

Analysis of Clinical Trials

Christy Chuang-Stein
Pfizer Global Research and Development

Alex Dmitrienko
Eli Lilly and Company

Geert Molenberghs
Universiteit Hasselt

1

Course Modules

- 8:30 – 10:15 Analysis of stratified data
- 10:15 – 10:30 *Morning break*
- 10:30 – 12:30 Multiple comparisons and endpoints
- 12:30 – 2:00 *Lunch*
- 2:00 – 3:15 Interim monitoring in clinical trials
- 3:15 – 3:30 *Afternoon break*
- 3:30 – 5:00 Handling incomplete data in
longitudinal studies

2

Objectives of the Course

- Learn common statistical approaches and statistical tests in the 4 topic areas, their strength and weaknesses.
- Identify the factors to consider when selecting appropriate statistical methods in the 4 topic areas.
- State the factors a clinical statistician or research scientist must keep in mind when applying the covered statistical methods to a clinical trial.
- Share the instructors' real-life experience with the course participants.

3

Analysis of Stratified Data

Christy Chuang-Stein

Statistical Research and Consulting Center
Pfizer Global Research and Development

4

Outline

- Type I, II, III analysis for continuous data
- Fixed vs random effects models
- Implications of enrollment on study designs
- CMH procedure for binary data
- Test for qualitative interaction
- Meta analysis for safety data
- Sample size in the presence of a covariate
- An example of the Bayes rule for binary data
- Concluding remarks

5

Analysis of Stratified Data

- Section I: Continuous data
 - ◆ Type I, II, III analysis; fixed vs random effect; implication of enrollment on estimates
- Section II: Binary data
 - ◆ CMH procedure for binary efficacy and safety data; test for qualitative interaction
- Section III: Sample size for a binary endpoint
 - ◆ When there is a major baseline covariate
- Section IV: Bayes rule for a diagnostic test

6

Notations for Continuous Case

- y_{ijk} represents the response of the k^{th} ($k=1, \dots, n_{ij}$), person receiving the i^{th} ($i=1, \dots, I$) treatment in the j^{th} ($j=1, \dots, J$) center. For simplicity, assume $I=2$.

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

$$n_i = \sum_{j=1}^J n_{ij} = \text{total number of subjects in } i^{\text{th}} \text{ group}$$

$$n = n_1 + n_2$$

- We will work with hypotheses expressed in (μ_{ij}) .

7

Type I Analysis

- Fit $\{\alpha_i\}$ after μ .

$$H_{0,I} : \sum_{j=1}^J \left(\frac{n_{1j}}{n_1} \right) \mu_{1j} = \sum_{j=1}^J \left(\frac{n_{2j}}{n_2} \right) \mu_{2j}$$

- For each treatment, find the weighted mean response over centers. The weights are proportional to the numbers of subjects receiving that treatment in the centers.
- Type I analysis tests if these two weighted means are the same or not.

8

Type II Analysis

- For Fit $\{\alpha_i\}$ after $\{\mu, \beta_j\}$.

$$H_{0,II} : \sum_{j=1}^J \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)^{-1} (\mu_{1j} - \mu_{2j}) = 0$$

- Find the difference in treatment response within each center. Calculate a weighted average of the differences.
- The weights are reciprocally proportional to the variances of sample mean differences within the strata, assuming that ε_{ijk} has a constant variance.

9

Type III Analysis

- Fit $\{\alpha_i\}$ after $\{\mu, \beta_j, (\alpha\beta)_{ij}\}$.

$$H_{0,III} : \frac{1}{J} \sum_{j=1}^J \mu_{1j} = \frac{1}{J} \sum_{j=1}^J \mu_{2j}$$

$$H_{0,III} : \frac{1}{J} \sum_{j=1}^J (\mu_{1j} - \mu_{2j}) = 0$$

- The (unweighted) average of $\{\mu_{1j}\}$ over centers is the same as the (unweighted) average of $\{\mu_{2j}\}$ over the centers.
- Alternatively, find the difference in treatment response within each stratum. $H_{0,III}$ states that the average of such differences is equal to 0.

10

Some Observations

- Type III analysis fits $\{\alpha_{ij}\}$ after $\{\mu, \beta_j, (\alpha\beta)_{ij}\}$, violating the marginality principle formulated by Nelder (1977, JRSS A, 140:48-76).
- Type III analysis weighs small and large centers equally. This completely differs from how one would normally do in a meta analysis.
- Type I analysis does not take stratification into consideration, which can lead to misleading conclusions as we will show later.
- Between Type II and Type III analyses, Type II analysis treats individual patients as experimental units and is closer to what we do when determining sample size.

11

Estimate of Treatment Effect

- The estimates correspond naturally to the hypotheses, obtained by replacing $\{\mu_{ij}\}$ with the observed sample means.

$$\hat{\Delta}_I = \sum_{j=1}^J \left(\frac{n_{1j}}{n_1} \right) \bar{y}_{1j} - \sum_{j=1}^J \left(\frac{n_{2j}}{n_2} \right) \bar{y}_{2j}$$

$$\hat{\Delta}_{II} = \frac{\sum_{j=1}^J \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)^{-1} (\bar{y}_{1j} - \bar{y}_{2j})}{\sum_{j=1}^J \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)^{-1}}$$

$$\hat{\Delta}_{III} = \frac{1}{J} \sum_{j=1}^J (\bar{y}_{1j} - \bar{y}_{2j})$$

12

Variance of the Estimate

- Assuming ε_{ijk} has the same variance σ^2 , the variances of these three estimates can be found to be

$$\text{Var}(\hat{\Delta}_I) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\text{Var}(\hat{\Delta}_{II}) = \sigma^2 \left(\sum_{j=1}^J \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)^{-1} \right)^{-1}$$

$$\text{Var}(\hat{\Delta}_{III}) = \frac{\sigma^2}{J^2} \sum_{j=1}^J \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)$$

$$\text{Var}(\hat{\Delta}_I) \leq \text{Var}(\hat{\Delta}_{II}) \leq \text{Var}(\hat{\Delta}_{III})$$

13

Minimizing Variance

- For all three analyses, the variances are the smallest when n_{ij} are the same (n_0). In this case,

$$\text{Var}(\hat{\Delta}_I) = \text{Var}(\hat{\Delta}_{II}) = \text{Var}(\hat{\Delta}_{III}) = \frac{4\sigma^2}{n}$$

- In the above, $n = J(2n_0)$ is the total number of patients. This minimum variance is the one we usually use to calculate the sample size for a study.

$$n = 2 \left\{ \frac{2\sigma^2 \left(z_{(1-\alpha/2)} + Z_{(1-\beta)} \right)^2}{\Delta_0^2} \right\}$$

14

Implications

- The sample size calculation assumes the optimal situation. That is – all centers enroll the same number of patients, perfectly balanced between the two treatment groups. The more $\{n_{ij}\}$ differ, the larger the variance is.
- Since centers typically differ in enrollment rate, are we under-estimating the variability at the design stage?
- So far, $\{\mu_{ij}\}$ are considered fixed and viewed as a collection of fixed and potentially distinct values. Trials conclusions are made specific to these $\{\mu_{ij}\}$.
- Is it more reasonable to view the set of $\{\mu_{ij}\}$ as a sample from a population of $\{\mu_{ij}\}$ corresponding to all centers?

15

Random Effects Models

- Defining y_{ijk} as before, random effects model assumes

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

$$E(\varepsilon_{ijk}) = 0, \quad \text{Var}(\varepsilon_{ijk}) = \sigma^2$$

$$E(\mu_{ij}) = \eta_i, \quad \text{Cov}(\mu_{1j}, \mu_{2j}) = V$$

$$\delta_j = \mu_{1j} - \mu_{2j}, \quad E(\delta_j) = \eta_1 - \eta_2$$

- We are interested in $\eta_1 - \eta_2 = \eta$. It can be considered as the mean treatment effect over a population of centers.

16

Variance under Mixed-Effects Models

- For a given J (# of centers) and n (total patients in a study), Fedorov et al (SMMR, 2005) showed that the variance of the estimate is the smallest when n_{ij} is the same for all centers. When this occurs,

$$Var(\hat{\eta}) = \frac{4\sigma^2}{n} + \frac{Var(\delta_j)}{J}$$

- The first term above is the variance under the fixed effect modeling approach. The second term relates to the random nature of $\{\delta_j\}$.

17

Fixed Effects vs Random Effects

- By choice, the treatment effect estimate under the random effects model has a larger variance than that under the fixed effects approach.
- Conclusions from random effects models are more general than those from fixed effects models.
- ICH E9 (Statistical Principles) on centers:
 - ◆ Should define centers clearly in the protocol. If pooling is to be done, should pre-specify and justify.
 - ◆ If randomization is stratified by center, center should be included in the model.
 - ◆ Mixed effects models may be used to explore the heterogeneity of the treatment effect, treating center and treatment by center effects as random.

18

What Does This Mean?

- Whether we treat center effect as fixed or random, the more imbalanced centers are with respect to enrollment, the higher the variance is. This affects the precision of estimates, and the power for testing a hypothesis related to the treatment effect.
- Enrollment pattern is surprisingly similar across trials for the same indication in a therapeutic area.
- When designing pivotal phase III studies, we should examine power relative to the enrollment pattern and adjust the sample size accordingly. Use simulations to help conduct the evaluation.

19

A Sepsis Study

- A confirmatory trial in severe sepsis, a double-blind placebo control trial; IV with 96 hours duration; randomization stratified by center.
- Primary analysis was 28-day mortality rate after treatment onset, stratified by 3 pre-specified covariates: APACHE II score, age and protein C activity.
- Trial was terminated by an independent DSMB for efficacy after 2nd interim analysis of 1520 patients.
- Many subgroup analyses were conducted, including APACHE II subgroups (4 defined by the observed quartiles), subgroups defined by the components of the APACHE II score, and subgroups defined by 1, or 2, or 3, or at least 4 organ dysfunctions.

20

Results from the Sepsis Study

APACHE II Score	New Treatment		Placebo	
	Mortality Rate	Total	Mortality Rate	Total
3 - 19 (1 st Q)	15%	218	12%	215
20 - 24 (2 nd Q)	22%	218	26%	222
25 - 29 (3 rd Q)	24%	204	36%	162
30-53 (4 th Q)	38%	210	49%	241

21

When Dealing with Binary Outcome

- Let π_{ij} ($i=1,2; j=1,2,3,4$) denote the mortality rate associated with treatment i in APACHE II stratum j . Denote the observed rate by p_{ij} .
- Three measures are commonly used to assess efficacy within the j^{th} APACHE II stratum
 - ◆ Risk difference $d_j : \pi_{1j} - \pi_{2j}$
 - ◆ Relative risk $r_j : \pi_{1j} / \pi_{2j}$
 - ◆ Odds ratio $o_j : \{ \pi_{1j} (1 - \pi_{2j}) \} / \{ (1 - \pi_{1j}) \pi_{2j} \}$
- We will focus on risk difference. Within each stratum, estimate $\pi_{1j} - \pi_{2j}$ by $p_{1j} - p_{2j}$. Need to combine them to get an overall treatment effect.

22

Combining Binary Results

- A common approach is to weigh the estimated d_i within a stratum by the inverse of its variance. The resulting estimate is the CMH-estimate. Asymptotic confidence intervals for the “weighted” treatment effect can be constructed using the normal approximation.

