Main statistical methods in RCTs

Raphaël Porcher

Centre d'Epidémiologie Clinique, Hôtel-Dieu

Complex Systems Science Master

Outline

Lecture

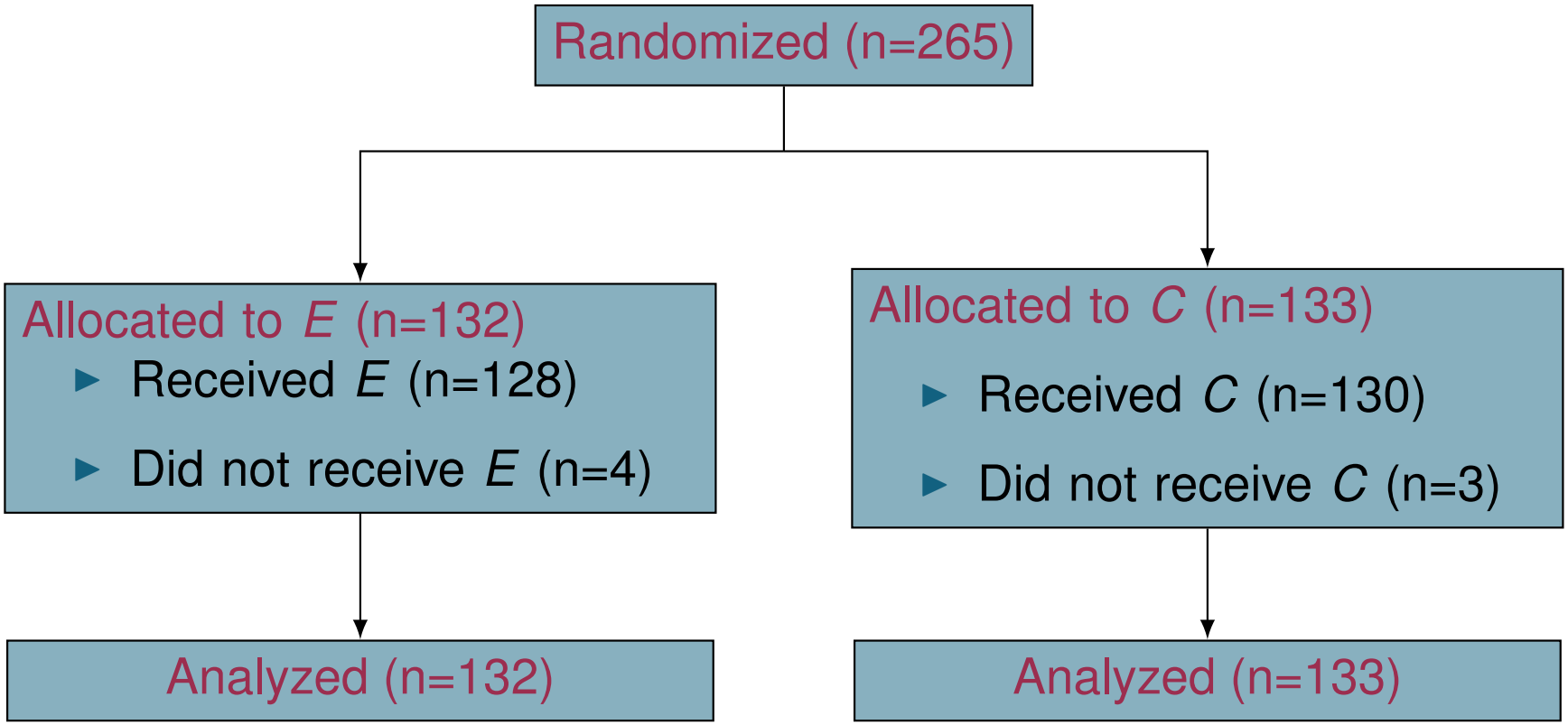
- Overview of usual statistical methods used in RCTs
- Each topic may deserve a lecture on its own (or even a full course)
- Make you aware of questioning we encounter, rather than explain everything in details
- Hopefully not a catalogue of tests

Statistical methods for RCTs

- Parallel two arm RCT: simple analysis at first sight
- Start from basics: two group comparison
- Analysis in practive: added levels of complexity encountered
- How to take them into account?

Outline

Two parallel arm RCT flow chart



Comparing two groups

- Two randomized (independent) groups of patients (E and C)
- One outcome measured on all patients (Y)
  - e.g. systolic blood pressure, PRO (scale, index, ...), success/failure (how defined?), occurrence of an event, ...
- Medical question: "Does the treatment have an effect of the outcome?"
- Statistical hypothesis: "No treatment effect on the probability distribution of the outcome"
- H_0: $\{f_E(y) = f_C(y)\}, \forall y \in I\!R$

Two-sample tests

- Depend on the type of outcome
- Continuous outcome
  - t-test or Welch's t-test: test $\{\mu_E = \mu_C\}$ assuming Y is Gaussian
  - Wilcoxon-Mann-Whitney test: non-parametric, rank-based test
- Binary outcome (H_0: $\{\pi_E = \pi_C\}$)
  - Pearson's χ^2 most frequent
  - Alternatives: Fisher's exact test, likelihood ratio test, some others
- Polytomous outcome
  - Pearson's χ^2 or Fisher's exact test
  - Not the best choice of an primary outcome anyway

Sketch of theory on statistical testing

- Medical researchers mix-up Fisher and Neyman–Pearson theories
- Suppose one observes $Y \sim f(y|\theta)$ and tests $H_0 : \theta = \theta_0$
- Fisher's significance testing
  - Choose $T = t(Y)$ so that large values of T reflect evidence against H_0
  - Compute $p = Pr_0(t(Y) \geq t(y))$ and reject H_0 if p is small
- Neyman–Pearson hypothesis testing
  - Also define an alternative hypothesis $H_1 : \theta = \theta_1$
  - Reject H_0 if $t(y) \geq c$ (predefined), and accept H_0 otherwise
  - With type I and II error rates $\alpha = Pr_0(t(Y) \geq c)$ and $\beta = Pr_1(t(Y) < c)$

Statistical testing in medical research

- Medical researchers use type I and II error rates
- α to detemine if the test is "significant" or not, i.e. if H_0 is rejected or not (usually $\alpha = 0.05$)
- β as the power $(1 - \beta)$ of study to detect an effect of a given size, or to choose in advance an appropriate sample size
- But they also consider the *p*-value as a strength of evidence against H_0
- Note that the cut-offs are arbitrary, and $p = 0.048$ or $p = 0.052$ should not indicate very different results
- Difference between "long-run" properties (NP) *vs* individual results (F)

Statistical testing with R

- *t*-test

```
t.test(outcome ~ tmt, var.equal = T)
```

- Welch *t*-test

```
t.test(outcome ~ tmt)
```

- Wilcoxon-Mann-Whitney test

```
wilcox.test(outcome ~ tmt)
```

- Pearson's χ^2 test

```
chisq.test(outcome,tmt,correct=F)
```

- Fisher's exact test

```
fisher.test(outcome,tmt)
```

Estimation of the treatment effect

- Statistical testing alone is not sufficient for inference on the treatment effect
- Fails to give any indication on its magnitude
- Point estimates and confidence intervals should be also provided
- Implies choosing a summary measure of the treatment effect

Treatment effect

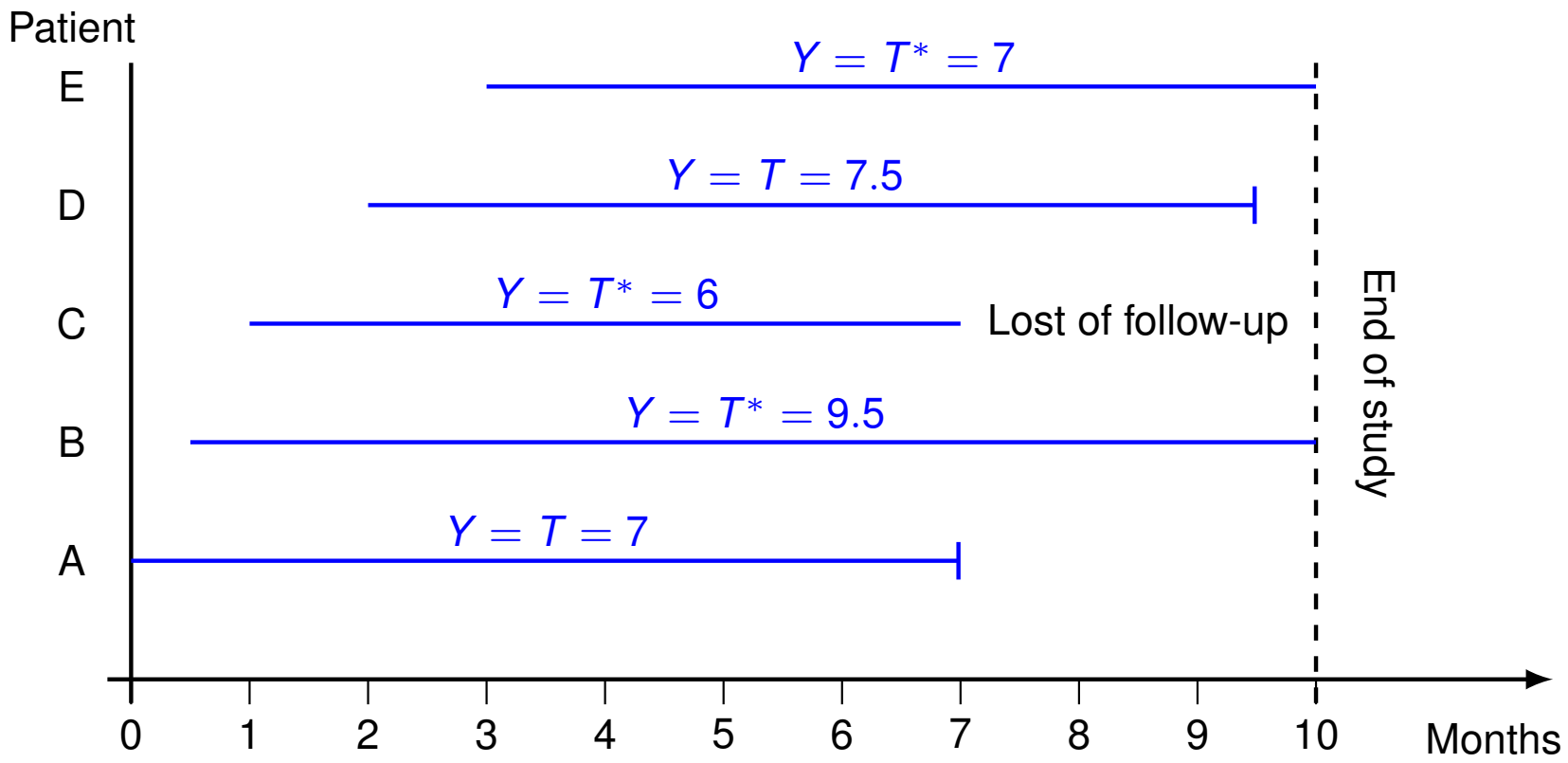
- Continuous outcome
  - $\delta = \mu_E - \mu_C$
  - Estimated by $d = \bar{y}_E - \bar{y}_C$
  - Standard error $s_d = \sqrt{\frac{s_E^2}{n_E} + \frac{s_C^2}{n_C}}$
  - $(1 - \alpha)$ confidence interval: $d \pm z_{\alpha/2} s_d$ by the CLT
- Binary outcome
  - Absolute effect $\delta = \pi_E - \pi_C$, inference as above by the CLT
  - Relative effect: $\theta = \frac{\pi_E}{\pi_C}$
  - Or the odds-ratio $\phi = \frac{\pi_E/(1-\pi_E)}{\pi_C/(1-\pi_C)}$

