Standard methods can no longer be used, as patient responses from the same cluster are **dependent**.
- The **intraclass correlation coefficient**, the correlation between outcomes within a cluster, needs to be taken into account.
- Methods of analysis:
 1. A simple approach is to construct a **summary measure** for each cluster and then analyse these summary values.
 2. To analyse data on patient level, a **mixed model**, a regression model with **cluster-specific random effects**, is needed. An alternative approach are so-called generalized estimating equations (GEEs).

Example

From Kelly and Bland (1998, BMJ)

- Investigation on the effect of guidelines for radiological referral on the referral practice of general practitioners (GPs)
- 17 practices in the intervention group received guidelines, 17 control practices were not sent anything.
- Outcome measure was the percentage of x-ray examinations requested that conformed to the guidelines.

Data

Data from five practices with and without intervention

```
head(CRT, 5)
```

##	Group	Practice	Total	Conforming	Percentage
## 1	Intervention	1	20	20	100.00
## 2	Intervention	2	7	7	100.00
## 3	Intervention	3	16	15	93.75
## 4	Intervention	4	31	28	90.32
## 5	Intervention	5	20	18	90.00

```
tail(CRT, 5)
```

##	Group	Practice	Total	Conforming	Percentage
## 30	Control	30	21	14	66.67
## 31	Control	31	126	83	65.87
## 32	Control	32	22	14	63.64
## 33	Control	33	34	21	61.76
## 34	Control	34	10	4	40.00

Summary Measure Analysis

Analysis on practice level

```
(mytTest <- t.test(Percentage ~ Group, var.equal=TRUE, data=CRT))

##
## Two Sample t-test
##
## data: Percentage by Group
## t = 1.8, df = 32, p-value = 0.07
## alternative hypothesis: true difference in means between group Intervention
## 95 percent confidence interval:
## -0.831 16.763
## sample estimates:
## mean in group Intervention      mean in group Control
##                81.53                73.56

(DifferenceInMeans <- mean(mytTest$conf.int))

## [1] 7.966
```

Summary Measure Analysis

Using the number of referrals as weight

```
result <- lm(Percentage ~ Group, data=CRT)
result.w <- lm(Percentage ~ Group, data=CRT, weight=Total)
tableRegression(result)
```

	Coefficient	95%-confidence interval	p-value
Intercept	73.56	from 67.34 to 79.78	< 0.0001
GroupIntervention	7.97	from -0.83 to 16.76	0.074

```
tableRegression(result.w)
```

	Coefficient	95%-confidence interval	p-value
Intercept	72.51	from 68.30 to 76.72	< 0.0001
GroupIntervention	6.98	from 0.14 to 13.82	0.046

Logistic Regression with Random Effects

Recommended analysis on patient level

Note: Binary patient level data can be aggregated to binomial counts within each practice.

```
library(lme4)
CRT$Outcome <- cbind(CRT$Conforming, CRT$Total-CRT$Conforming)
result.glmm <- glmer(Outcome ~ Group + (1|Practice), family=binomial, data=CRT)
(summary(result.glmm)$varcor)

## Groups      Name          Std.Dev.
## Practice (Intercept) 0.309

(ttable <- coef(summary(result.glmm)))

##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.019    0.1259   8.092 5.859e-16
## GroupIntervention  0.409    0.1964   2.082 3.731e-02

## treatment effect on patient level (odds ratio)
printWaldCI(ttable[2,1], ttable[2,2], FUN=exp)

##      Effect 95% Confidence Interval P-value
## [1,] 1.505  from 1.024 to 2.212      0.037
```

Ignoring the Clustering

Analysis on patient level, **not recommended**

- The following analysis ignores the cluster structure and acts as if patients had been randomly assigned to treatment groups.
- Aggregation to 2×2 table
- This results in confidence intervals which are too narrow and P values which are too small.

```
## aggregate data to 2x2 table ignoring cluster membership
Sums <- lapply(split(CRT, CRT$Group), function(x) c(Sums = colSums(x[,3:4])))
Total <- c(Sums$Intervention["Sums.Total"], Sums$Control["Sums.Total"])
Conf <- c(Sums$Intervention["Sums.Conforming"], Sums$Control["Sums.Conforming"])
notConf <- Total-Conf
Group <- factor(c("Intervention", "Control"), levels = c("Intervention", "Control"))
tab <- xtabs(cbind(Conf, notConf) ~ Group)
print(tab)

##
## Group      Conf notConf
## Intervention 341      88
## Control     509     193
```