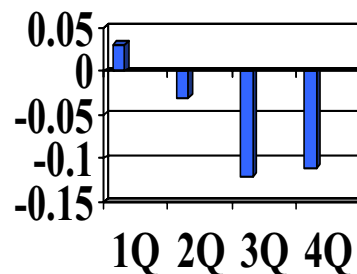
$$\hat{d} = \frac{\sum_{j=1}^4 \left(\text{Var}(\hat{d}_j) \right)^{-1} \hat{d}_j}{\sum_{j=1}^4 \left(\text{Var}(\hat{d}_j) \right)^{-1}}$$
$$\text{Var}(\hat{d}) = \left(\sum_{j=1}^4 \left(\text{Var}(\hat{d}_j) \right)^{-1} \right)^{-1}$$

A 95% Confidence interval is (-0.081,-0.001). The weights used are in the ratios of 2 : 1.3 : 0.9 : 1.

23

Comparing across Strata

- The difference in the mortality rates (new treatment – placebo) in the 4 APACHEII strata range between 3% to -12%.
- The graph suggests an interaction which might be qualitative in nature.
- We will look at an approach proposed by Gail and Simon (1985, Biometrics, 41:361-372) to test for qualitative interaction.



24

Test for Qualitative Interaction

- Let $O^+ = \{d_j \geq 0\}$ = set of non-negative differences
- Let $O^- = \{d_j \leq 0\}$ = set of non-positive differences

$$Q^+ = \sum_{j=1}^J \frac{\hat{d}_j^2}{s_j^2} I(\hat{d}_j > 0), \quad Q^- = \sum_{j=1}^J \frac{\hat{d}_j^2}{s_j^2} I(\hat{d}_j < 0)$$

$$Q = \min(Q^+, Q^-)$$

- $Q > c$ can be used to test the null hypothesis of no qualitative interaction.
- Gail and Simon showed that Q follows a fairly complex distribution based on a weighted sum of chi-square distribution. SAS codes are available in the text book.

25

Test for Qualitative Interaction

- Q^+ can be used to test the null hypothesis of all differences being negative. Q^- can be used to test the null hypothesis of all differences being positive.
- For the sepsis study, the two-sided Gail-Simon test has a P-value of 0.4822.
- The one-sided P-value for H_0 of positive differences is 0.0030. The one-sided P-value for H_0 of negative differences is 0.4822.
- Like other interaction tests, G-S test requires strong evidence before we can reject the no qualitative interaction hypothesis.

26

In the End...

- Data from this single study led to the approval of Xigris®
- Xigris® INDICATIONS AND USAGE

Xigris is indicated for the reduction of mortality in adult patients with severe sepsis (sepsis associated with acute organ dysfunction) who have a high risk of death (e.g., as determined by APACHE II).

Safety and efficacy have not been established in adult patients with severe sepsis and lower risk of death.

27

Table in the Package Insert

APACHE II Quartile score	Xigris		Placebo	
	Total	Mortality rate	Total	Mortality rate
1 st + 2 nd (3-24)	436	18.8%	437	19.0%
3 rd + 4 th (25-53)	414	30.9%	403	43.7%

- Patients who have a high risk for death are represented by an APACHE II score in the 3rd and 4th APACHE II score categories.
- Treatment effects need to differ more than what shown in this case for Gail-Simon test to conclude interaction.

28

The LIFE Study

- Losartan Intervention For Endpoint Reduction in Hypertension Study.
 - ◆ Conducted at 945 sites in 7 countries.
 - ◆ Enrolled 9193 hypertensive patients with left ventricular hypertrophy (LVH)
 - ◆ The primary endpoint is a composite endpoint of cardiovascular deaths, stroke, and myocardial infarction.
- Results reviewed by the FDA Cardiovascular and Renal Drugs AC on Jan 6 2003 for a new proposed indication

Cozaar is indicated to reduce the risk of cardiovascular morbidity and mortality as measured by the combined incidence of cardiovascular death, stroke, and myocardial infarction in hypertensive patients with left ventricular hypertrophy.

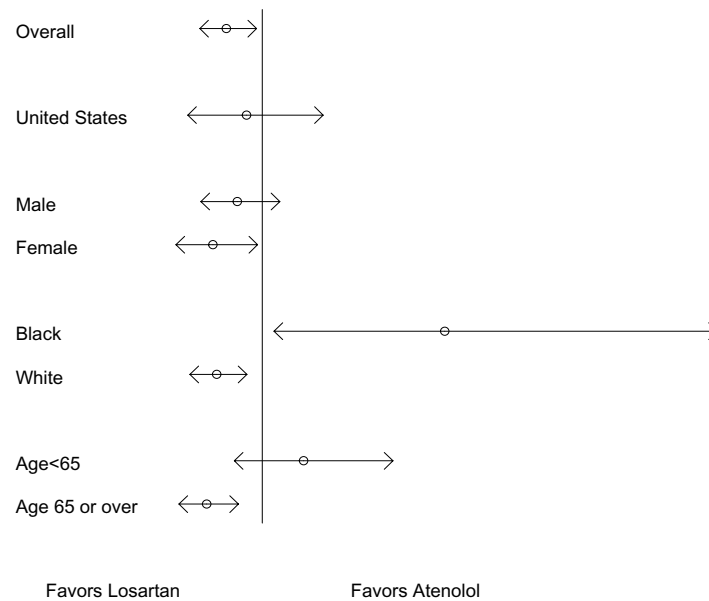
29

Some Background

- Losartan's then label states that the effect in blood pressure reduction in blacks was somewhat less than in that in whites (a common statement for beta-blockers).
- FDA statistician quoted data from three endpoint studies of other drugs. These studies demonstrated less or no treatment effect in blacks when compared to whites.
- On the primary endpoint, when compared to atenolol, losartan had a hazards ratio of 0.869 (95% CI from 0.772 to 0.979) with a P-value of 0.021. The effect came primarily from the stroke component of the composite.
- The issue of how losartan compared to atenolol in blacks came up.

30

Hazard Ratio and 95% CIs - Primary Endpoint



31

Gail-Simon Test

- Nominal p-value for Black vs. Non-Black Qualitative Interaction = 0.016.
- Impossible to correctly adjust this p-value for multiple comparisons post hoc.
 - ◆ 3 subgroups pre-specified for special importance (U.S. region, Diabetics, ISH)
 - ◆ To do it correctly, the formal analysis plan would need to list all important subgroups and specify a method to correctly adjust for number of tests.

Source: John Lawrence's (FDA Statistical Reviewer) slides at the January 6 2003 FDA AC meeting. For more discussion, see <http://www.fda.gov/ohrms/dockets/ac/03/slides/3920s1.htm>

32

COZAAR[®] Package Insert

Indications and Usage

... COZAAR is indicated to reduce the risk of stroke in patients with hypertension and left ventricular hypertrophy, but there is evidence that this benefit does not apply to Black patients. ...

Clinical Pharmacology

In the LIFE study, Black patients treated with atenolol were at lower risk of experiencing the primary composite endpoint compared with Black patients treated with COZAAR.... This finding could not be explained on the basis of differences in the populations other than race or on any imbalances between treatment groups... the LIFE study provides no evidence that the benefits of COZAAR on reducing the risk of cardiovascular events in hypertensive patients with left ventricular hypertrophy apply to Black patients.

33

Observations

- In the case of Xigris, subgroups defined by APACHE II score were pre-specified. Statistical significance was not achieved by the Gail-Simon test at the 5% level.
- In the case of COZAAR, race subgroups were not among the pre-specified primary subgroups of special importance. They are, however, among the “usual” demographic subgroups we routinely report descriptive statistics. A data-driven Gail-Simon test produced a value less than 0.05.
- The end results (language in the product package insert) are similar.
- Regulatory actions in both cases were not completely based on statistical test results.

34

Clinical Summary of Safety

Study	Drug A	# of Pts	Drug B	# of Pts
1	8%		4%	
2	7%		6%	
3	1%		1%	
4	1%		2%	
5	21%		20%	
6	8%		10%	
Total Avg	13%	1000	9.5%	750

13% vs 9.5%: a two-sided P-value of 0.023.

35

Clinical Summary of Safety

Study	Drug A	# of Pts	Drug B	# of Pts
1	8%	100	4%	100
2	7%	100	6%	100
3	1%	100	1%	100
4	1%	100	2%	100
5	21%	500	20%	250
6	8%	100	10%	100
Total Avg	13%	1000	9.5%	750

95% CMH confidence interval for the diff (Drug A – Drug B) is (-0.017, 0.018). The point estimate is 0.001.

36

Simpson's Paradox

Trt	Moderate			Severe		
	R	NR	Total	R	NR	Total
A	32 (80%)	8 (20%)	40	12 (20%)	48 (80%)	60
B	48 (80%)	12 (20%)	60	8 (20%)	32 (80%)	40

- With the moderate stratum, response rate is 80% for both groups. Within the severe stratum, response rate is 20% for both group.
- There is a moderate imbalance: 60% of Trt A patients were severe while 40% of Trt B patients were severe.

37

Simpson's Paradox

Trt	R	NR	Total
A	44 (44%)	56 (56%)	100
B	56 (56%)	44 (44%)	100

- The chi-square statistic for homogeneity has a value of 2.88 with 1 d.f.
- If the sample size doubles with the same configurations, the chi-square statistic will be 5.76 (2x2.88). The 95th percentile of chi-square with 1 df is 3.84.
- Conducting un-stratified analysis in this case could lead to an erroneous conclusion.

38

Observations

- CMH procedure is frequently used to combine efficacy results from multiple studies in a meta analysis. It can be used to combine results from strata within a study.
- CMH procedure performs well in sparse stratifications.
- Consider applying CMH procedure when summarizing safety data from different studies, especially when studies have very different patient populations and the randomization ratio varies. The naïve approach of pooling data from multiple can produce misleading results.
- Need to stratify randomization or analysis over covariates highly correlated with response.
- Remember the Simpson Paradox.

39

Designing the Sample Size

Disease Severity Score	Treatment Group		Difference (Treatment – Placebo)
	Treatment	Placebo	
1Q	12%	12%	0%
2Q	21%	24%	-3%
3Q	27%	36%	-9%
4Q	36%	48%	-12%

When the response (e.g. mortality) rate is low, there is not much room to improve. Most of the benefit is in the high-risk population.

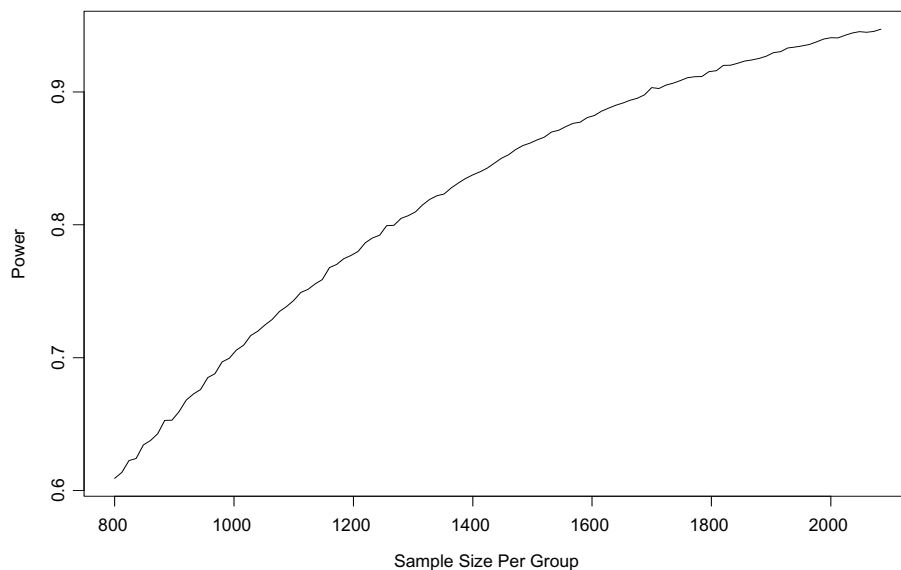
40

Implications on Sizing a Study

- Treatment effect under the un-stratified analysis is 6%. Need 860 patients per group for an 80% power to detect a treatment effect, based on a two-sided test at the 2-sided 5% level.
- Under the CMH procedure, effects in different strata (equal stratum size) will be combined using weights roughly in the ratios of 2.23 : 1.38 : 1.20 : 1.00. The combined effect is 4.5%.
- 860 patients per group will only have a 64% power to conclude a treatment effect under the assumed efficacy configuration.
- Can use simulations to determine sample size.