Outline

Survival data

- Aka time-to-event data
- Time T from an origin (randomization) to the occurrence of an event
- Event = (e.g.)
  - Death,
  - Cardiovascular event (death from cardiovascular cause or stroke or cardiac arrest)
  - Disease progression, improvement (of a given level), ...
- Specificity: time to event not observed for all patients (right censoring)

Example of a (small) trial



Censoring

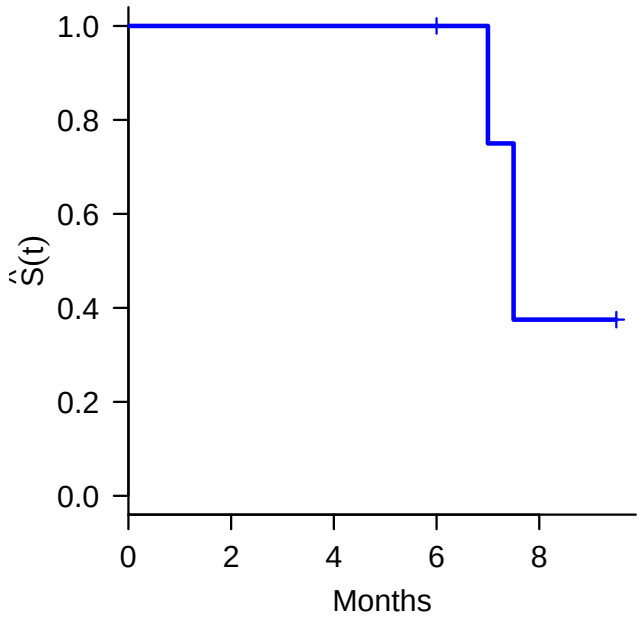
- Patients still alive (free of event) at the end of study
- Some patients lost to follow-up
- Mathematically, we observe $Y < T$ for censored patients
- Other censoring type: interval censoring ($Y_l < T < Y_u$)
- Issue: what inference for $f(t)$? How do we test H_0: $f_E(t) = f_C(t)$?

Kaplan–Meier estimator of the survivor function

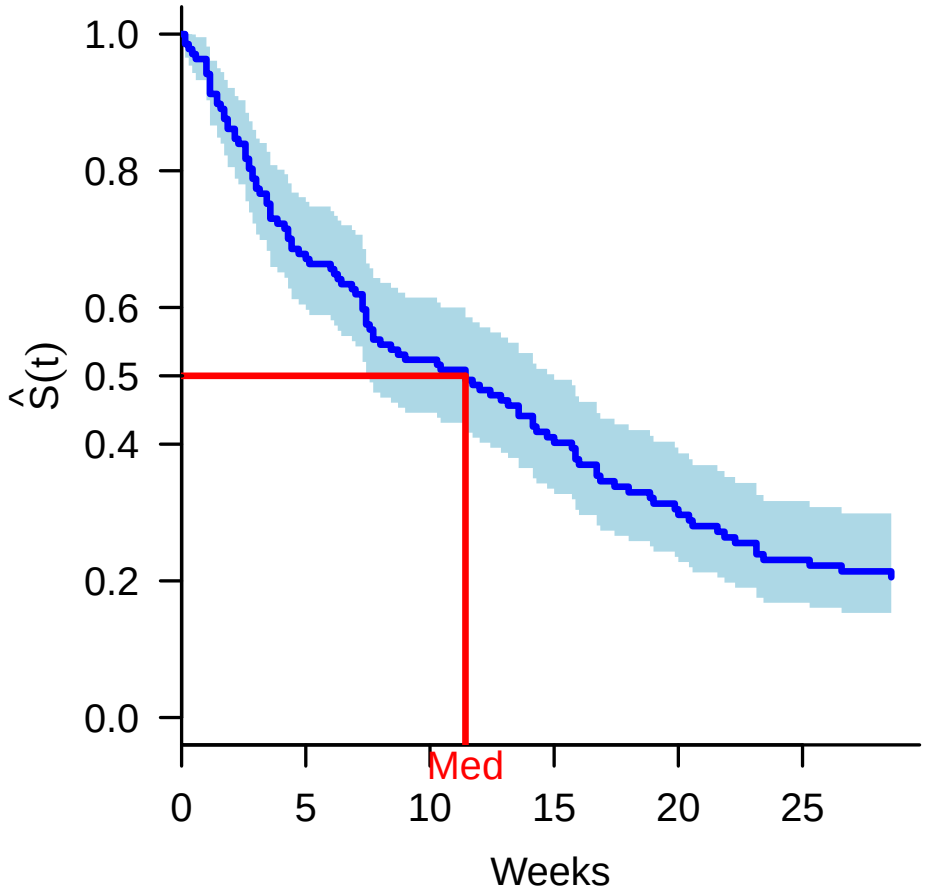
- $S(t) = 1 - F(t) = \Pr(T > t) = \int_t^{+\infty} f(u)du$
- $\hat{S}_{KM}(t) = \prod_{k:t_k \leq t} \frac{N_k - D_k}{N_k}$, where t_k are event times, N_k no. patients at risk of event just before t_k and D_k no. events at t_k
- Formulas for its variance (Greenwood) and confidence interval (Rothman)
- R syntax: `survfit(Surv(time,status) ~ 1, data=survive)` or `survfit(Surv(time,status) ~ levo, data=survive)`

Toy example

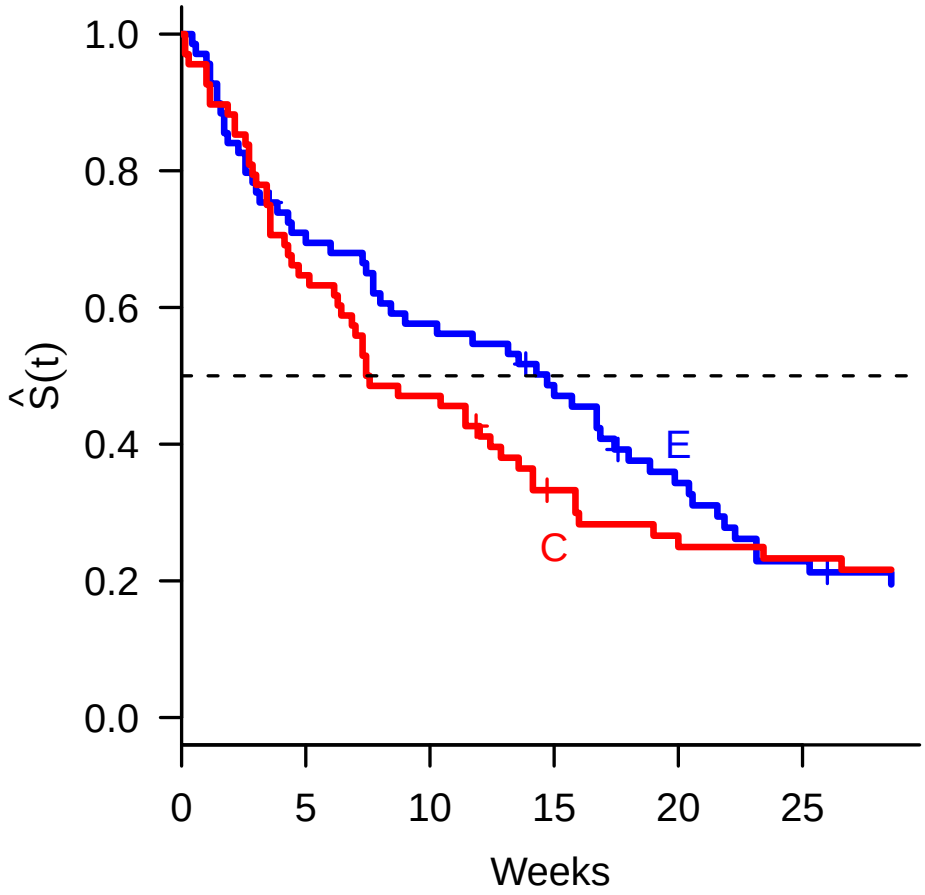
t_k	N_k	D_k	$\hat{S}(t_k)$
0	5	0	1
6	5	0	1
7	4	1	$1 \times 0.75 = 0.75$
7.5	2	1	$0.75 \times 0.5 = 0.375$
9.5	1	0	0.375



Example: lung cancer trial



Example: lung cancer trial (cont'd)



Hypothesis testing

- Log-rank test of $H_0 : f_E(t) = f_C(t) \Leftrightarrow H_0 : S_E(t) = S_C(t)$
- Member of a more general family of tests (G_ρ family of Harrington and Fleming)
- Detailed formulation not given here (unless you want it)
- Can be seen as a generalization of Mantel–Haenszel (stratified) χ^2 test
- R syntax: `survdiff(Surv(time,status) ~ levo, data=survive)`