Ignoring the Clustering

Analysis on patient level, **not recommended**

```
twoby2(tab)

## 2 by 2 table analysis:
## -----
## Outcome      : Conf
## Comparing    : Intervention vs. Control
##
##              Conf notConf      P(Conf) 95% conf. interval
## Intervention  341      88      0.7949   0.7540   0.8305
## Control       509     193      0.7251   0.6908   0.7568
##
##                                     95% conf. interval
##              Relative Risk: 1.0963   1.0260   1.1713
##              Sample Odds Ratio: 1.4693  1.1027   1.9577
##              Conditional MLE Odds Ratio: 1.4688  1.0935   1.9828
##              Probability difference: 0.0698  0.0182   0.1191
##
##              Exact P-value: 0.0087
##              Asymptotic P-value: 0.0086
## -----
```

Equivalence and Non-Inferiority Trials

- Up to now we discussed **superiority** studies.
- Aim of **equivalence trials** is not to detect a difference, but to establish equivalence of the two treatments.
- An equivalence trial needs pre-specification of an **interval of equivalence** $I = (-\delta, \delta)$ for the treatment difference.
- If δ is only specified in one direction, then we have a **non-inferiority trial**.
- Non-inferiority trials are based on one-sided hypothesis tests to check whether one group is **almost as good** (not much worse) than the other group.

Why Non-Inferiority and Equivalence Studies?

Examples

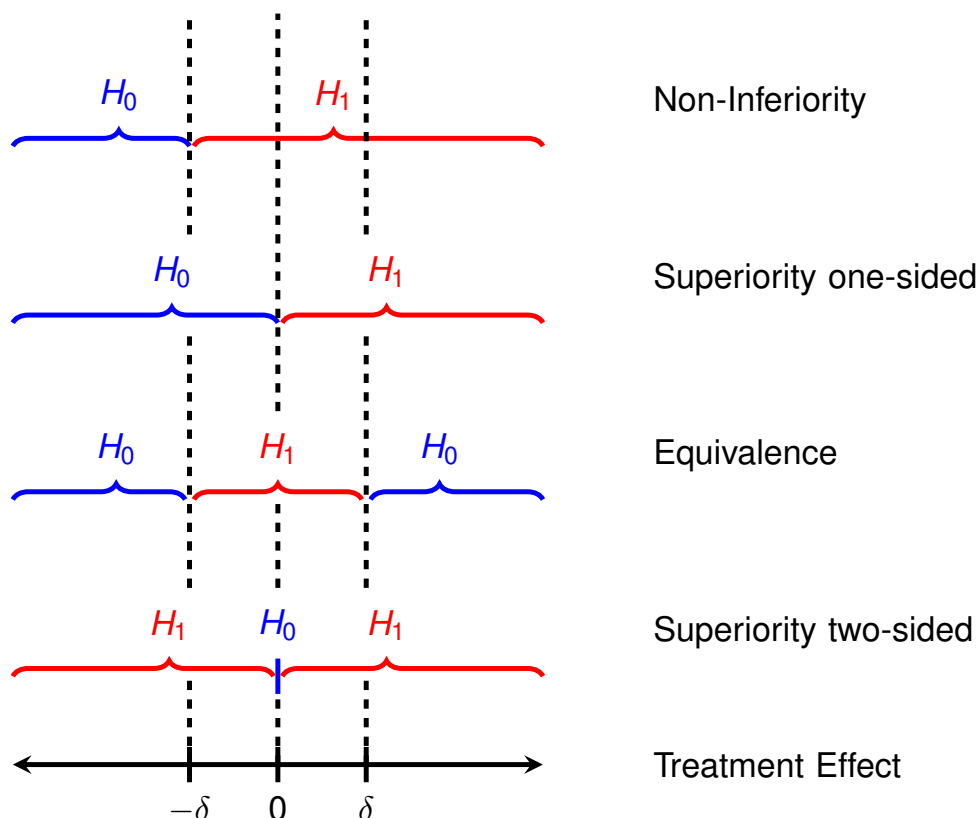
1. Intervention to be assessed:

- Similar to standard-of-care w.r.t. primary endpoint.
- **But:** may have **advantages in secondary endpoints:**
 - fewer side effects,
 - lower production costs,
 - more convenient formulation (tablet instead of infusion),
 - fewer doses,
 - better quality of life.

2. Bio-equivalence of drugs with identical active ingredient:

- different formulations of the same drug,
- **Generics.**

Comparison of H_0 and H_1



Assessing Equivalence

1. Compute a confidence interval at level γ for the difference in the treatment means
2. The treatments are considered equivalent if **both ends** of the confidence interval lie within the prespecified **interval of equivalence** $(-\delta, \delta)$.

If this does not occur, then equivalence has not been established.