41

Power as a Function of Sample Size



42

Reminder

- Since variance is a function of the response rate, the CMH estimate will be affected by the response rates in different strata.
- The discussion is also relevant when
 - ◆ The absolute difference is the same across strata, but the control group has different response rates in different strata.
 - ◆ The relative improvement (death rate is reduced by 20% on the relative scale) is the same across categories defined by an influential baseline covariate.
- As long as there exists a baseline covariate that is associated with the binary response, we have an issue.

43

A Diagnostic Test Example

Diagnostic Test Result	Case (Disease)	Control (No Disease)	Total
Positive	72	6	78
Negative	8	74	82
	80	80	160

$P(\text{Positive} \mid \text{Case}) = 72/80$ (90%) \rightarrow *Sensitivity*

$P(\text{Positive} \mid \text{Control}) = 6/80$ (7%)

$P(\text{Negative} \mid \text{Control}) = 74/80$ (93%) \rightarrow *Specificity*

What is $P(\text{Disease} \mid \text{Positive})$ (positive predictive value)? Is it $72/78$ (92%)?

44

Bayes Theorem

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

- This shows how the initial probability for event A (i.e., $P(A)$) is updated when taking into account information contained in event B.
- A formal mechanism for learning from experience.

45

Diagnostic Test Example

Diagnostic Test Result	Case (Disease)	Control (No Disease)	Total
Positive	72	24	96
Negative	8	296	304
	80	320	400

Increasing the size of “control” by 4-fold, $72/96 = 75\%!!!$

$$\begin{aligned} P(D | \text{Positive}) &= P(D) \frac{P(\text{Positive} | D)}{P(\text{Positive})} \\ &= \{P(D) (72/80)\} / \{P(\text{Pos} | D) P(D) + P(\text{Pos} | \text{ND}) P(\text{ND})\} \\ &= \{P(D) (72/80)\} / \{(72/80) P(D) + (24/320) P(\text{ND})\} \end{aligned}$$

We need $P(D)$, the disease prevalence rate.

46

Diagnostic Test Example

- If $P(D) = 1/10,000$, then $P(D | \text{Pos}) = 0.0013$
- If $P(D) = 1/1,000$, then $P(D | \text{Pos}) = 0.013$
- If $P(D) = 1/100$, then $P(D | \text{Pos}) = 0.12$
- In each case, getting a positive result increases the likelihood of having the disease by about 13 folds.
- One major factor in the sensitivity ($P(D | \text{positive test result})$) is the prevalence of the disease in the population. Might not be worthwhile to run the diagnostic test in a extremely low-risk population.

47

CHMP/EWP/2863/99 - Baseline Adjustment

Adopted by CHMP in May 2003

- Variables known *a priori* to be strongly, or at least moderately, associated with the primary outcome should be considered as covariates in the primary analysis. These covariates should be pre-specified in the protocol or the statistical analysis plan. Pre-specification is key to credibility.
- The baseline of a continuous primary endpoint, if available, should usually be included in the model as a covariate.
- Only a few covariates should be included in a primary analysis. The primary model should not include treatment by covariate interactions.
- Whenever adjusted analyses are presented, results of the treatment effect in relevant subgroups should be presented.
- Model assumptions should be checked and sensitivity analysis conducted, whenever possible.

48

Multiple comparisons and endpoints

Alex Dmitrienko
Eli Lilly and Company

Slide 49

Outline

Multiple testing in clinical trials

Popular multiple tests

- Single-step, closed, fixed-sequence tests

Analysis of multiple endpoints

Gatekeeping testing strategies

Resources

- Analysis of Clinical Trials Using SAS (Chapter 2)
- Annotated bibliography of multiple comparison papers (<http://biopharmnet.com/doc/doc13001.html>)

Slide 50

Part I: Multiple testing in clinical trials

Multiple treatment comparisons

- Several treatment groups (several doses of a drug) are compared to a control

Multiple primary endpoints

- Multiple criteria for assessing efficacy or safety of a drug

Multiple secondary analyses

- Multiple secondary endpoints or subgroup analyses

Multiple interim looks

Slide 51

Type I error rate inflation

Control of Type I error rate

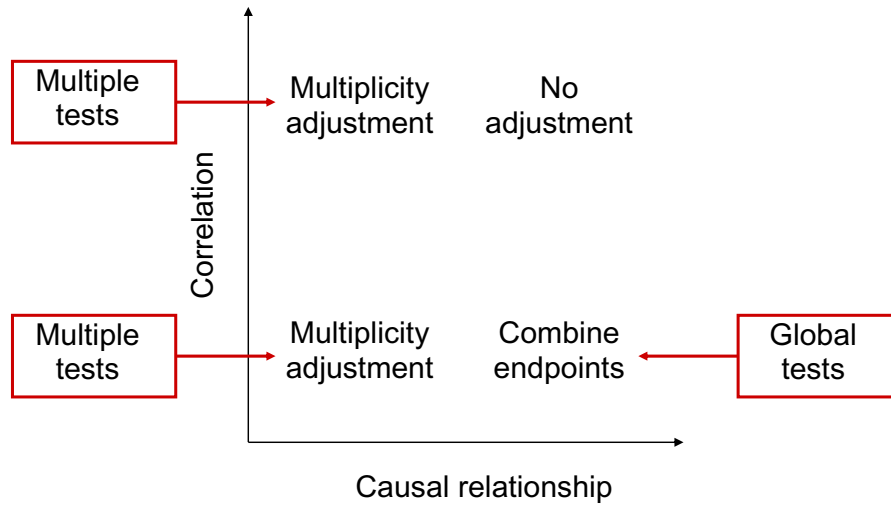
- Univariate tests increase the overall false-positive rate
- Control of the false-positive rate is mandated in registration trials
- False positive findings could lead to approval of inefficacious drugs

CPMP guidance document

- “Points to consider on multiplicity issues in clinical trials” (Sep 19, 2002)
- “A clinical study that requires no adjustment of the Type I error is one that consists of **two treatment groups**, that uses a **single primary variable**, and has a confirmatory statistical strategy that pre-specifies just **one single null hypothesis** relating to the primary variable and **no interim analysis**.”

Slide 52

Multiplicity adjustment



Slide from a presentation given by Dr. Mo Huque, FDA, in 2004

Slide 53

Type I error rate: Strong control

Dose-ranging hypertension clinical trial

- Three doses (D1, D2 and D3) are compared to placebo (D0)
- Continuous primary endpoint
- Three null hypotheses, $H_{01}: \mu_1 = \mu_0$, $H_{02}: \mu_2 = \mu_0$, $H_{03}: \mu_3 = \mu_0$

Strong control (control of familywise error rate, FWER)

- Probability of erroneously concluding D3 is efficacious regardless of the drug effect at D1 or D2
- This definition is suitable for making specific claims
- Dunnett test provides a strong control of Type I error rate

Slide 54

Part II: Popular multiple tests

Use of marginal p-values

- Rely on the marginal distribution of individual test statistics
- Less powerful than tests that account for correlation
- Bonferroni, Holm, Hochberg, Hommel tests

Single-step tests

- Examine each individual null hypothesis

Stepwise tests

- Hypotheses are tested in a sequential manner
- Some may be retained or rejected by implication

Compare multiple and global tests

Slide 55

Bonferroni test

Notation

- Consider m null hypotheses, H_{01}, \dots, H_{0m}
- Associated p-values, p_1, \dots, p_m

Bonferroni test

- Reject H_{01} if $p_1 \leq \alpha/m$
- Reject H_{02} if $p_2 \leq \alpha/m$, etc

Control of familywise error rate

- Controls familywise error rate for arbitrary dependence structures

Slide 56

Simes test

Global tests

- Global null hypothesis is the intersection of H_{01}, \dots, H_{0m}
- Consider ordered p-values, $p_{(1)} \leq \dots \leq p_{(m)}$

Simes test is a global test

- Rejects global null hypothesis if
 $p_{(1)} \leq \alpha/m$ or $p_{(2)} \leq 2\alpha/m$ or $p_{(3)} \leq 3\alpha/m$ or ... or $p_{(m)} \leq \alpha$
- Uniformly more powerful than Bonferroni global test

Control of familywise error rate

- Only under additional assumptions, e.g., independent or positively dependent test statistics
- Maximum FWER is 1.5α if $m=2$ and 2.08α if $m=4$

Slide 57

Adjusted p-values

Multiplicity adjustment

- Significance levels are adjusted downward or p-values are adjusted upward
- An adjusted p-value is the smallest significance level for which one would reject the corresponding null hypothesis
- Adjusted p-values are more convenient to work with (can be used with any α)

Dose-ranging clinical trial

Comparison	Raw p-value	Adjusted p-value	
		Bonferroni	Simes
D1 vs D0	0.0473	0.1419	
D2 vs D0	0.0167	0.0501	0.0251
D3 vs D0	0.0152	0.0456	

Slide 58

Exercise

Schulz and Grimes (2005) *The Lancet*

- A new antibiotic versus a standard antibiotic

Two endpoints

- A 50% reduction in the incidence of fever ($p=0.048$)
- A 52% reduction in the incidence of wound infections ($p=0.041$)
- Biologically related and statistically correlated
- No significant effect after a multiplicity adjustment (Bonferroni test)

Conclusion

- Multiplicity adjustments sabotage interpretation of clinical trials

Your response?