Inference for treatment effect

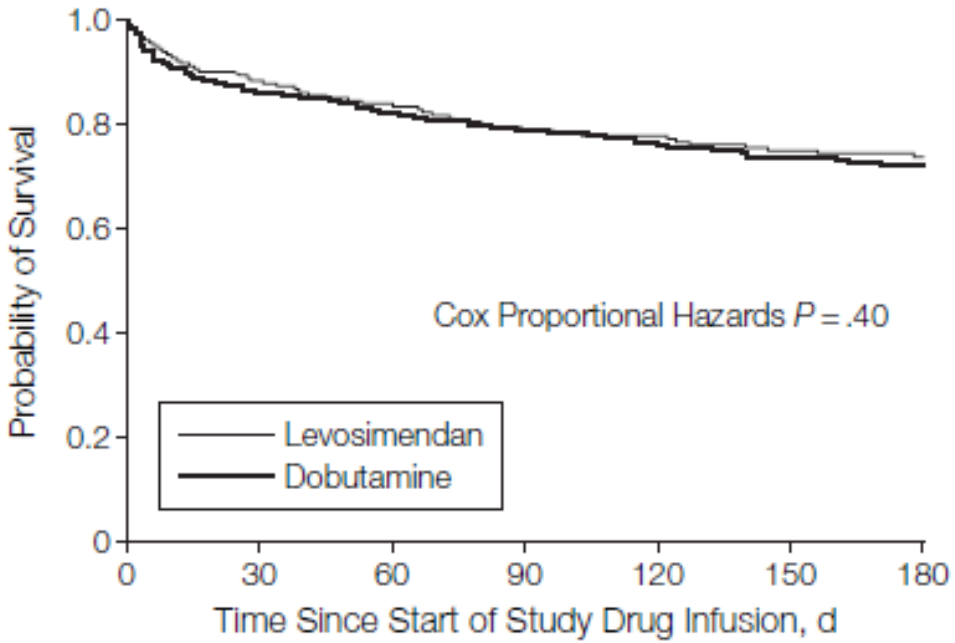
- For censored data, the mean is difficult to estimate in general
- Median survival usually given (with confidence interval or IQR — very different)
- Difficult to interpret in terms of treatment effect (difference of medians is not a median of differences!)
- Often hazard ratios estimated (HR)
- Hazard $\lambda(t) = f(t)/S(t) = -\frac{d\log(S(t))}{dt} = \lim_{dt\to 0} \Pr\left(t \leq T < t + dt | T \geq t\right)/dt$
- HR: $\theta = \lambda_E(t)/\lambda_C(t)$, assumed constant (PH)

Example: SURVIVE study

- Double-blind RCT comparing levosimendan *vs* dobutamine in patients with acute decompensated heart failure
- 1327 patients randomized (664 in the levosimendan arm — E, 663 in the dobutamine arm — C)
- Primary endpoint: all-cause mortality during the 180 days following randomization (i.e. survival)

SURVIVE: survival curves

Figure 2. Effect of Dobutamine and Levosimendan Treatment on All-Cause Mortality During 180 Days Following the Start of Study Drug Infusion



No. at Risk						
Levosimendan	664	608	586	525		462
Dobutamine	663	596	568	519		454

SURVIVE: treatment effect estimate(s)

Table 2. Primary, Secondary, and Post Hoc All-Cause Mortality End Points*

	No. (%) of Patients†			
	Levosimendan (n = 664)	Dobutamine (n = 663)	HR (95% CI)	*P* Value
Primary end point				
All-cause mortality at 180 d	173 (26)	185 (28)	0.91 (0.74-1.13)	.40‡
Secondary end point				
All-cause mortality at 31 d	79 (12)	91 (14)	0.85 (0.63-1.15)	.29‡
Mean change in BNP at 24 h from baseline, pg/mL	(n = 628) −631	(n = 611) −397		<.001§
Mean No. of days alive and out of the hospital during 180 d	120.2	116.6		.30‖
Dyspnea assessed at 24 h; ≥mild improvement¶	544 (82)	550 (83)		.96‖
Global assessment at 24 h; ≥mild improvement¶	531 (80)	537 (81)		>.99
Cardiovascular mortality during 180 d	157 (24)	171 (26)	0.90 (0.72-1.12)	.33‡
Post hoc all-cause mortality				
5 d	29 (4)	40 (6)	0.72 (0.44-1.16)	.17‡
14 d	59 (9)	69 (10)	0.84 (0.59-1.19)	.33‡
90 d	139 (21)	138 (21)	0.99 (0.78-1.25)	.91‡

Abbreviations: BNP, B-type natriuretic peptide; CI, confidence interval; HR, hazard ratio.
*Survival differences were tested for significance by the Cox proportional hazard regression model with treatment as the only covariate. Comparison of categorical variables such as dyspnea assessment, patients' global assessment, and days alive and out of the hospital were performed by the Cochran-Mantel-Haenszel test with effect for treatment only. Changes in BNP levels were analyzed using the Kruskal-Wallis test.
†Unless otherwise indicated.
‡Cox proportional hazards model was used for treatment effect only.
§Analysis of covariance model used with baseline value as covariate and treatment for main effect.
‖Cochran-Mantel-Haenszel mean score test with effect for treatment only.
¶Distribution from markedly improved to markedly worse.

Outline

Definition

- Still the two parallel arm RCT
- Objective: to study the effect of a treatment (E) relative to C in a certain type of patients
- Conclusions should apply to any patient verifying eligibility criteria
- Outcomes can however be affected by other factors than treatment (prognostic factors)
- The treatment effect itself may be affected by some patients' characteristics

Why take it into account?

- Accounting for heterogeneity may be valuable
- To determine if the trial's result equally apply to all patients' types recruited
- To obtain an unbiased or more precise estimate of the treatment effect

Sources of heterogeneity

- Differing treatment effect in various subgroups (e.g. age, previous medical condition, gene polymorphism, . . .)
- Even when the (true) treatment effect is constant over subgroups, imbalance on a major prognostic factor may bias the treatment effect estimate
- Centre effect: due to local recruitment, practices, SoC, . . .
- Accounting for heterogeneity
  - Adjusted analyses
  - Estimation and test of heterogeneity

No heterogeneity (interaction) on the treatment effect

- When the treatment effect is the same for all patients
- But randomization groups are unbalanced on one (or more) prognostic factor
- May bias the treatment effect estimate
- Corrected by adjusted analyses (multiple regression)
- If groups are balanced, adjustment may improve the precision of the treatment effect estimate

What should we adjust for?

- From a theoretical point-of-view, randomization guarantees the absence of bias (i.e. on average over repeated trials)
- But substantial imbalance may arise by chance
- Adjustment may be recommended when
  - Imbalance on important prognostic factors has occurred
  - A factor has a huge effect on the outcome ($\forall$ imbalance)
  - Demonstration is needed that the observed effect is not artificially caused by a prognostic factor
  - We want to illustrate or quantify the effect of known factors
- Objective: to obtain an estimate of the treatment effect independent from prognostic factors

Regression model

- Mathematical model of the relationship between the distribution of the outcome and the prognostic factors (covariates)
- General model: $g(\mathrm{E}(Y|X = x)) = h(x)$
- More simple model: $Y = h(x) + \epsilon$, where ϵ is a random variable with mean 0, which represents the residual error of the model
- Model (and in particular g) depending on the outcome

Usual regression models

- Continuous outcome (Gaussian) :
  linear model

$$y_i = \mu + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$$

  or AN(C)OVA

$$y_i = \mu + \alpha_{k(i)} + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$$

- Binary endpoint: logistic model

$$\ln\left(\frac{\pi}{1-\pi}\right) = \mu + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Censored outcome: Cox proportional hazards model

$$\lambda(t; x_1, \ldots, x_p) = \lambda_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)$$

R syntax

- Linear model

```
lm(bnp ~ levo + bnp0 + chf + gender, data=survive)
```

- Logistic model

```
glm(dysp ~ levo + gender, family="binomial", data=survive)
```

- Cox model

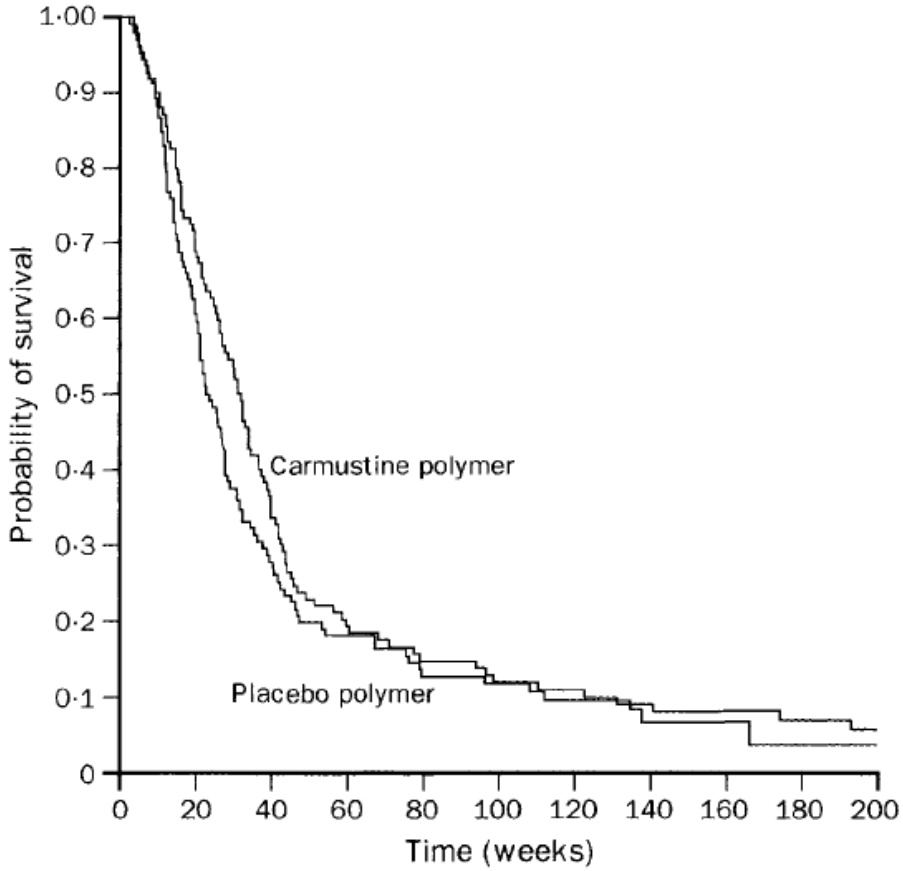
```
coxph(Surv(ftime,fs) ~ bcnu + ps>=70 + local + age,
                                    data=BCNU)
```

Example: BCNU trial[1]

- Double-blind RCT comparing carmustine-impregnated (BCNU) *vs* placebo biodegradable polymers in patients with recurrent malignant glioma
- 222 patients randomized (110 in the BCNU arm, 112 in the placebo arm)
- Primary endpoint: survival from polymer implantation

[1] Brem et al. *Lancet* 1995;345(8956):1008–12.