The Type I error rate of this procedure is

$$\alpha \approx (1 - \gamma)/2.$$

For $\gamma = 90\%$ we have $\alpha \approx 0.05$.

For $\gamma = 95\%$ we have $\alpha \approx 0.025$.

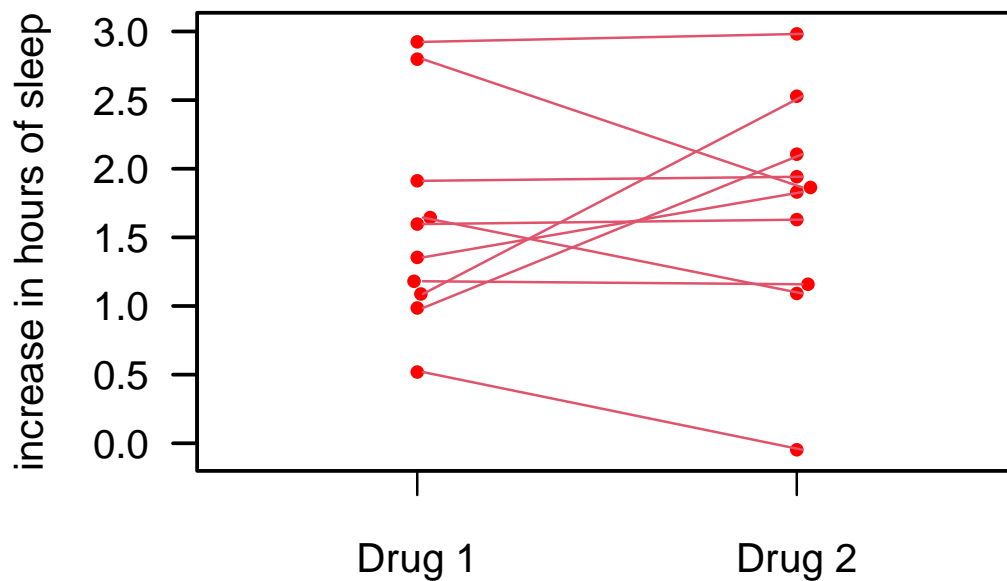
The TOST Procedure

An alternative approach to assess equivalence is as follows:

1. Apply **T**wo separate standard **O**ne-**S**ided significance **T**ests (**TOST**) at level α :
 - Test 1 for $H_0: \Delta < -\delta$ vs. $H_1: \Delta \geq -\delta$
 - Test 2 for $H_0: \Delta > \delta$ vs. $H_1: \Delta \leq \delta$.
2. If both one-sided tests can be rejected, we can conclude equivalence at level α .

Example: Comparison of Two Soporific Drugs

Increase in hours of sleep compared to control



Suppose goal is to show equivalence at margin of $\delta = 0.5h$.

Solution 1: Confidence Interval

```
extra1 <- sleep$extra[sleep$group==1]
extra2 <- sleep$extra[sleep$group==2]
res <- t.test(x=extra1, y=extra2, paired=TRUE, conf.level=0.9)
print(res$conf.int)

## [1] -0.5379  0.3216
## attr(,"conf.level")
## [1] 0.9
```

Limits of 90% CI do not lie within the interval of equivalence (-0.5h, 0.5h). Equivalence cannot be established.

Solution 2: TOST Procedure

```
tost1 <- t.test(x=extra1, y=extra2, paired=TRUE, mu=-0.5,
               alternative="greater", sig.level=0.05)
print(tost1$p.value)

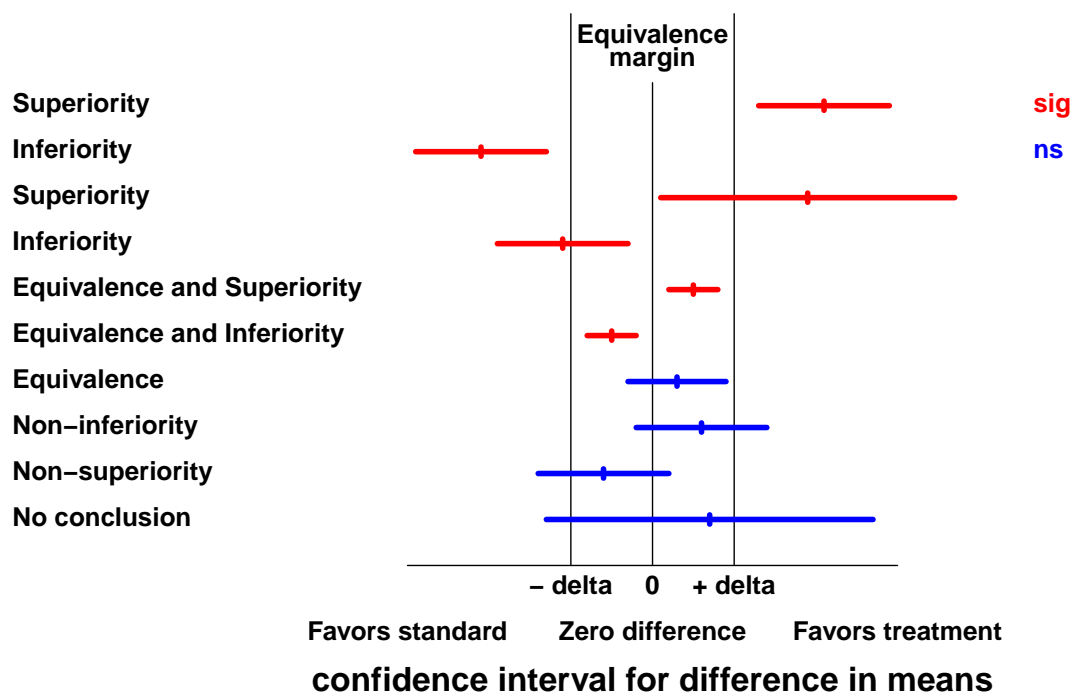
## [1] 0.06448

tost2 <- t.test(x=extra1, y=extra2, paired=TRUE, mu=0.5,
               alternative="less", sig.level=0.05)
print(tost2$p.value)

## [1] 0.01451
```