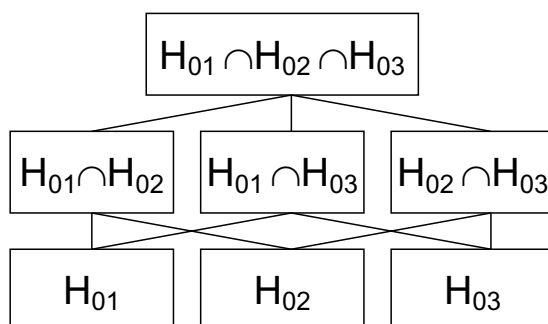
Slide 59

Closed tests

Closed testing principle

- Very powerful method for creating multiple tests
- Any single-step or stepwise test can be written as a closed test

Dose-ranging clinical trial



Principle

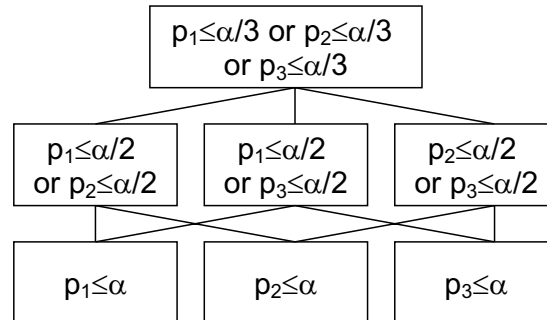
- Test each intersection hypothesis at α level
- A null hypothesis is rejected if its test and all tests for hypotheses implying it are significant
- Resulting test controls FWER at α level

Slide 60

Holm test as a closed test

Holm test

- Equivalent to a closed test in which each intersection hypothesis is tested using Bonferroni test

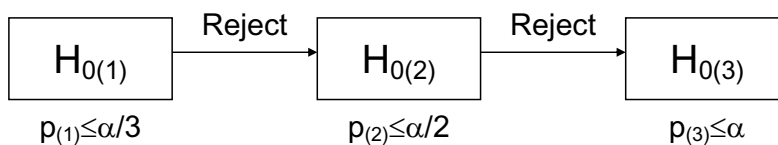


Slide 61

Holm test as a stepwise test

Stepwise version

- Ordered p-values, $p_{(1)} < p_{(2)} < p_{(3)}$
- Associated hypotheses $H_{0(1)}$, $H_{0(2)}$, $H_{0(3)}$



Holm test

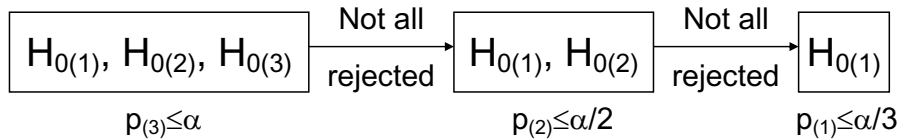
- Closed test derived from Bonferroni test
- Uniformly more powerful than Bonferroni test

Slide 62

Other closed testing procedures

Hochberg test

- Stepwise test derived from Simes test
- Mirror image of Holm test but uniformly more powerful than Holm test



Hommel test

- Closed test derived from Simes test
- No stepwise version but software implementation is available
- Uniformly more powerful than Hochberg test (equivalent to Hochberg test when there are two hypotheses)

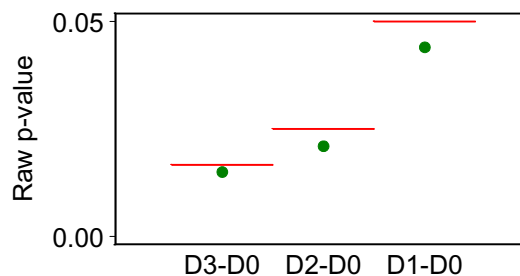
Slide 63

Dose-ranging clinical trial example

Dose-ranging clinical trial

- D1 vs D0, $p_1=0.044$
- D2 vs D0, $p_2=0.021$
- D3 vs D0, $p_3=0.015$

Compare Bonferroni and Holm tests



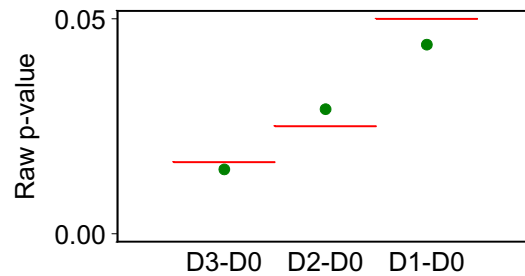
Slide 64

Dose-ranging clinical trial example

Dose-ranging clinical trial

- D1 vs D0, $p_1=0.044$
- D2 vs D0, $p_2=0.029$
- D3 vs D0, $p_3=0.015$

Compare Holm and Hochberg tests

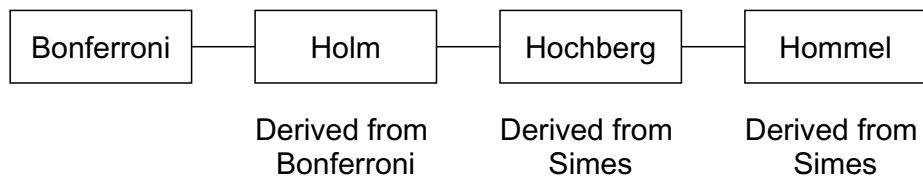


Slide 65

Comparison of multiple tests

Less power

More power



Software implementation

- Marginal tests are easy to carry out using PROC MULTTEST
- BONFERRONI option to request Bonferroni test
- STEPBON to request Holm test, etc

Slide 66

Power and Type I error rate

Power loss

- Marginal tests perform poorly when test statistics are correlated or a large number of null hypotheses
- Hommel test in the analysis of normally distributed endpoints with the correlation of 0.9
 - Type I error rate is 0.04 when 3 endpoints
 - Type I error rate is 0.03 when 10 endpoints

Type I error rate inflation

- Hochberg and Hommel tests do not always control Type I error rate
- Difficult to characterize cases when inflation occurs
- Generally accepted by regulatory agencies

Slide 67

Exercise

Cardiovascular clinical trial

- Study start: A single primary endpoint
- Interim analysis: A co-primary endpoint was added
- How to control the Type I error rate?

Proposal

- Bonferroni test is too conservative and alternatives will be considered (Holm, Hochberg and Hommel tests)
- Power comparison: Holm < Hommel < Hochberg
- Hommel test does not always control Type I error rate but Hochberg test does
- Hochberg test is superior to Hommel test and will be used in the study

Slide 68

Fixed-sequence tests

Multiple tests without a multiplicity adjustment

- A natural ordering among the null hypotheses

Doses are compared to a control in a sequential manner

- Find the minimally effective dose

Longitudinal measurements are analyzed in a sequential manner

- Find the earliest timepoint when the experimental drug shows a statistically significant effect over the control

Slide 69

Dose-ranging clinical trial example

Fixed-sequence testing

- Reject H_1 if $p_1 \leq \alpha$
- If H_1 was rejected, reject H_2 if $p_2 \leq \alpha$, otherwise stop
- If H_2 was rejected, reject H_3 if $p_3 \leq \alpha$, otherwise stop, etc

Dose-ranging clinical trial

- D3 vs D0, $p_1=0.015$
- D2 vs D0, $p_2=0.021$
- D1 vs D0, $p_3=0.044$
- All three dose-placebo comparisons are significant

Slide 70

Exercise

Allergen-induced clinical trial

- One dose of an experimental drug vs. placebo
- Patients inhale an allergen that causes bronchoconstriction
- Primary endpoint is FEV1 (forced expiratory volume in 1 sec)

Time	T statistic	P-value
15 min	0.90	0.189
30 min	1.82	0.043
45 min	1.42	0.086
1 h	2.71	0.007
2 h	4.08	0.001
3 h	3.32	0.002

When does the experimental drug show a statistically significant effect over placebo for the first time?

Slide 71

Comparison with other multiple tests

Comparison with stepwise tests (Holm, Hochberg)

- Order of null hypotheses is driven by the data (more flexibility)
- Difficult to construct associated confidence intervals

Fixed-sequence tests

- Order of null hypotheses is pre-specified (more power)
- Easy to set up confidence intervals (Hsu and Berger, 1999)
- Confidence intervals are important in non-inferiority analyses and risk-benefit assessment

Slide 72

Summary

Multiple tests based on marginal p-values

- Easy to carry out and communicate to non-statisticians
- Easy to implement using PROC MULTTEST

Closed tests

- A rich family of tests derived using the closed testing principle
- Stepwise tests with a data-driven testing order: More flexibility when multiple comparisons are difficult to order
- Fixed-sequence tests: More power when multiple comparisons are naturally ordered

Slide 73

Part III: Analysis of multiple endpoints

Multiple endpoints in clinical trials

- Difficult to fully characterize efficacy of a drug using a single endpoint
Diseases may have a complex etiology
- Numerous conditions/indications in which multiple instruments are used (e.g., patient or physician assessments)

A simultaneous analysis of all endpoints

- Global tests for examining the drug's effect on all endpoints
- Improved power to detect the overall effect
- No multiplicity adjustment (only one test is carried out)

Correlation among endpoints

- Important to account for correlation when pooling the information across the endpoints

Slide 74

Global tests

Statement of problem

- g treatment groups, n patients per group, m continuous endpoints
- X_{ijk} is a normally distributed response (ith group, jth patient, kth endpoint)
- μ_{ik} and σ_k are mean and standard deviation of X_{ijk}

Null hypothesis

- $H_0: \mu_{1k} = \dots = \mu_{gk}$ for all endpoints (all k)

Directional alternative hypothesis

- $H_1: \mu_{tk} \leq \mu_{uk}$ for all endpoints (all k) with at least one strict inequality for some u and t
- More relevant than the direction-less Hotelling T^2 test

Slide 75

O'Brien ordinary least squares (OLS) test

Dimensionality reduction

- Assume a common effect size within each group,
 $\mu_{i1}/\sigma_1 = \dots = \mu_{im}/\sigma_m = \lambda_i$

O'Brien OLS test

- Compare $\lambda_1, \dots, \lambda_g$ using a one-way analysis of variance

Two-arm studies: special case

- O'Brien OLS statistic is based on a sum of t-statistics with equal weights
 $t_{OLS} = wt_1 + \dots + wt_m$

Slide 76

Clinical trial in rheumatoid arthritis

Proof-of-concept study

- Small clinical trial to compare an experimental drug to placebo
- Four endpoints
 - number of swollen joints (SJC)
 - number of tender joints (TJC)
 - patient global assessment (PTA)
 - physician global assessment (PHA)

Endpoint	Raw	Effect size		O'Brien test	Simes test
	p-value	Drug	Placebo	p-value	p-value
SJC	0.0403	1.20	0.29		
TJC	0.0375	1.12	0.22	0.0134	0.0403
PTA	0.0239	1.03	0.20		
PHA	0.0205	1.20	0.08		

Slide 77

Inferences for individual endpoints

Global test

- No inferences for individual endpoints
- A multiple test is required to examine individual endpoints

Multiple test based on closed testing principle

- Consider all intersection hypotheses
- Apply a global test to all intersection hypotheses to compute adjusted p-values for original null hypotheses

Endpoint	Raw p-value	O'Brien test p-value	Adjusted p-value
SJC	0.0403		0.0403
TJC	0.0375	0.0134	0.0375
PTA	0.0239		0.0239
PHA	0.0205		0.0205

Slide 78

Summary

Global tests

- Global tests for examining the drug's effect on all m endpoints
- O'Brien OLS and related tests are widely used (most powerful under the assumption of equal effect sizes)
- O'Brien rank-sum test is a nonparametric version of OLS test

Inferences for individual endpoints

- Individual endpoints can be analyzed using closed tests
- Common to see non-significant results in univariate analyses following a significant global test

Software implementation

- SAS macro for implementing global tests in *Analysis of Clinical Trials Using SAS* (Chapter 2)

Slide 79

Part IV: Gatekeeping testing strategies

Primary and secondary analyses

- Product labels typically focus on primary findings
- Secondary analyses (secondary endpoints or subgroup analyses) provide much useful information to prescribing physicians, patients, hospital administrators

FDA guidance “Clinical studies section of labeling for prescription drugs and biologics”

- “The CLINICAL STUDIES section should present those endpoints that are **essential to establishing the effectiveness of the drug** (or that show the limitations of effectiveness) and those that provide **additional useful and valid information about the activities of the drug.**”

Slide 80

Gatekeeping strategies

Dilemma

- What secondary findings should be included in the product label?
- Are pharmaceutical companies guilty of cherry-picking?