Example: survival results



Example: effect of adjustment

Variable	HR	95%CI	*p*
BCNU *vs* placebo	0.83	0.63–1.10	0.19
BCNU *vs* placebo	0.69	0.52–0.91	0.01
Karnofsky >70 *vs* ≤ 70	0.66	0.49–0.91	0.01
Local radiation	0.59	0.42–0.83	0.003
"Active" *vs* "quiescent"	1.93	1.26–3.78	0.02
Previous nitrosoureas	1.53	1.13–2.08	0.006
White *vs* other race	1.75	1.03–2.99	0.04
$>75\%$ resection *vs* $\leq 75\%$	0.67	0.49–0.93	0.02
Age (/decade)	1.25	1.11–1.40	<0.001

- Prognostic factors were balanced between groups
- But several were strong predictors of outcome
- Drug approved by the FDA on the basis of this trial (and some other data)

Heterogeneity of the treatment effect

- Different treatment effect according to patients subgroup
- From a statistical point-of-view: interaction between the treatment and subgroup indicator variables
- Several methods to test an interaction
  - Interaction effect in a regression model
  - Compare the treatment effect estimate obtained in the different subgroups

Regression model with interactions

- Just the same, with a variable indicating the interaction
- E.g. logistic model: $\log\left(\frac{\pi}{1-\pi}\right) = \mu + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2$ with $x_1 = 0$ if C et $x_1 = 1$ if E
- R syntax:

```
glm(dysp ~ levo + gender + levo:gender, family="binomial",
data=survive)
```

or `glm(dysp ~ levo*gender, family="binomial", data=survive)`

- Be careful with the coding of treatment effect (e.g. dose effect or age)

Interaction test in the model

- As for main effects
- $H_0 : \gamma = 0$
- ANOVA: Fisher's F test ($\neq$ Fisher's exact test)
- Other model: Wald test, score test or likelihood ratio test
- Wald test: $\left(\frac{\hat{\gamma}}{se(\hat{\gamma})} \right)^2 \sim \chi^2_1$ under H_0

Interaction test using subgroup results

- RCT comparing two combinations(PFT and PF) for adjuvant therapy of resecable primitive breast cancer (Fisher et al. *J Clin Oncol* 1983)
- The response to PFT may depend on age and progesterone receptor

	Age <50 **PR** <10		**Age** ≥ 50 **PR** <10		**Age** <50 **PR** ≥ 10		**Age** ≥ 50 **PR** ≥ 10	
	PF	**PFT**	**PF**	**PFT**	**PF**	**PFT**	**PF**	**PFT**
3y DFS	0.599	0.436	0.526	0.639	0.651	0.698	0.639	0.790
se	0.0542	0.0572	0.0510	0.0463	0.0431	0.0438	0.0386	0.0387
d_k	0.163		-0.114		-0.047		-0.151	
s_k	0.0788		0.0689		0.0614		0.0547	

Interaction tests

- "Quantitative" interaction (roughly the same as previously)
  - K subgroups $\rightarrow H_0 : \delta_k = \delta, \forall k = 1, \ldots, K$
  - Obtain d_k in each subgroup, with stadard error s_k
  - Test statistic: $Q = \sum \frac{(D_k - \bar{d})^2}{s_k^2} \sim \chi^2_{K-1}$ under H_0
  - Where $\bar{d} = \sum \frac{d_k / s_k^2}{1 / s_k^2}$
- Qualitative interaction[2]
  - The direction of the effect varies betwen subgroups
  - $H_0 : \delta_k \geq 0, \forall k = 1, \ldots, K$ or $\delta_k \leq 0, \forall k = 1, \ldots, K$
  - Test statistic: $K = \min(Q^+, Q^-)$ to be compared to a critical value
  - $Q^- = \sum (D_k^2 / s_k^2) 1_{[D_k > 0]}$ and $Q^+ = \sum (D_k^2 / s_k^2) 1_{[D_k < 0]}$

[2] Gail, Simon. *Biometrics* 1985;41(2):361–72

Preceding example

- $\bar{d} = -0.062$
- $Q = 11.43$, df=3, $p = 0.0096$
- $Q^+ = 10.94$, $Q^- = 4.28 \rightarrow K = 4.28$
- $c_{\alpha=0.05,K=4} = 5.43 > K$

Subgroup analyses

- Companion to the heterogeneity tests
- Objective: to find the the treatment effect may be different according to subgroup
- Scientifically sound (at least may be)
- Statistically problematic

Issues with subgroup analyses

- Validity and rationale of the definition (not further developed here)
- Multiplicity (of tests)
- Power (to detect meaningful differences)

Multiplicity

- Problems caused by multiple statistical testing
- In the absence of heerogeneity, if tests with $\alpha = 0.05$ are repeated, we increase the probability to find at least one "significant" test by chance:

No. tests	Probability
1	0.05
2	0.10
3	0.14
5	0.23
10	0.40
20	0.64

How to handle multiplicity?

- Perform a single test ;-)
- Correct p-values for multiplicity
- Bonferoni correction: $p' = \min(p \times k, 1)$ if k tests
- Better corrections: Holm, Hochberg, ...
- R syntax: `p.adjust(p, method="holm")`

Power of interaction tests

- The trials are generally powered for main test of the primary hypothesis with adequate power
- Interaction tests and tests within subgroup have (much) less power
- All the more when a correction for multiple testing is applied!

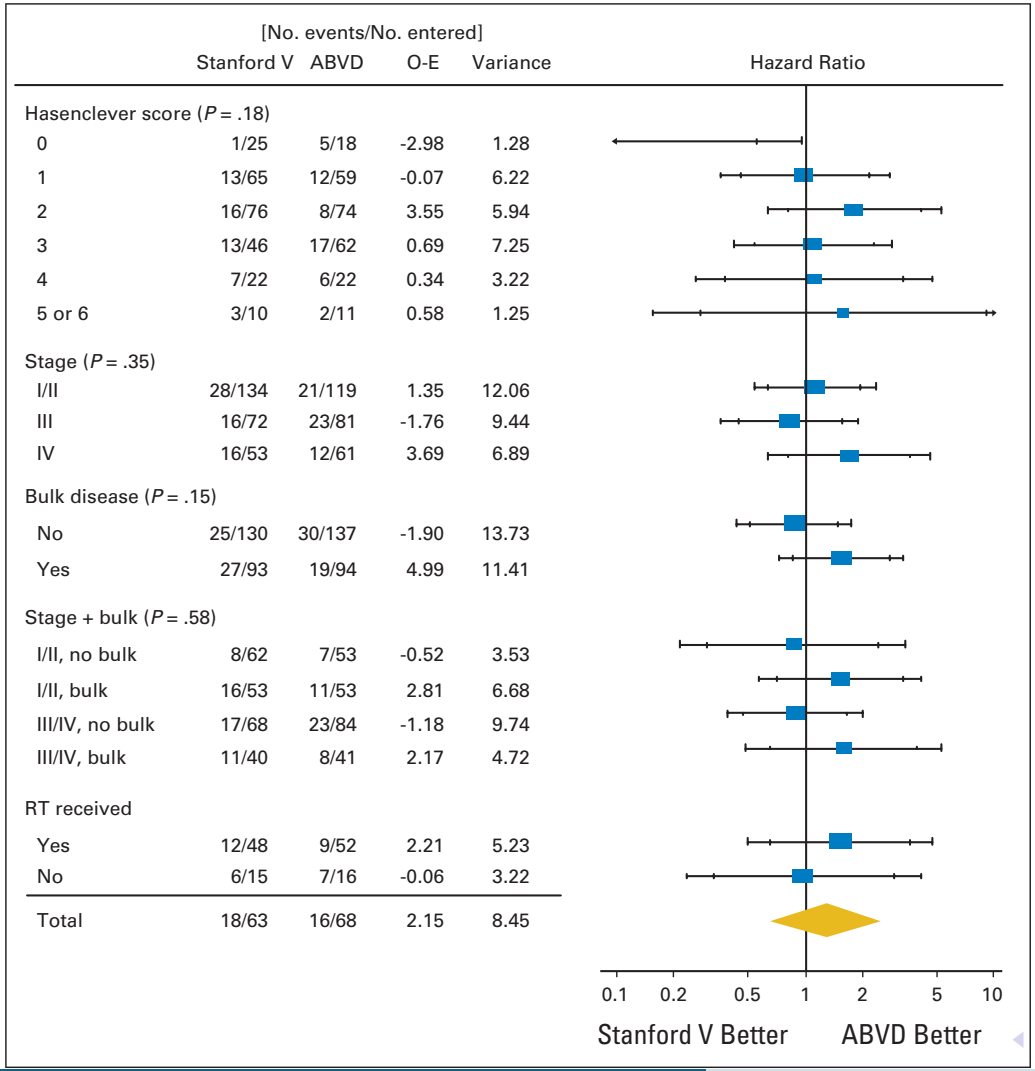
Example: SURVIVE trial

Table 4. Influence of Prespecified Baseline Characteristics on Survival at 180 Days*

	Total No. of Patients	Mortality Rate, No./Total (%)		HR (95%CI)	*P* Value for Interaction
		Levosimendan	Dobutamine		
Sex					
Female	371	45/171 (26)	58/200 (29)	0.89 (0.60-1.32)	.86
Male	956	128/493 (26)	127/463 (27)	0.93 (0.73-1.19)	
Age, y					
<65	501	54/237 (23)	56/264 (21)	1.08 (0.74-1.57)	.25
≥65	826	119/427 (28)	129/399 (32)	0.83 (0.65-1.06)	
History of CHF					
Yes	1171	148/586 (25)	165/585 (28)	0.87 (0.70-1.09)	.19
No	156	25/78 (32)	20/78 (26)	1.31 (0.73-2.37)	
Prior use of β-blocker					
Yes	669	65/336 (19)	72/333 (22)	0.87 (0.62-1.22)	.69
No	658	108/328 (33)	113/330 (34)	0.95 (0.73-1.24)	
Prior use of ACE inhibitor or ARB					
Yes	914	104/463 (23)	98/451 (22)	1.04 (0.79-1.38)	.15
No	413	69/201 (34)	87/212 (41)	0.77 (0.56-1.06)	
AMI as primary cause of hospitalization					
Yes	178	27/83 (33)	40/95 (42)	0.72 (0.44-1.17)	.23
No	1149	146/581 (25)	145/568 (26)	0.98 (0.78-1.23)	
Serum creatinine level, mg/dL (μmol/L)					
≤2.5 (≤221)	1237	149/620 (24)	164/617 (27)	0.88 (0.71-1.10)	.48
>2.5 (>221)	87	24/44 (55)	21/43 (49)	1.10 (0.61-1.99)	
Oliguria					
Yes	101	20/49 (41)	27/52 (52)	0.69 (0.39-1.24)	.26
No	1226	153/615 (25)	158/611 (26)	0.95 (0.76-1.19)	
Dyspnea at rest					
Yes	1184	151/595 (25)	161/589 (27)	0.92 (0.73-1.14)	.80
No	133	21/64 (33)	22/69 (32)	0.99 (0.54-1.80)	
Mechanical ventilation for heart failure					
Yes	24	6/12 (50)	5/12 (42)	1.05 (0.32-3.49)	.75
No	1302	166/651 (26)	180/651 (28)	0.90 (0.73-1.12)	
Heart rate, beats/min					
<83	705	78/339 (23)	88/366 (24)	0.93 (0.69-1.26)	.80
≥83	622	95/325 (29)	97/297 (33)	0.88 (0.66-1.17)	
Systolic blood pressure, mm Hg					
<100	271	53/135 (39)	62/136 (46)	0.80 (0.55-1.15)	.39
≥100	1051	129/527 (23)	121/524 (23)	0.97 (0.76-1.25)	

Abbreviations: ACE, angiotensin-converting enzyme; AMI, acute myocardial infarction; ARB, angiotensin II receptor blocker; CHF, congestive heart failure; CI, confidence interval; HR, hazard ratio.
*The Cox model was used to examine potential treatment × subgroup interactions using treatment × subgroup interaction as covariates.