One p -value is larger than $\alpha = 5\%$, the other one is smaller.
Equivalence cannot be established.

Superiority, Equivalence and Non-Inferiority



Non-Inferiority Trials

- Useful if a proven active treatment exists and placebo-controls are not acceptable for ethical reasons.
- Technically: Just perform one of the two TOST tests, say

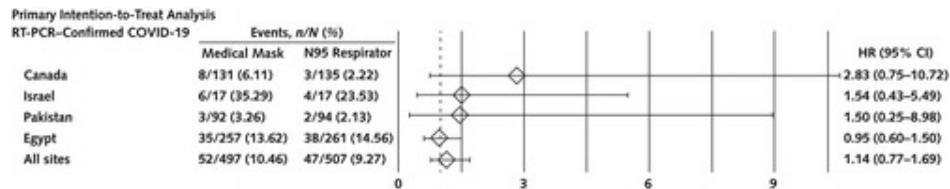
$$H_0 : \Delta < -\delta \text{ vs. } H_1 : \Delta \geq -\delta$$

- One-sided superiority trial corresponds to $\delta = 0$.
- Alternative procedure based on confidence intervals:
 - Compute confidence interval at level γ .
 - Reject H_0 of inferiority if upper bound is smaller than δ .
 - Type I error rate is $\alpha = (1 - \gamma)/2$.

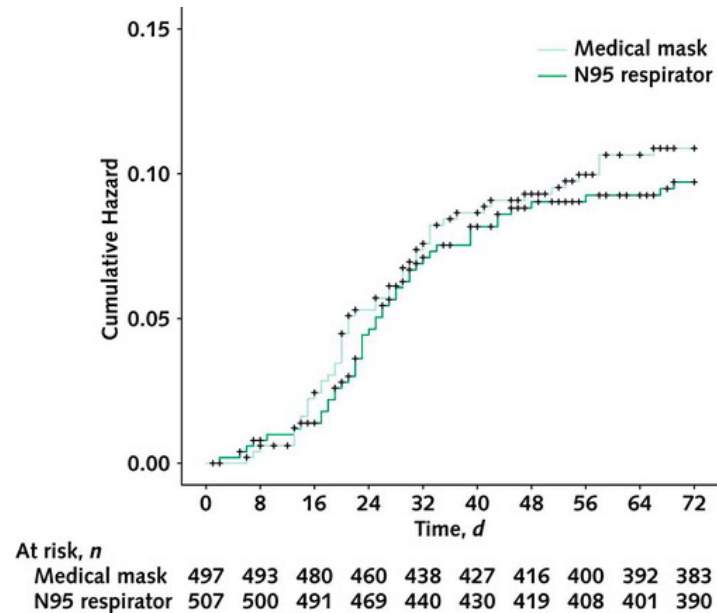
An interesting recent NI trial



Results



Non-inferiority margin $HR=2$



Conclusion

Conclusion:

Among health care workers who provided routine care to patients with COVID-19, the overall estimates rule out a doubling in hazard of RT-PCR-confirmed COVID-19 for medical masks when compared with HRs of RT-PCR-confirmed COVID-19 for N95 respirators. The subgroup results varied by country, and the overall estimates may not be applicable to individual countries because of treatment effect heterogeneity.