Gatekeeper strategies offer one potential solution to the dilemma

- They account for the hierarchical structure of multiple analyses
- Preserve the overall Type I error rate

Clinical trial applications

- Registration trials with multiple primary and secondary endpoints
- Dose-ranging trials

Slide 81

Types of gatekeeping strategies

Serial testing

- All significant results in the gatekeeper to proceed to the next family of analyses

Parallel testing

- At least one significant result in the gatekeeper to proceed to the next family of analyses

Introduce general gatekeeping framework

- Based on the closed testing principle
- Focus on strategies derived using Bonferroni's test

Slide 82

One primary endpoint

Depression trial

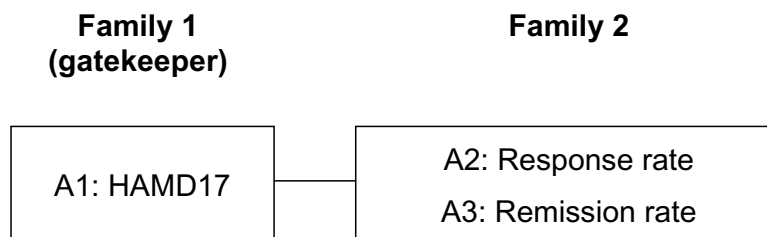
- Single primary endpoint
 - 17-item Hamilton depression rating scale (HAMD17)
 - Successful outcome if the drug is superior to placebo
- Two important secondary endpoints
 - Response rate based on HAMD17
 - Remission rate based on HAMD17

Serial gatekeeping strategy

- Propose including the secondary findings in the product label if the primary endpoint is significant

Slide 83

Serial gatekeeping strategy



Step 1: Primary analysis at α level

- No adjustment for multiplicity

Step 2: Secondary analyses **if the primary analysis yielded a significant result**

- Stepwise Holm test to adjust for multiplicity within Family 2
- No adjustment for the primary endpoint (memory-less method)

Slide 84

Serial gatekeeping strategy

Endpoint	Raw p	Adjusted p
Primary: HAMD17	0.046	0.046
Secondary: Response rate	0.048	0.048
Secondary: Remission rate	0.021	0.042

Primary and secondary endpoints are significant at 5% level

- Justification for including the secondary endpoints in the product label

Slide 85

Multiple primary endpoints

Clinical trial in patients with acute lung injury (ALI)

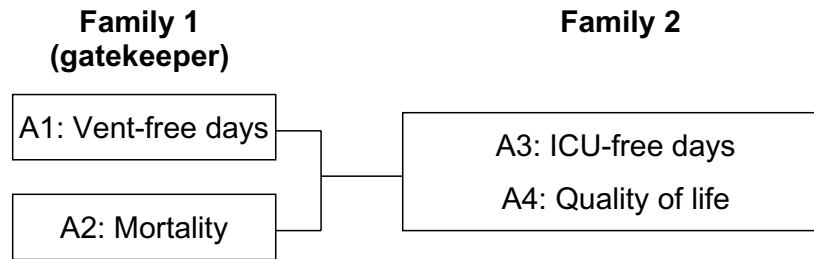
- Two primary endpoints
 - Number of days patients are off mechanical ventilation (vent-free days)
 - 28-day all-cause mortality rate
 - Successful outcome if the drug is superior to placebo with respect to **either** endpoint
- Two important secondary endpoints
 - Number of days patients are out of ICU (ICU-free days)
 - Overall quality of life at the end of the study

Parallel gatekeeping strategy

- Propose including the secondary findings in the product label provided at least one primary endpoint is significant

Slide 86

Parallel gatekeeping strategy



Step 1: Primary analysis at overall α level

- Adjustment for multiplicity within Family 1

Step 2: Secondary analyses **if at least one primary analysis** yielded a significant result

- Adjustment for multiplicity within Family 2 will depend on the number of significant primary outcomes (not memory-less anymore)

Slide 87

Parallel gatekeeping test

Closed testing principle

- Gatekeeping tests are constructed using the closed testing principle

Stepwise representation

- Family 1: Bonferroni test at overall α level
 - k is the number of significant outcomes
- Family 2: Stepwise Holm test
 - Overall significance level is $\alpha k/2$
 - No multiplicity adjustment for the primary endpoints (memory-less method) if both primary endpoints are significant ($k=2$)
 - Penalty if only one primary endpoint is significant ($k=1$)
 - Secondary analyses are not performed if the primary endpoints are not significant ($k=0$)

Slide 88

ALI clinical trial: Scenario 1

Two significant primary variables

- Significant improvement in the mean number of ventilator-free days and 28-day all-cause mortality

Endpoint	Raw p	Adjusted p
Primary: Vent-free days	0.024	0.027
Primary: Mortality	0.003	0.030
Secondary: ICU-free days	0.026	0.029
Secondary: Quality of life	0.002	0.027

All analyses are significant at 5% level

- Justification for including the secondary endpoints in the product label

Slide 89

ALI clinical trial: Scenario 2

Single significant primary variable

- Significant improvement in 28-day all-cause mortality but not in mean number of ventilator-free day

Endpoint	Raw p	Adjusted p
Primary: Vent-free days	0.084	0.093
Primary: Mortality	0.003	0.030
Secondary: ICU-free days	0.026	0.093
Secondary: Quality of life	0.002	0.040

Primary mortality analysis and secondary quality of life analysis are significant at 5% level

- Justification for including the secondary endpoints in the product label

Slide 90

Dose-ranging study

Clinical trial in patients with hypertension

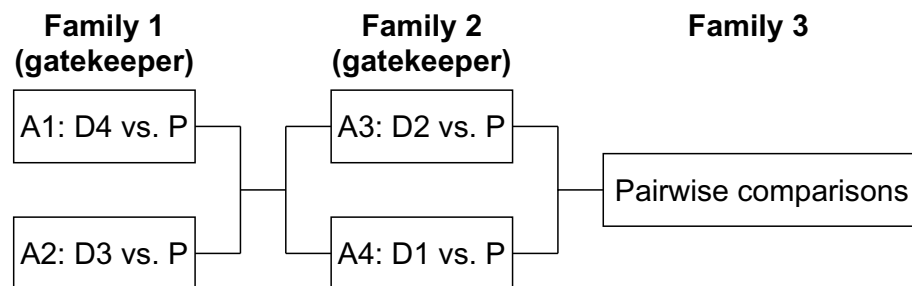
- Four doses of an experimental drug are compared to placebo
 - Doses are labeled as D1, D2, D3 and D4
- Primary endpoint
 - Reduction in diastolic blood pressure

Objectives of the study

- Find the doses with a significant reduction in diastolic blood pressure compared to placebo
- Study the shape of the dose-response curve

Slide 91

Parallel gatekeeping strategy



Step 1: Compare doses D3 and D4 to placebo

Step 2: Compare doses D1 and D2 to placebo **if at least one comparison at Step 1 is significant**

Step 3: Pairwise dose comparisons **if at least one comparison at Step 2 is significant**

Slide 92

Parallel gatekeeping strategy

Comparison	Raw p	Adjusted p		
		Gatekeeping procedure	Holm procedure	Dunnett procedure
D4 vs. P	0.0008	0.0016	0.0055	0.0030
D3 vs. P	0.0135	0.0269	0.0673	0.0459
D2 vs. P	0.0197	0.0394	0.0787	0.0656
D1 vs. P	0.7237	1.0000	1.0000	0.9899
D4 vs. D1	0.0003	0.0394	0.0021	
D4 vs. D2	0.2779	1.0000	0.8338	
D3 vs. D1	0.0054	0.0394	0.0324	
D3 vs. D2	0.8473	1.0000	1.0000	

Doses D2, D3 and D4 are significantly different from placebo at 5% level

Slide 93

Comments

Basic gatekeeping framework

- Focused on gatekeeping procedures based on Bonferroni test

More powerful gatekeeping tests

- Based on more powerful tests, e.g., Simes test
- Based on tests accounting for the correlation among the endpoints
 - Exact parametric tests such as Dunnett test and approximate resampling-based Westfall-Young tests

Software implementation

- SAS macros for performing gatekeeping inferences in *Analysis of Clinical Trials Using SAS* (Chapter 2)

Slide 94

Summary

Gatekeeping strategies can be successfully used in

- Pivotal trials with multiple primary and secondary endpoints
- Dose-ranging studies

Registration trials

- A priori designation of gatekeeping strategy allows additional data useful to physician and patient to be presented in the product label

Dose-ranging studies

- Efficient tests of dose-response relationship

Slide 95

Interim monitoring in clinical trials

Alex Dmitrienko
Eli Lilly and Company

Slide 96

Outline

Introduction to interim monitoring and group sequential trials

Repeated significance tests

Stochastic curtailment tests

Slide 97

Part I: Introduction to interim monitoring

Sequential monitoring of safety and efficacy data

- Interim monitoring has become an integral part of clinical trials
- Early stopping as soon as enough information is accumulated to reach a conclusion about the drug's properties (positive and negative)

Group sequential trials

- Interim looks are taken after groups of patients completed the study
 - Pre-specified number of patients (traditional approach)
 - Pre-specified amount of information (information-based approach)
- Focus on the traditional approach

Slide 98

Interim assessments in clinical trials

Ethical requirements

- Imperative to ensure that patients are not exposed to inferior or harmful therapies
- Stop the study as soon as
 - Experimental therapy is found to cause unacceptable side effects
 - Experimental therapy is shown to be superior to control

Example

- Interim evaluations are generally mandated in clinical trials with non-reversible outcomes (e.g., mortality trials)
 - Monthly safety assessments to avoid harmful effects

Slide 99

Interim assessments in clinical trials

Financial considerations

- Futility analyses to make the optimal use of R&D dollars
 - Early proof-of-concept studies and larger Phase II, III and IV trials
- Early stopping if the study is unlikely to achieve its objectives at the planned end

Example

- Several recent severe sepsis trials were terminated early due to lack of therapeutic benefit

Slide 100

Interim assessments in clinical trials

Underlying assumptions

- Stable patient population
- No changes in standards of care or seasonal trends
- No patient selection trends (patients with mild disease are enrolled first)

Non-statistical issues

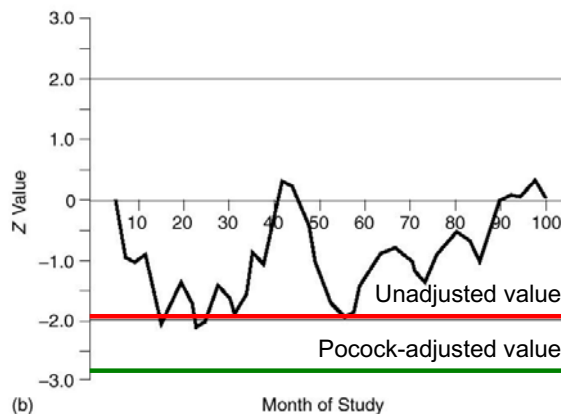
- More to decision-making at interim looks than computing a p-value
- Risk-benefit ratios
- Operational aspects (data monitoring committees, independent data analysis groups)
- Will not be discussed

Slide 101

Statistical methods

Statistical methods are required to properly interpret interim findings

- Clinical researchers must account for **multiple analyses**



Example from Ellenberg, Fleming and DeMets (2002)

Coronary Drug Project trial (clofibrate vs placebo)

Unadjusted two-sided 0.05 critical value is 1.96

Pocock-adjusted critical value is 2.76 (36 equally spaced looks)

Slide 102

Statistical methods

Statistical methods are required to properly interpret interim findings

- Eliminate bias caused by **sequential sampling**

Adjustment for sequential sampling

- Van Den Berghe et al (2001). *The New England Journal of Medicine*
- A mortality trial was stopped early due to superior efficacy
- Observed reduction in overall mortality
42% [95% CI 22% -- 62%]
- Adjusted reduction in overall mortality
32% [95% CI 2% -- 55%]
Treatment effect estimate was pulled toward 0 to account for sequential sampling

Slide 103

Statistical methods in interim monitoring

Repeated significance testing

- Designs with equally-spaced interim analyses (Pocock, 1977; O'Brien and Fleming, 1979)
- Flexible sequential monitoring (Lan and DeMets, 1983)

Boundaries approach

- Triangular and related tests (Whitehead, 1997)
- Conceptually similar to repeated significance testing
- Will not be discussed

Stochastic curtailment

- Emphasis on predictive inferences whereas repeated significance tests focus on currently available data

Slide 104

Part II: Repeated significance testing

How to design group sequential trials?