Example: another trial (lymphoma)



	[No. events/No. entered] Stanford V	ABVD	O-E	Variance
Hasenclever score ($P = .18$)				
0	1/25	5/18	-2.98	1.28
1	13/65	12/59	-0.07	6.22
2	16/76	8/74	3.55	5.94
3	13/46	17/62	0.69	7.25
4	7/22	6/22	0.34	3.22
5 or 6	3/10	2/11	0.58	1.25
Stage ($P = .35$)				
I/II	28/134	21/119	1.35	12.06
III	16/72	23/81	-1.76	9.44
IV	16/53	12/61	3.69	6.89
Bulk disease ($P = .15$)				
No	25/130	30/137	-1.90	13.73
Yes	27/93	19/94	4.99	11.41
Stage + bulk ($P = .58$)				
I/II, no bulk	8/62	7/53	-0.52	3.53
I/II, bulk	16/53	11/53	2.81	6.68
III/IV, no bulk	17/68	23/84	-1.18	9.74
III/IV, bulk	11/40	8/41	2.17	4.72
RT received				
Yes	12/48	9/52	2.21	5.23
No	6/15	7/16	-0.06	3.22
Total	18/63	16/68	2.15	8.45

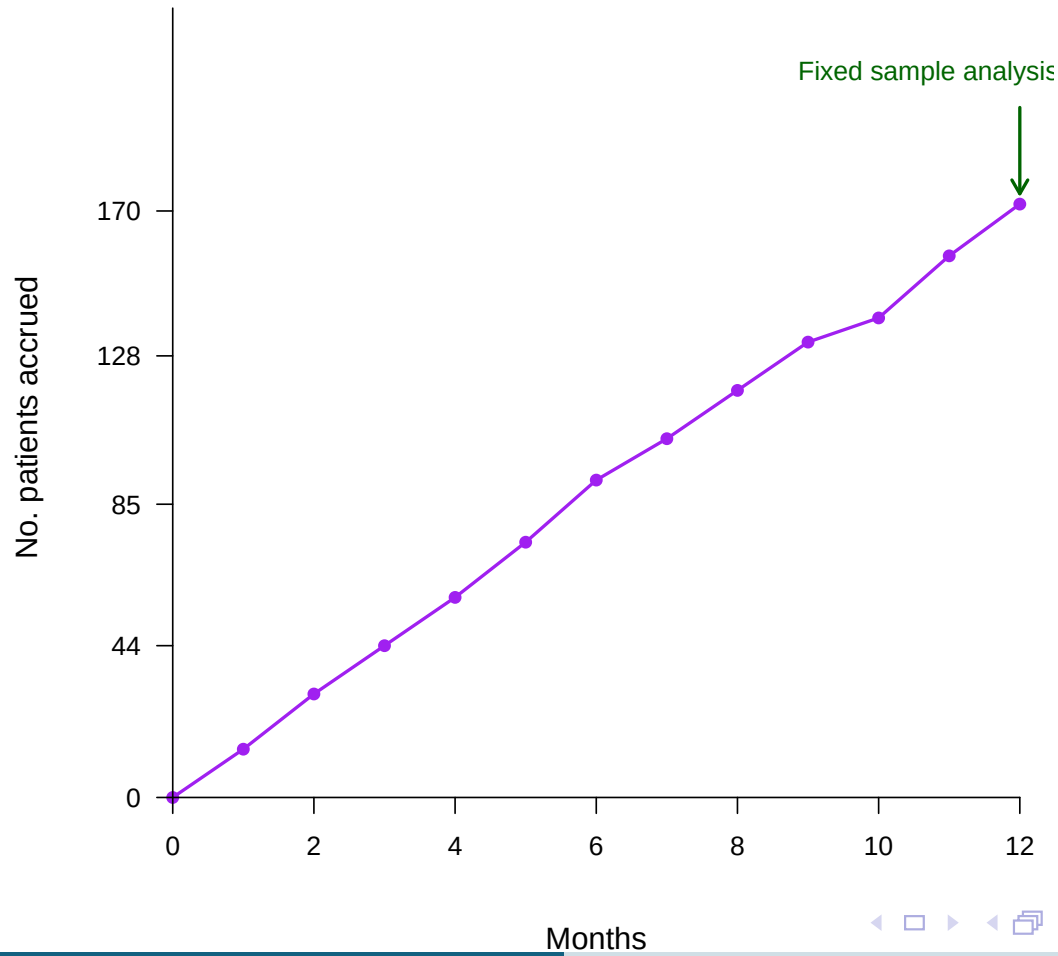
Fig 4. Subgroup analyses on progression-free survival. Hasenclever score ($P = .18$), stage ($P = .35$), bulk disease ($P = .15$), stage plus bulk disease ($P = .58$), and radiotherapy (RT) received (phase II data only). O, observed; E, expected; ABVD, doxorubicin, bleomycin, vinblastine, and dacarbazine.

Outline

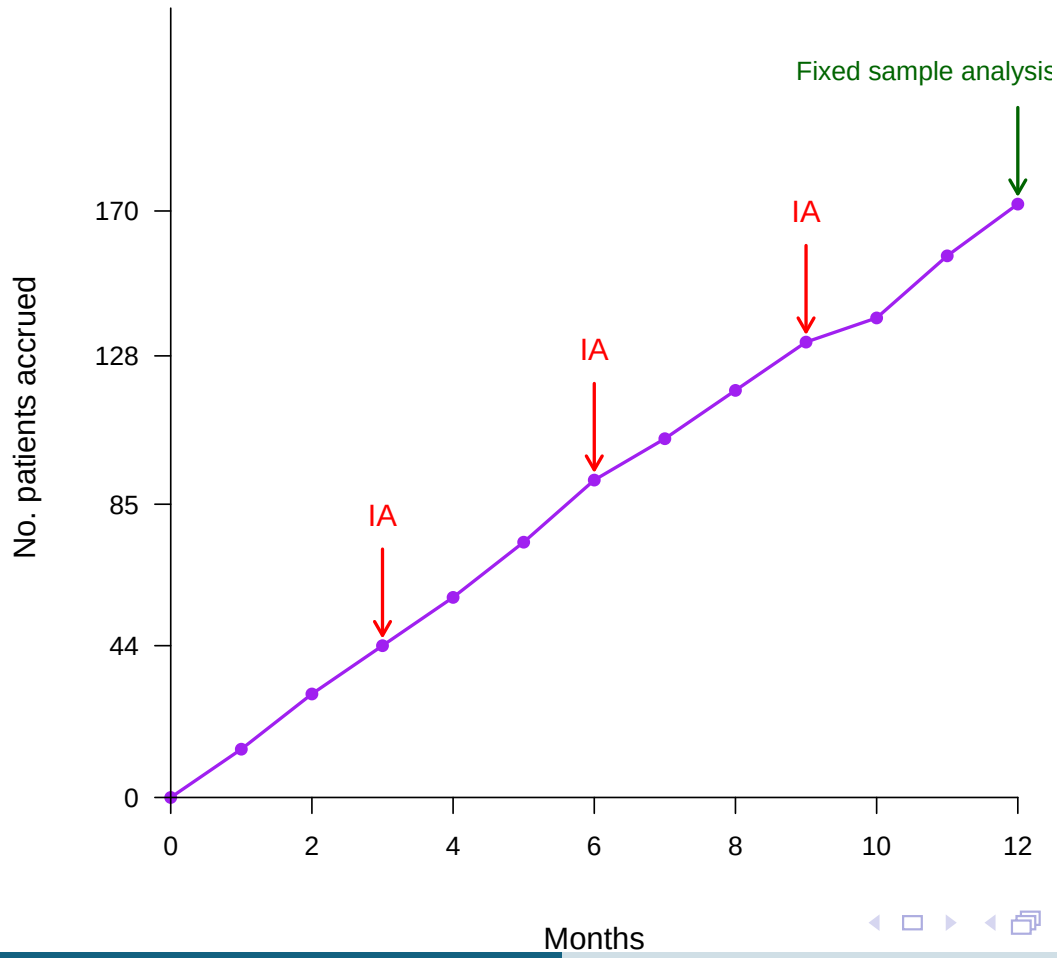
Accounting for accumulated information during the trial

- Classical design for RCT
  - Plan a trial (α, β, n, ...)
  - Accrue prespecified no. of patients
  - Analyze the results at the end of follow-up
- Has the advantage of simplicity (really?)
- Does not account for accumulated data $\rightarrow$ but relevant information!
- Solution: sequential (repeated) analyses

Interim analyses

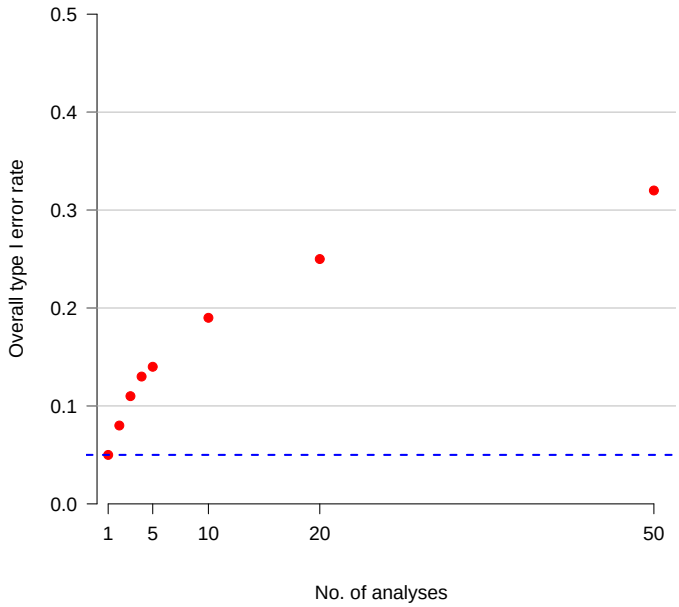


Interim analyses



Repeated statistical testing

- Naive repetition of interim analyses compromises the overall type I error rate



$\rightarrow$ Methods have been developed for a proper control of α

Why a group sequential design?

- Ethical reasons
  - Experiment in humans!
  - Patients should not be exposed to inferior or toxic treatments
  - Equipoise at the beginning of the trial, but after 75% accrual?
- Economic reasons
  - Early stopping = less expensive
  - For efficacy: treatment on the market sooner
  - For futility: resources saved
- Scientific

Pro and cons early stopping

Stop the trial

- Minimizing the number of subjects
- Minimizing the number of subjects in the inferior arm
- Minimizing the costs
- Publish rapidly, market rapidly, . . .

Carry on with the trial

- Better estimate of treatment effect (precision)
- Reduce the risk of error +++
- Increase power
- Allow for subgroup analyses
- More data on secondary outcomes

Group sequential tests

- Two seminal works (Pocock 1977, O'Brien & Fleming 1979)
- Other methods (Wand & Tsiatis, α-spending approach, ...) not developed here
- Test statistics Z_k computed at each analysis k ($\leq K$)
  - if $|Z_k| \geq c_k$, stop and reject H_0
  - si $|Z_k| < c_k$, $k < K$, continue
  - si $|Z_K| < c_K$, stop and accept H_0
- Decision boundaries $\{c_1, \ldots, c_K\}$ chosen to control α
  $\Pr_{H_0}(\text{reject at analysis } k = 1, k = 2, \ldots \text{or } k = K) = \alpha$
- Sample size modifed as compared to fixed design for power $1 - \beta$

Pocock *vs* OBF

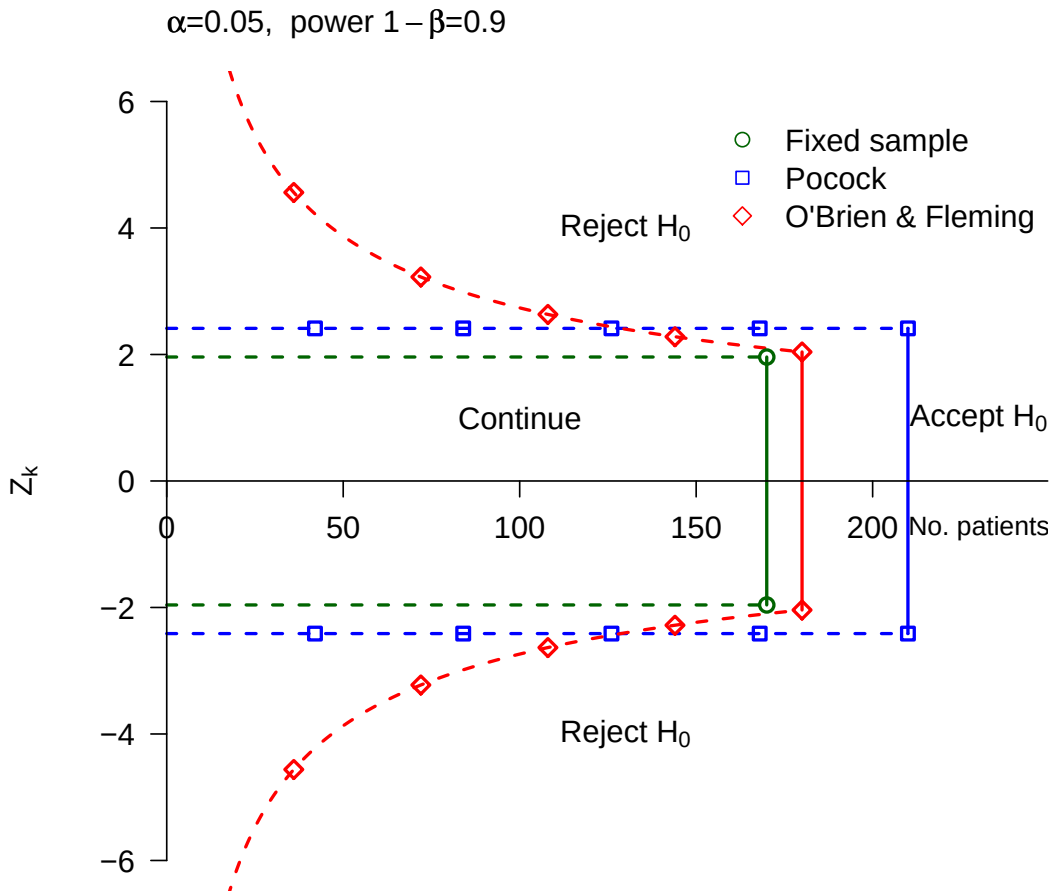
Pocock

- $c_k = C_P(K, \alpha)$ fixed $\forall k$
- Equivalent to perform repeated tests at a level $\alpha' = 2[1 - \Phi\{C_P(K, \alpha)\}]$
- Sample size multiplier $R_P(K, \alpha, \beta)$

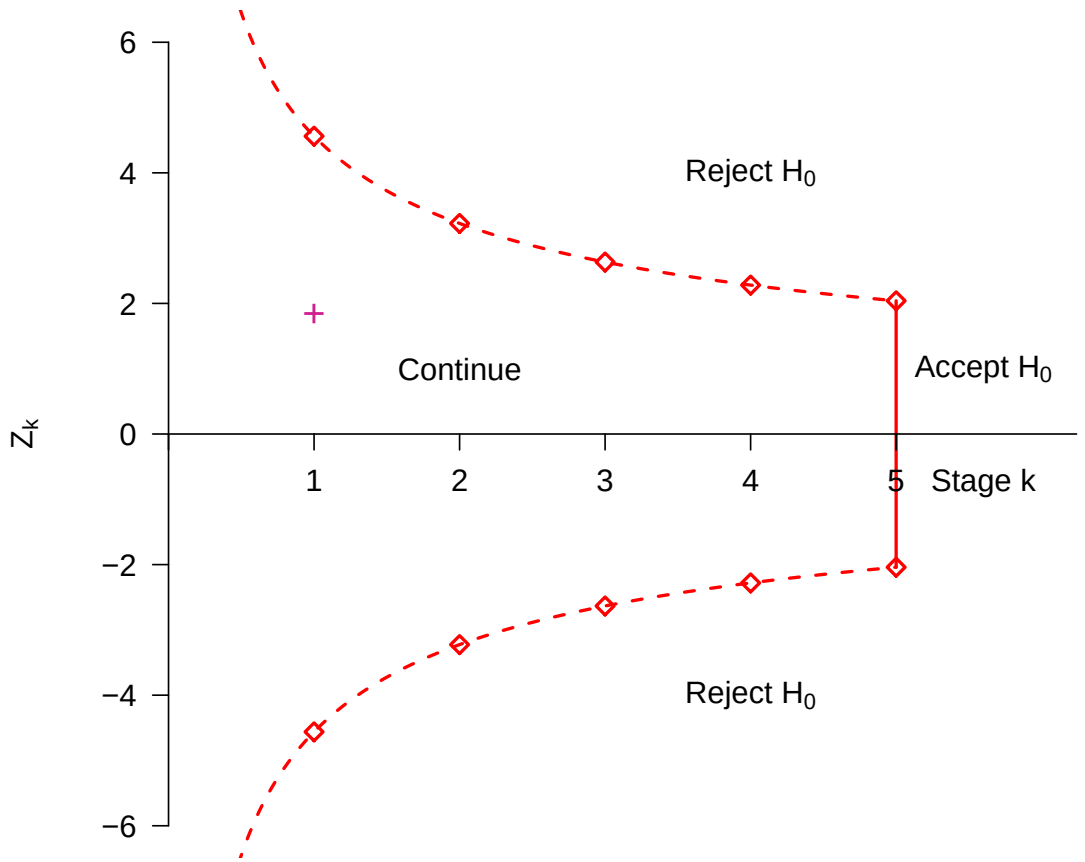
O'Brien & Fleming

- $c_k = C_B(K, \alpha) \times \sqrt{\frac{K}{k}}$
- Equivalent to perform repeated tests at levels $\alpha'_k = 2\left[1 - \Phi\left\{C_B(K, \alpha)\sqrt{\frac{K}{k}}\right\}\right]$
- $R_B(K, \alpha, \beta) < R_P(K, \alpha, \beta)$