How to execute group sequential trials?

Slide 105

Design stage

Step 1: Select a design that reflects the trial's objective

- Efficacy testing
- Simultaneous efficacy and futility testing
- Note: Safety stopping rules are often difficult to quantify

Step 2: Select stopping boundaries

- Pocock, O'Brien and Fleming, Wang-Tsiatis boundaries

Step 3: Set up a group sequential plan

- Compute stopping boundaries and operating characteristics (stopping probabilities, expected sample size)

Slide 106

Notation

Two-arm clinical trial

- m data analyses (including final analysis)
- Z_1, \dots, Z_m are test statistics (jointly normally distributed)
- δ is the true treatment difference

Hypothesis testing problem

- Null hypothesis H_0 : $\delta=0$
- One-sided alternative H_1 : $\delta>0$

Error probabilities

- Type I error rate is α and Type II error rate is β

Slide 107

Step 1: Stopping boundaries

Efficacy testing

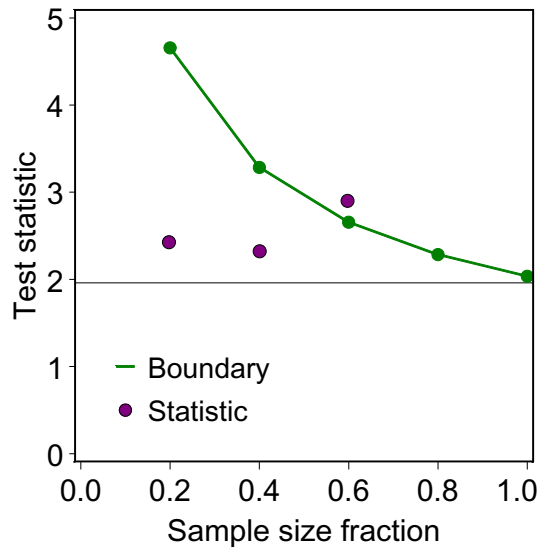
- Z_k is compared to the upper critical value $u(\alpha)_k$
- $Z_k \geq u(\alpha)_k$: stop early due to superior efficacy
- $Z_k < u(\alpha)_k$: continue to next look

Simultaneous efficacy and futility testing

- Z_k is compared to the lower critical value $l(\alpha, \beta)_k$ and upper critical value $u(\alpha, \beta)_k$
- $Z_k \geq u(\alpha, \beta)_k$: stop early due to superior efficacy
- $l(\alpha, \beta)_k < Z_k < u(\alpha, \beta)_k$: continue to next look
- $Z_k \leq l(\alpha, \beta)_k$: stop early due to futility

Slide 108

Step 1: Detecting superior efficacy

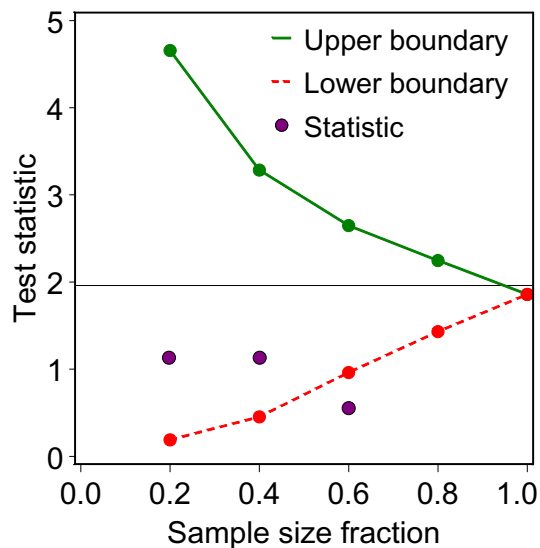


A clinical trial with five equally-spaced looks

Stop at the third interim analysis due to superior efficacy

Slide 109

Step 1: Simultaneous efficacy/futility testing



A clinical trial with five equally-spaced looks

Stop at the third interim analysis due to lack of therapeutic benefit (futility)

Lower (futility) boundary pulls the upper (efficacy) boundary down

Slide 110

Step 2: Stopping boundaries

Wang-Tsiatis family

- Family of sequential plans indexed by the ρ parameter (0 -- 0.5)

Efficacy testing

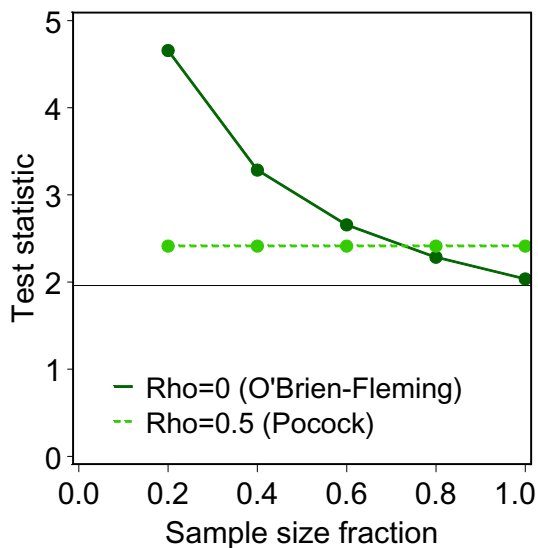
- Upper boundary $u(\alpha)_k = ck^{\rho-0.5}$
c is chosen to ensure Type I error rate is α

Simultaneous efficacy and futility testing

- Upper boundary $u(\alpha, \beta)_k = c_1 k^{\rho-0.5}$
- Lower boundary $l(\alpha, \beta)_k = \eta k^{0.5} - c_2 k^{\rho-0.5}$
 η is chosen to ensure the boundaries meet at the last analysis
 c_1 and c_2 are chosen to ensure Type I and II error rates are α and β

Slide 111

Step 2: Stopping boundaries



O'Brien-Fleming boundary

- Earlier analyses are performed in a conservative manner
- Later tests are carried out at significance levels close to the nominal level

Pocock boundary

- Same critical value at each interim look

Slide 112

How to choose stopping boundaries

Clinical considerations

- Decision making is a balancing act
 - Delay early termination unless the treatment effect is big
 - Very early stopping complicates safety/secondary efficacy analyses
- O'Brien-Fleming boundary is a most commonly used boundary

Data management/data quality considerations

- First interim analysis: Test the process
 - Test data management, check compliance, ensure data quality
 - Early stopping is undesirable
- Subsequent interim analyses: Detect early evidence of efficacy
 - Early stopping is desirable
- O'Brien-Fleming boundary fits this consideration well

Slide 113

Step 3: Set up a group sequential plan

Compute stopping boundaries

- Test statistic, p-value or treatment difference scales
 - Statisticians often favor the test statistic scale
 - Physicians prefer p-values or treatment difference (this requires standard deviation)

Compute maximum and average number of patients

- Average number of patients under H_0 and H_1

Compute stopping probabilities

- At each interim analysis under H_0 and H_1

Compute power function

- Power as a function of the true treatment difference

Slide 114

Clinical trial example

Trial in patients with severe sepsis

- 28-day all-cause mortality
- 80% power and 0.025 Type I error rate (one-sided)

Two interim analyses and final analysis

- Interim looks at 20% and 66%

First interim analysis

- Futility analysis
- Efficacy stopping only if a highly beneficial treatment effect

Second interim analysis

- Examine both efficacy and futility

Slide 115

Trial in patients with severe sepsis

Step 1: Simultaneous efficacy and futility testing

Step 2: O'Brien-Fleming boundary for efficacy testing

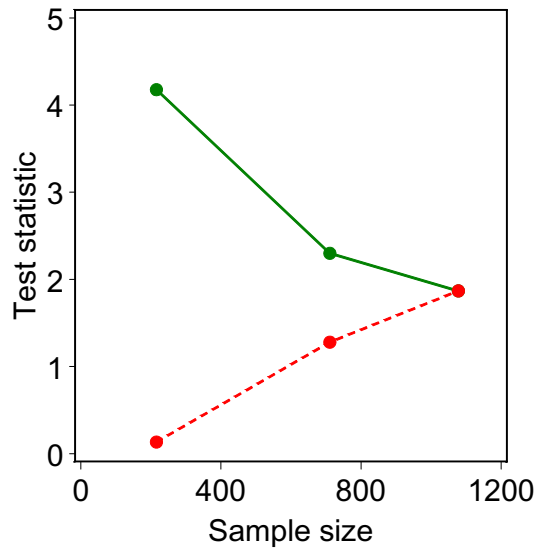
- Difficult to exceed O'Brien-Fleming boundary early in the trial
Efficacy stopping is unlikely at the first look
- Higher power at the second interim look and final analysis
Compared to Pocock boundary

Step 2: Pocock boundary for futility analyses

- Futility rule will be sensitive to small negative treatment difference
- Patients should not be exposed to harmful treatments

Slide 116

Step 3: Set up a group sequential plan



Fixed-sample design

- 859 patients per group

Group sequential design

- Maximum 1078 patients per group
- On average, 464 patients per group under H_0
- On average, 753 patients per group under H_1

Two interim looks

- After 216 and 711 patients in each group

Slide 117

Step 3: Set up a group sequential plan

Stopping probabilities at each analysis

Analysis	Under H_0	Under H_1
Interim 1	55.3%	10.5%
Interim 2	37.4%	64.0%
Final	7.4%	25.5%

Futility characteristics

- 92.7% chance of early stopping if the drug does not improve survival

Efficacy characteristics

- 74.5% chance of early stopping if the drug is efficacious

Slide 118

Execution stage

Compute adjusted critical values

Approach 1 (little flexibility)

- Time points at which the data will be reviewed are set in stone
- Pocock and O'Brien-Fleming boundaries were proposed for this type of inspection plans

Approach 2 (very flexible)

- Pre-specify an error spending function
- Choose the timing and frequency of interim looks (as long as they are not data dependent)
- Introduced by Lan and DeMets (1983)

Slide 119

Error spending approach

Type I error spending function

- Non-decreasing function defined over $0 \leq t \leq 1$ with $\alpha(0)=0$ and $\alpha(1)=\alpha$
t is the sample size fraction