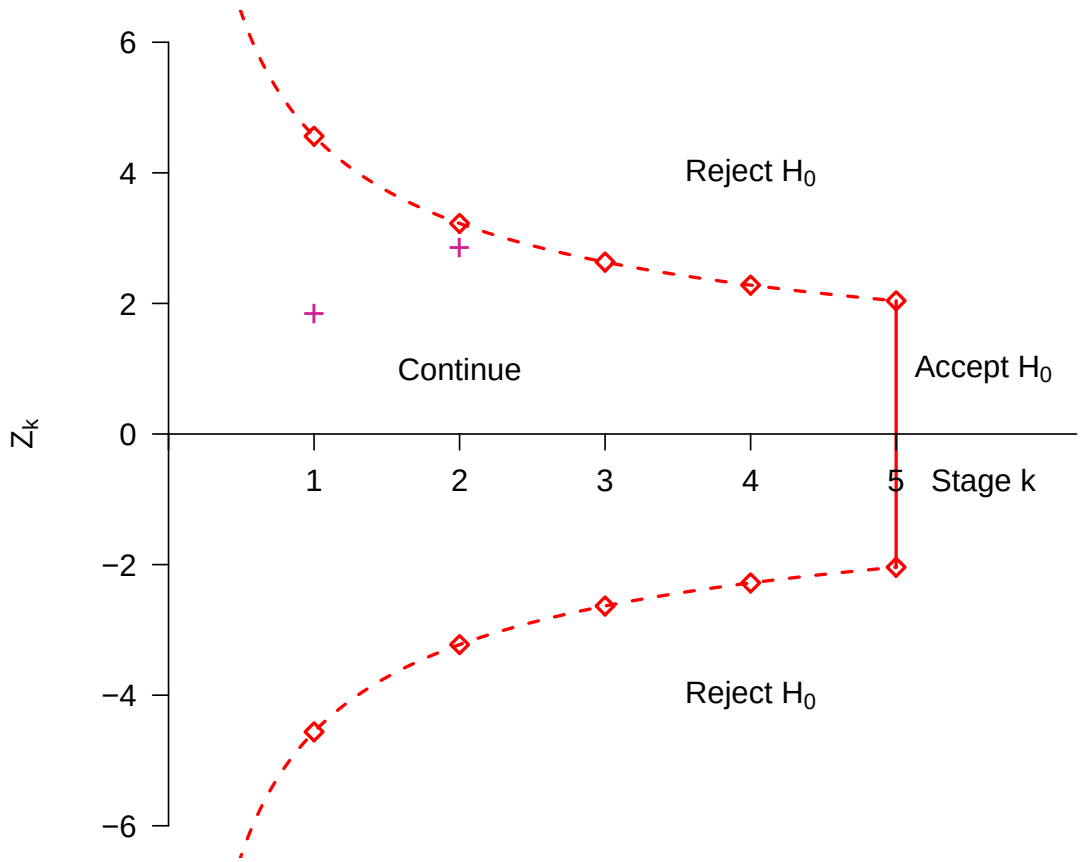
Boundaries *vs* no. subjects



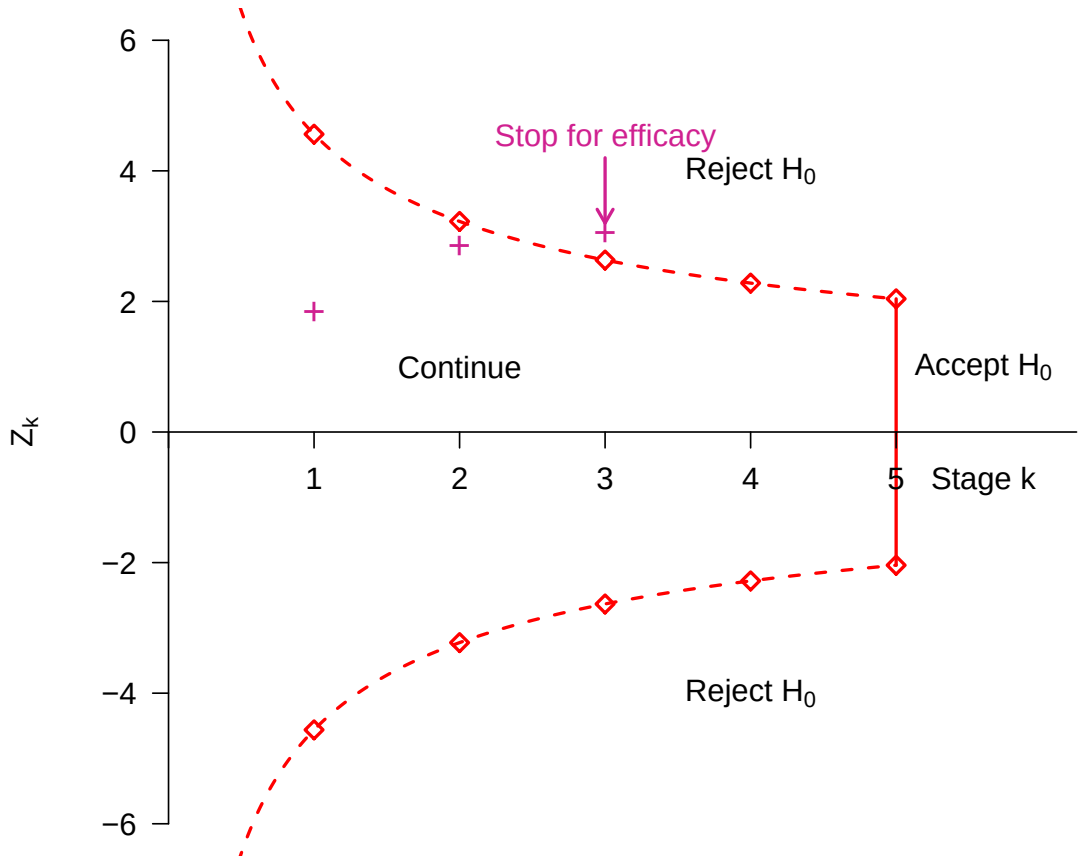
Fictive example: first IA



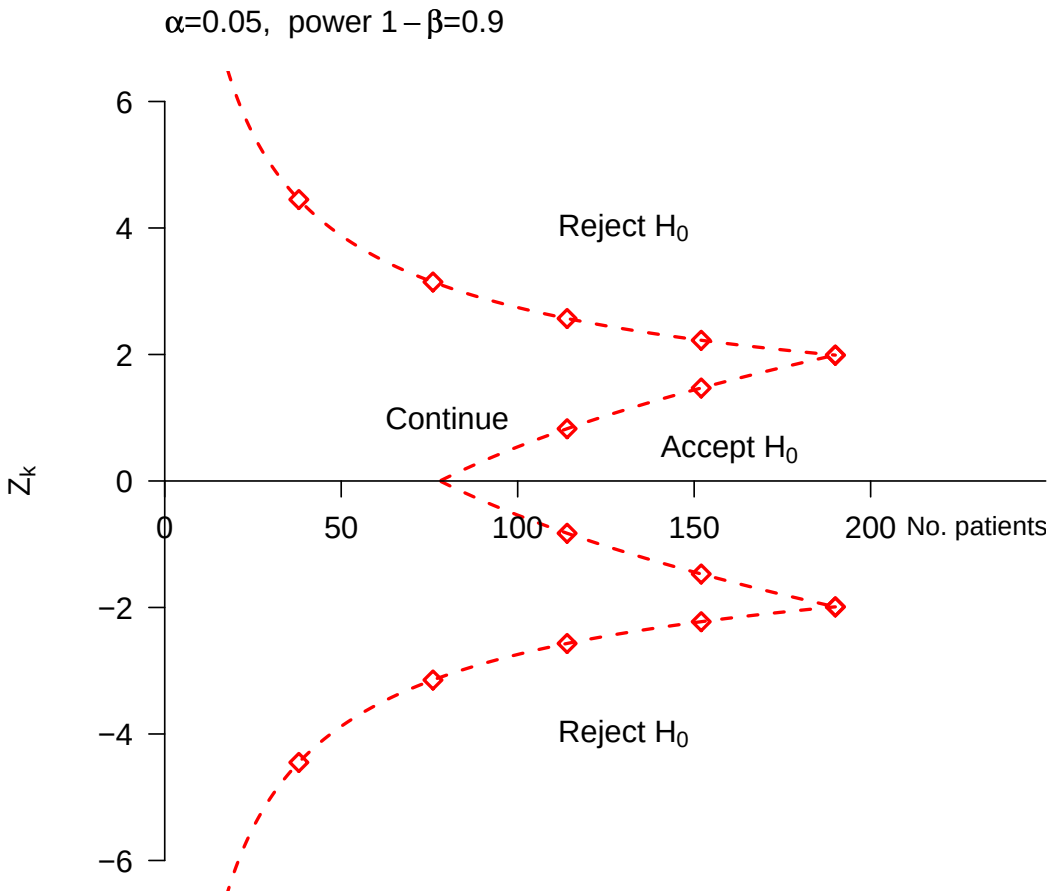
Fictive example: second IA



Fictive example: third IA



Stopping for futility

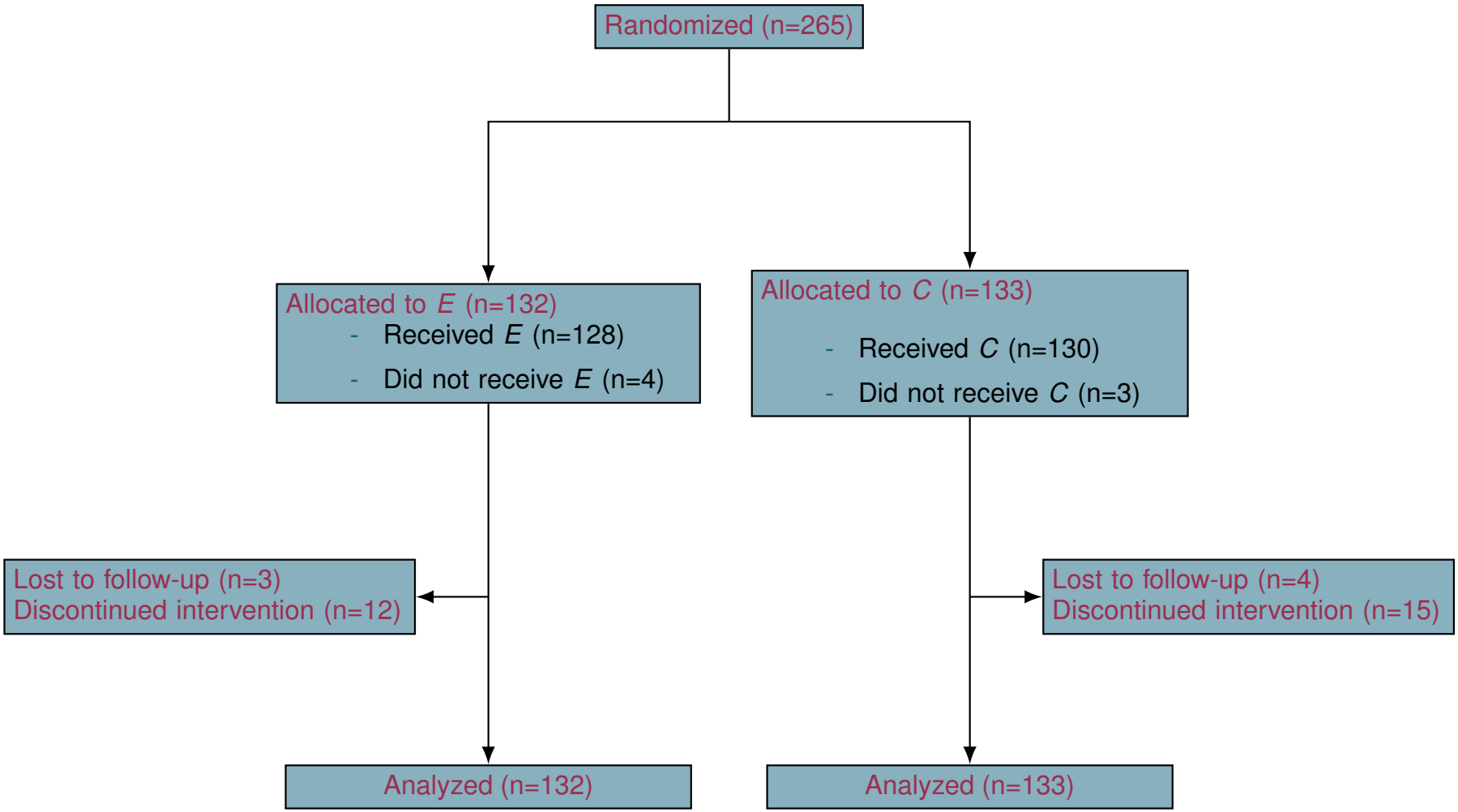


To be used?

- Shortens time to decision
- Little gain beyond $K = 5$ to 10
- OBF usually preferred to Pocock (less probability of early stopping, final test close to fixed sample)
- Need short-term outcomes
- Logistic constraints (get the data clean in time)
- Estimation of treatment effect problematic (optional sampling effect)
- Adaptive trials also exist (extension of GST)

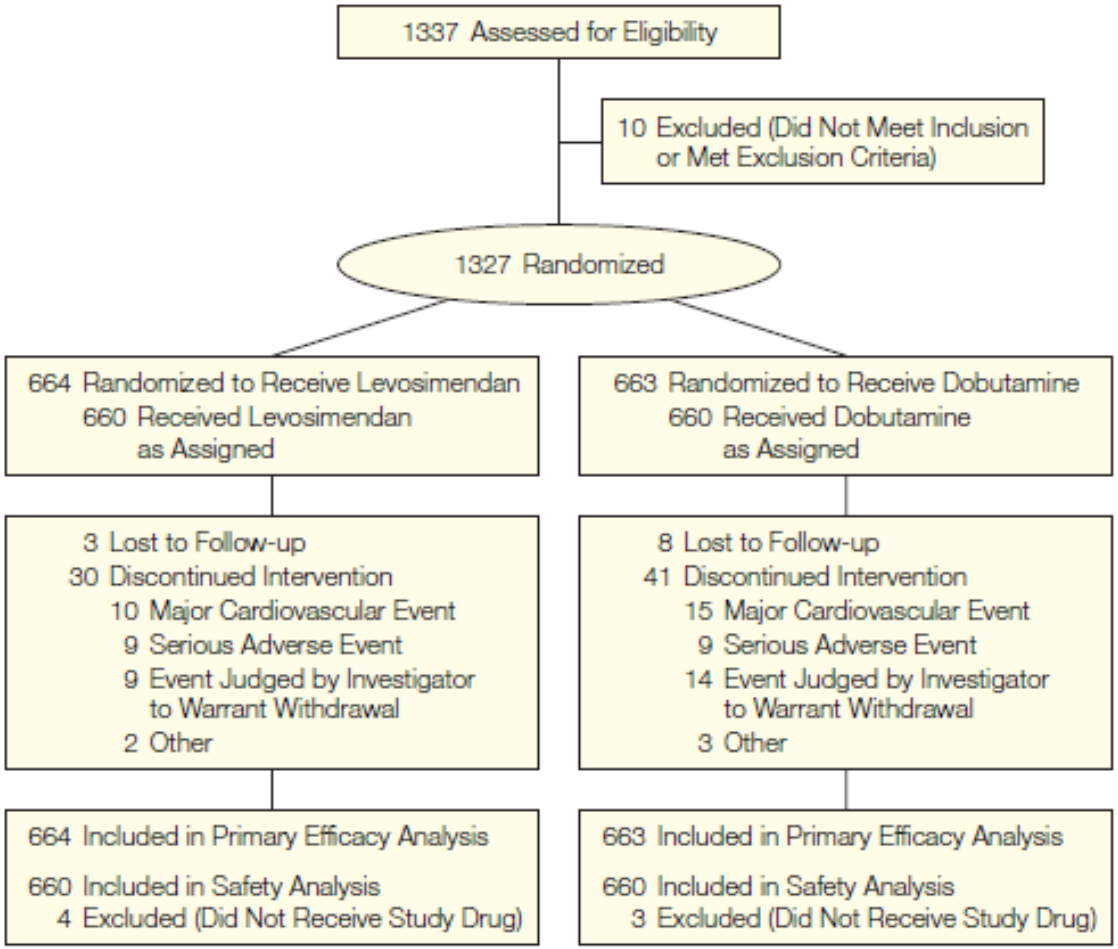
Outline

A more realistic flow chart



Example: SURVIVE trial

Figure 1. Participant Flow Through the Study



Missing outcome data

- We already know how to address loss to follow-up for the survival outcomes (censoring)
- But how to handle other endpoints (e.g. dyspnea or changes in BNP)?
- "Complete case" or "available case" analysis: discard patients with missing outcome
- ITT principle: as many patients analyzed as randomized (see flow charts)
- Moreover missingness may be related to response to treatment
- Find a value for the missing outcomes $\rightarrow$ impute a value

Imputing missing values in RCTs

- Complete case analysis would not be a major concern if very few outcomes are missing (few as compared to what?)
- Several strategies to impute missing outcomes
- Some are quite simple to undertake, but arguable
- Preferred methods are more complex
- In all cases, the mechanism underlying missing outcomes should be scrutinized

Simple strategy: worst case analysis

- For a binary outcome only
- Consider missing outcomes as failure if in E and as success if in C
- Likely to yield a biased estimate of treatment effect, but at least a lower bound
- If conclusions are in favor of E, then we know the analysis without any missing outcome would also yield the same
- Should be used at least as a sensitivity analysis

Simple strategy: LOCF

- Last Value Carried Forward: use the last recorded value for the patient
- OK for any type of outcome
- Assumption unlikely to be true even on average
- Dropping out is likely to be associated with response to treatment (e.g. failure to respond or adverse events)
- Despite arguable properties, often encountered in practice

Mechanisms for missing data

- MCAR
  - Missing data mechanism independent from the observed or unobserved value
  - Example: blood sample was lost by lab
  - Example for dop-out: patient moved to another location where no medical center participates to the trial
- MAR
  - Missing data are unrelated to the unobserved response but may be related to the observed ones
  - Example for dop-out: patients stopped the trial before visit 2 because his/her condition worsened at previous visit
- MNAR
  - The missing-value mechanism is related to the unobserved value
  - Previous example, but patient stopped before visit 1 because of worsening, and this could not be recorded because visit 1 was unattended

Other inadequate strategies

- Impute by the mean (or median or whatever) $\rightarrow$ OK **on average** if MCAR only
- Why not a model for the individual evolution of the outcome?
- Impute the missing value by the model-based prediction
- Both cases artificially decrease the variability of data
- Single imputation is a bad idea in general
- Need to also account for the variability of potential outcomes

Multiple imputation

- Based on (regression) models for the outcome
- Generates a series of imputed values for a single missing outcome, from a an estimated distribution
- The set of imputed (complete) datasets are then analyzed separately and results are combined afterwards
- Can be valid with MAR data, but not MNAR data
- More complex methods also exist (e.g. selection models, pattern-mixture models, . . .), not further developed here

Mixed models for repeated measures (MMRM)

- Regression model for the series of repeated measurements over time (see next section)
- Predicted values (average) depend on time and on other covariates
- Multiple imputations are obtained from the predicted values and the probability distribution of residuals
- Works well
  - If a "good" (reasonable?) model can be found
  - If other covariates are not missing

Multiple imputation by chained equations (MICE)

- Iterative algorithm
- First impute all missing values on each variable with missing values from their conditional distribution given the others
- Update the model and then imputations given the values obtained at the previous step
- Until convergence (rather quick in practice)
- R package `mice`

Combining results from mutliple datasets: Rubin's rule

- m imputed datasets: estimate $\hat{\delta}_i$ with variance $\hat{v}_i$
- Combined estimate $\bar{\delta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\delta}_i$
- Within imputation variance: $\bar{v}_w = \frac{1}{m} \sum_{i=1}^{m} \hat{v}_i$
- Between-imputation variance: $\bar{v}_b = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\delta}_i - \bar{\delta})^2$
- The total variance is then: $\bar{v}_t = \bar{v}_w + \left(1 + \frac{1}{m}\right) \bar{v}_b$

Outline

Outcomes measured at several timepoints

- Frequently encountered in RCTs
- Most often for continuous outcomes
- Example: SBP, BNP (SURVIVE trial), weight, ...
- Very frequently used as "changes from baseline"
- Sometimes more repeated measurements over the follow-up period

Changes from baseline

- Baseline measurment Y_0 and follow-up measurement Y_1
- Four ways to analyze the data
  A. Compare only post-treatment values Y_1 across groups
  B. Compare absolute changes $A_d = Y_1 - Y_0$
  C. Compare relative changes $R_d = (Y_1 - Y_0)/Y_0$
  D. ANCOVA: $Y_1 = \mu + \beta \times Y_0 + \delta \times 1_{\{E\}} + \epsilon$, and test $\delta = 0$, adjusted treatment effect

Example

- Main outcome = pain scale (0 to 100)

	n	baseline (Y_0)	f-up (Y_1)	A_d	R_d
E	30	42.2 (9.7)	20.8 (11.4)	21.5 (8.2)	52.6% (19.7)
C	30	39.7 (9.8)	31.1 (9.7)	8.4 (6.1)	21.1% (14.7)

- ANCOVA: $\hat{\delta} = -12.1$ (1.2)

Example: possible conclusions

A. Post-treatment pain is on average 10.3 pts lower with *E* (95%CI 4.9–15.7)

B. Pain decrease 13.1 pts higher on average with *E* (95%CI 9.3–16.8)

C. Relative decrease 31.6% hiher on average with *E* (95%CI 22.6–40.6)

D. Adjusted for initial pain, post-treatment pain on average 12.1 pts lower with *E* (95%CI 8.9–16.3)

Comparing the analyses

- Power of the different approaches according to the correlation between Y_0 and Y_1 (ρ)

Method	$\rho = 0.2$	$\rho = 0.35$	$\rho = 0.5$	$\rho = 0.65$	$\rho = 0.8$
A	70.5%	70.5%	70.5%	70.5%	70.5%
B	50.7%	59.2%	70.5%	84.8%	97.7%
C	45.1%	56.4%	67.0%	82.7%	97.1%
D	72.3%	76.1%	82.3%	90.8%	98.6%

$\rightarrow$ ANCOVA more powerful

Short mathematical proof

- Assume Y_0 and Y_1 have the same variance σ^2
- $\text{Var}(A_d) = 2(1-\rho)\sigma^2$, which is $< \sigma^2$ iff $\rho > 0.5$
- $\text{Var}(\hat{\delta}) = \text{Var}(Y_1 - \beta Y_0) = \text{Var}(Y_1 - \rho Y_0) = \sigma^2(1-\rho^2) \leq \sigma^2$

$\rightarrow$ Explains the comparison of A, B and D

- Less evident for C ...

Repeated measurements during follow-up

- Variation of an outcome over several timepoints
- Seldom the primary outcome: often one preferred timepoint
- Repeated measurements = correlated measurements
- Adequate modelling approaches exist

Two approaches

- Choose a summary measure
  - E.g. mean, area under the curve, slope, ...
  - Allows to get rid of the correlation
- Mixed regression model
- Don'ts
  - Repeat two-sample tests at the various timepoints (multiplicity + correlated results + interpretation!)
  - Use repeated measures ANOVA (for that purpose)

Mixed regression model

- Regression model (as before)
- I.e. a model with fixed effects
- But with random effects also
- Mixed linear model (random intercept):
  $Y_{ij} = \mu + b_i + \beta t_{ij} + \epsilon_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ with
  $b_i \sim \mathcal{N}(0, \sigma_b^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$
- Random slope model for the trial data:
  $Y_{ij} = \mu + b_{i0} + \beta t_{ij} + b_{i1} t_{ij} + \gamma 1_{\{x_i = E\}} t_{ij} + \epsilon_{ij}$, with
  $b_{i0} \sim \mathcal{N}(0, \sigma_0^2)$, $b_{i1} \sim \mathcal{N}(0, \sigma_1^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

Inference in mixed regression models

- More complex than what we have seen until know
- Even more complex for logistic and Cox models
- Also used to incorporate a center effect
- Center = level of clustering of data (as patient in previous slide)
- Assumes participating centers were drawn from a larger center population at random

Conclusion

- Short overview of methods used for analyzing RCTs
- No emphasis on planning (e.g. sample size, choice of a design), even if it is important
- Not exhaustive
- No emphasis (enough) on inference methods (likelihood, partial likelihood for Cox, . . .)
- Bayesian methods overlooked (but uncommon at best in RCTs)