Computation of critical values

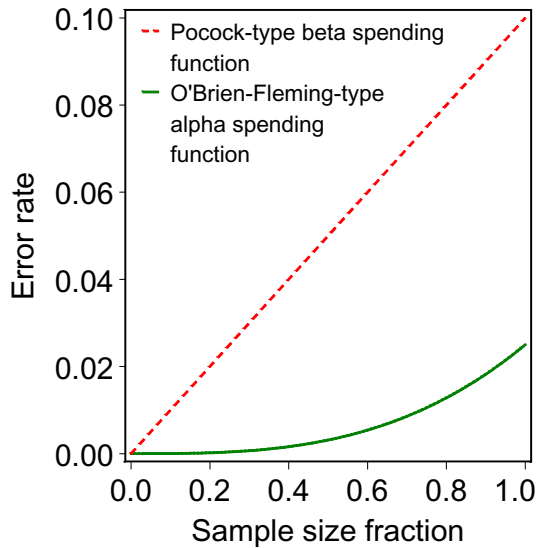
- First interim look at t_1
Compute u_1 from $P(Z_1 > u_1 | H_0) = \alpha(t_1)$
- Second interim look at t_2
Compute u_2 from $P(Z_1 \leq u_1, Z_2 > u_2 | H_0) = \alpha(t_2) - \alpha(t_1)$, etc

Examples

- Approximation to O'Brien-Fleming boundary: $\alpha(t) = \alpha t^3$
- Approximation to Pocock boundary: $\alpha(t) = \alpha t$
- Many other approximations

Slide 120

α - and β -spending functions



α -spending function

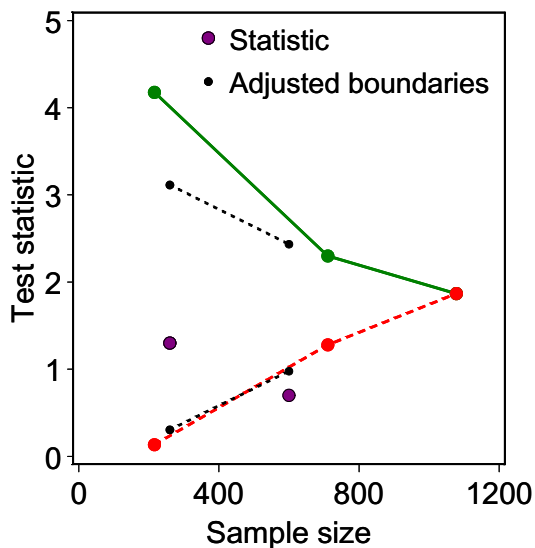
- Rate at which Type I error probability is spent
- Used in computing the **upper** critical value (efficacy testing)

β -spending function

- Rate at which Type II error probability is spent
- Used in computing the **lower** critical value (futility testing)

Slide 121

Trial in patients with severe sepsis



Planned interim looks

- Look 1 after 216 patients in each group
- Look 2 after 711 patients in each group

Actual interim looks

- Look 1 after 260 patients in each group, $Z=1.301$
- Look 2 after 600 patients in each group, $Z=0.707$

Slide 122

Summary

Repeated significance tests

- Trials for detecting superior efficacy or superior efficacy/futility
- O'Brien-Fleming boundary for superior efficacy testing and Pocock boundary for futility testing are commonly used
- Implement error spending functions to achieve flexibility

Software implementation

- SAS macros for designing and executing group sequential trials in *Analysis of Clinical Trials Using SAS* (Chapter 4)

Slide 123

Part III: Stochastic curtailment tests

Stochastic curtailment is based on prediction

- Predict the distribution of the treatment effect at the end of the trial
- Compute the probability of observing a statistically significant treatment difference

Three types of stochastic curtailment tests

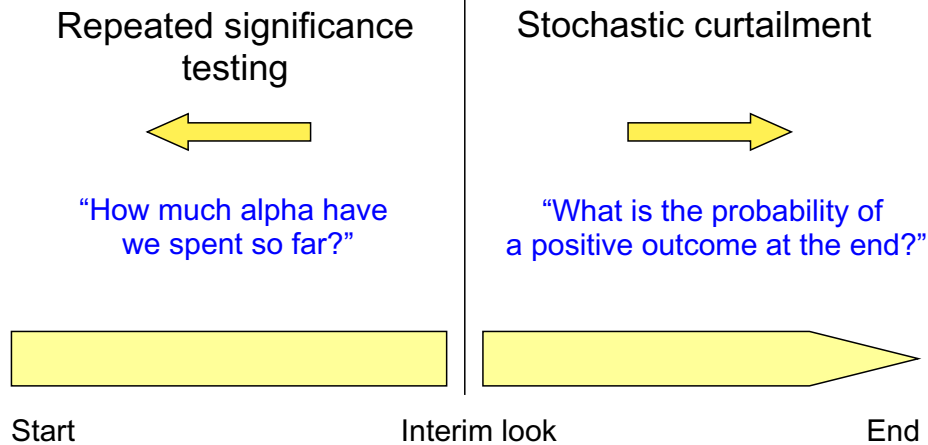
- Frequentist approach (Conditional power)
- Mixed Bayesian-frequentist approach (Predictive power)
- Bayesian approach (Predictive probability)
Will not be discussed

Futility testing versus efficacy testing

- Stochastic curtailment tests are often used in futility rules

Slide 124

Comparison of statistical methods



Slide 125

Conditional power method

Power calculation

- Run before the study start
- Become less relevant once the data are observed

Conditional power

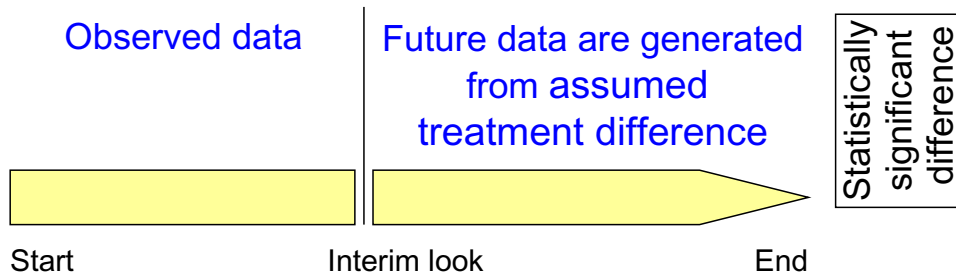
- Observed test statistic (Z_n) and test statistic at the planned end (Z_N)
- Likelihood of a statistically significant result given the interim data

Conditional power function $CP_n(\delta) = P(Z_N \geq z_\alpha | Z_n, \delta)$

- δ is assumed treatment difference
- Examples: treatment difference under the alternative ($\delta = \delta_1$) or observed treatment difference

Slide 126

Conditional power method



Futility stopping rule

- Stop early if $CP_n(\delta)$ is low (0.2 or lower)

Efficacy stopping rules are unlikely to be accepted in registration trials

Slide 127

Properties of conditional power method

Probability of early stopping

- Conditional power is influenced by assumed treatment difference (δ)
- More likely to stop early due to futility if δ is small

Futility testing under alternative hypothesis ($\delta=\delta_1$)

- Generally optimistic about the final outcome in futility testing
- Tends to delay termination if the drug is not efficacious

Adaptive conditional power method ($\delta=d$)

- Future data are consistent with interim data rather than alternative hypothesis
 - More reasonable in futility testing compared to $\delta=\delta_1$
-

Slide 128

Clinical trial example

Trial in patients with severe sepsis

- Single dose of a drug versus placebo
- Primary endpoint is 28-day all-cause mortality

Survival data were monitored on a monthly basis

- Survival rates in the experimental group were consistently low compared to placebo

Slide 129

Futility test in a severe sepsis trial

Regular and adaptive conditional power futility rules

- One-sided 0.1 significance level

Look	Survival rate		Conditional power method	
	Exp drug	Placebo	Regular	Adaptive
1	60.0%	64.4%	55.5%	22.8%
2	64.6%	66.2%	47.4%	23.5%
3	64.4%	65.3%	37.8%	17.5%
4	64.1%	67.0%	16.4%	2.8%
5	64.7%	65.7%	14.3%	4.3%
6	63.9%	65.6%	3.5%	0.6%

Adaptive futility rule appears more attractive

- Patients should not be exposed to harmful treatments

Slide 130

Predictive power method

Conditional power function $CP_n(\delta) = P(Z_N \geq z_\alpha | Z_n, \delta)$

- Depends heavily on the assumed treatment difference

Mixed Bayesian-frequentist method

- Average conditional power function

Predictive power function $PP_n = \int CP_n(\delta) f(\delta | Z_n) d\delta$

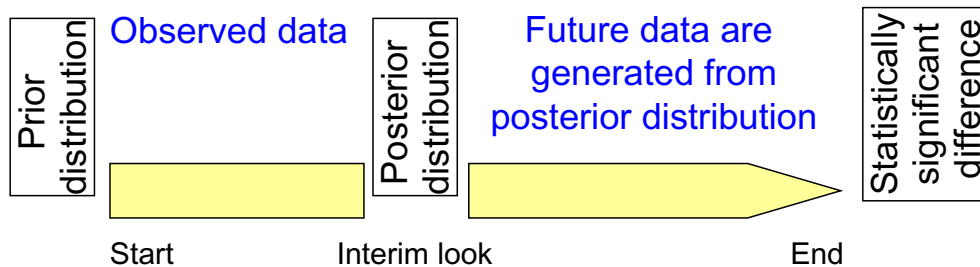
- $f(\delta | Z_n)$ is posterior density of treatment difference δ given the interim data

Futility stopping rule

- Stop early if PP_n is low (0.2 or lower)

Slide 131

Predictive power method



Prior distributions

- Uniform priors are advocated in clinical trial literature
- Choice of prior distributions is driven by objective of interim monitoring, development phase and patient population

Slide 132

Conditional and predictive power

Comparison of conditional and predictive power approaches

- At each look compute conditional and predictive power
- Statistical significance is defined at a one-sided 0.1 level

Prior distributions

- Uniform prior
- Strong aggressive prior
 - 79% survival rate in experimental group
 - 70% survival rate in placebo group
 - Small variance (CV=0.1)

Slide 133

Conditional and predictive power

Predictive and conditional power utility rules

– Severe sepsis trial

Look	Predictive power		Conditional power method	
	Uniform prior	Aggressive prior	Regular	Adaptive
1	11.1%	19.9%	55.5%	22.8%
2	10.9%	18.7%	47.4%	23.5%
3	8.6%	12.6%	37.8%	17.5%
4	1.8%	3.4%	16.4%	2.8%
5	2.4%	3.5%	14.3%	4.3%
6	0.4%	0.5%	3.5%	0.6%

Predictive power method is more sensitive to negative treatment differences than conditional power method

Slide 134

Summary

Stochastic curtailment

- Stochastic curtailment methods are based on prediction
- Predict the treatment effect at the end of the trial given interim data
- Commonly used in futility testing
 - Also as an efficacy/futility diagnostic tool in a non-registration setting or for internal decision-making

Three popular stochastic curtailment methods

- Frequentist, mixed Bayesian-frequentist and fully Bayesian (not discussed)
- Adaptive conditional power method is more sensitive to observed data than regular conditional power method
- Predictive power method incorporates prior information

Handling Incomplete Data in Longitudinal Studies

For: Analysis of Clinical Trials

July 30, 2007

Geert Molenberghs

**Universiteit
Hasselt**

Handling Incomplete Data in Longitudinal Studies – p. 136/178



Growth Data

- Taken from Potthoff and Roy, *Biometrika* (1964)
- Research question:

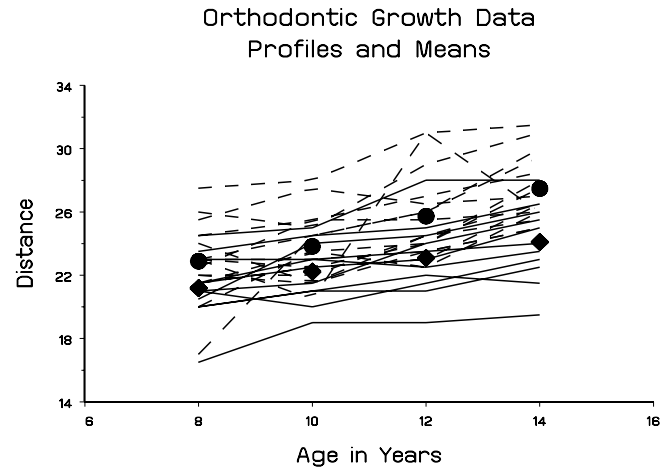
Is dental growth related to
gender ?

- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 11 girls and 16 boys

Handling Incomplete Data in Longitudinal Studies – p. 137/178

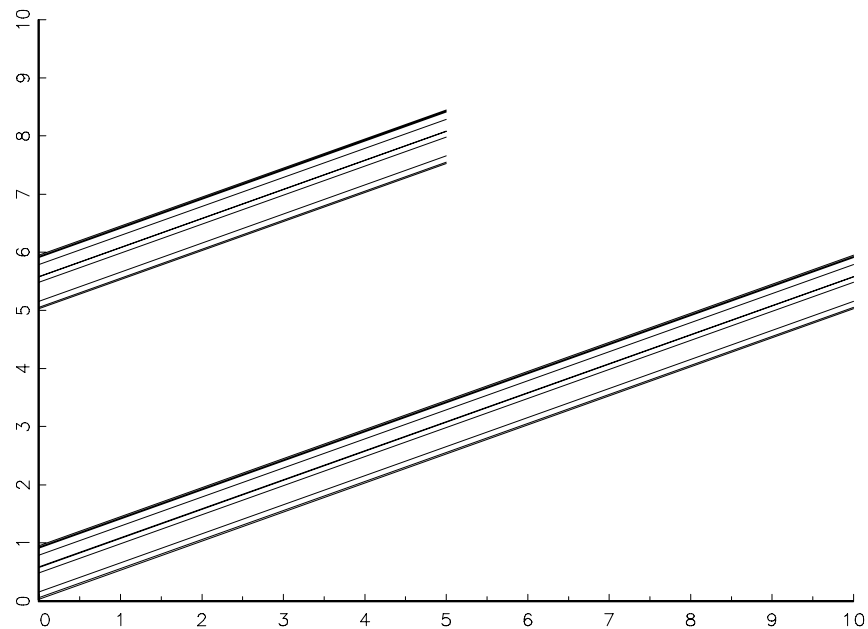
Growth Data: Individual Profiles

- Much variability between girls / boys
- Considerable variability within girls / boys
- Fixed number of measurements per subject
- Measurements taken at fixed time points



Handling Incomplete Data in Longitudinal Studies – p. 138/178

Incomplete Longitudinal Data



Handling Incomplete Data in Longitudinal Studies – p. 139/178

Scientific Question

View 1 In terms of **entire longitudinal profile**

View 2 In terms of **last *planned* measurement**

View 3 In terms of **last *observed* measurement**

Notation

- Subject i at occasion (time) $j = 1, \dots, n_i$
- **Measurement** Y_{ij}
- **Missingness indicator** $R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$
- Group Y_{ij} into a vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})' = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$

$$\begin{cases} \mathbf{Y}_i^o & \text{contains } Y_{ij} \text{ for which } R_{ij} = 1, \\ \mathbf{Y}_i^m & \text{contains } Y_{ij} \text{ for which } R_{ij} = 0. \end{cases}$$
- Group R_{ij} into a vector $\mathbf{R}_i = (R_{i1}, \dots, R_{in_i})'$
- D_i : time of dropout: $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$

Framework

$$f(Y_i, D_i | X_i, \theta, \psi)$$

- **Selection Models:** $f(Y_i | X_i, \theta) f(D_i | X_i, Y_i^o, Y_i^m, \psi)$

MCAR

→

MAR

→

MNAR

CC ?

direct likelihood !

joint model !?

LOCF ?

expectation-maximization (EM)

sensitivity analysis

imputation ?

multiple imputation (MI) !

⋮

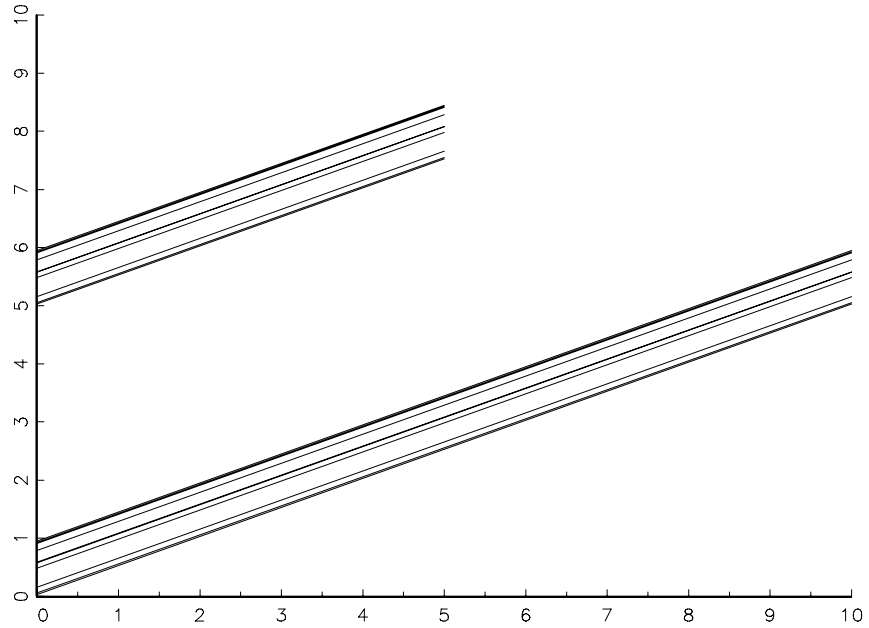
(weighted) GEE !

- **Pattern-Mixture Models:** $f(Y_i | X_i, D_i, \theta) f(D_i | X_i, \psi)$
- **Shared-Parameter Models:** $f(Y_i | X_i, b_i, \theta) f(D_i | X_i, b_i, \psi)$

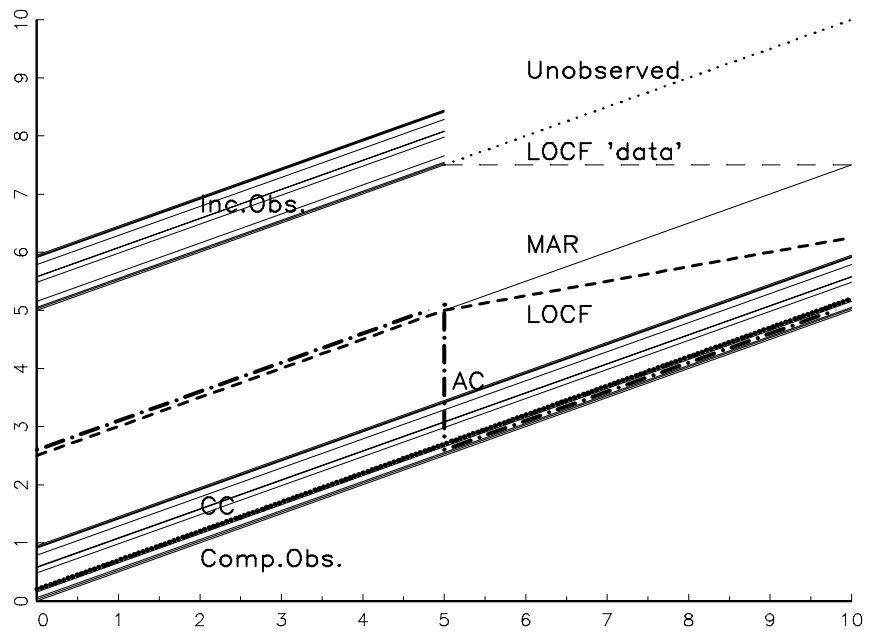
Overview

- Simple methods
- Bias for LOCF and CC
- Direct likelihood inference
- Multiple Imputation
- Weighted generalized estimating equations
- Sensitivity analysis

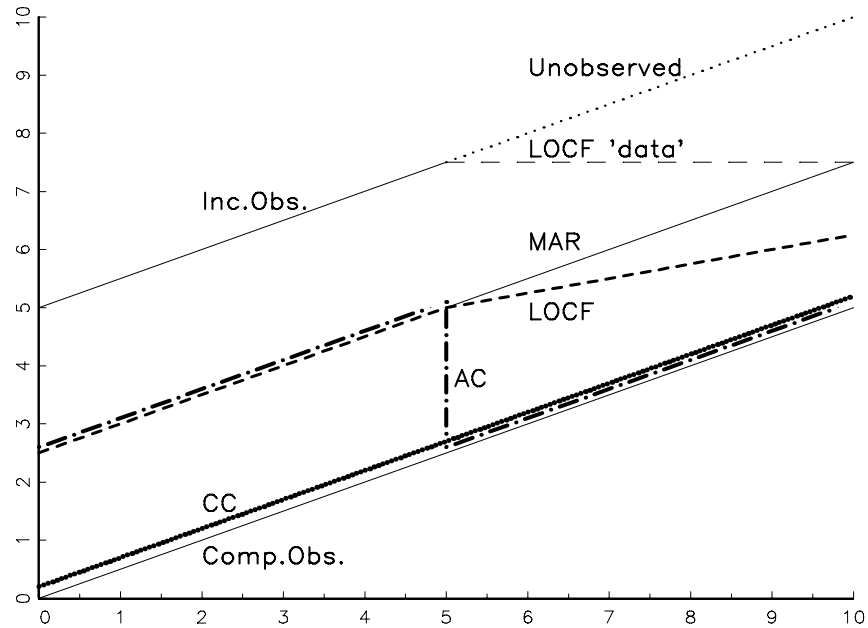
Incomplete Longitudinal Data



Data and Modeling Strategies



Modeling Strategies



Handling Incomplete Data in Longitudinal Studies – p. 146/178

Simple Methods

MCAR

- Rectangular matrix by **deletion**: *complete case analysis*
- Rectangular matrix by **completion** or imputation
 - *Vertical*: Unconditional mean imputation
 - *Horizontal*: Last observation carried forward
 - *Diagonal*: Conditional mean imputation
- Using data **as is**: *available case analysis*
 - Likelihood-based MAR analysis: simple and correct

Handling Incomplete Data in Longitudinal Studies – p. 147/178

Quantifying the Bias

<p style="text-align: center;">Dropouts $t_{ij} = 0$</p> <p style="text-align: center;">Probability p_0</p> <p style="text-align: center;">$E(Y_{ij}) =$ $\beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij}$</p>	<p style="text-align: center;">Completers $t_{ij} = 0, 1$</p> <p style="text-align: center;">Probability p_1</p> <p style="text-align: center;">$E(Y_{ij}) =$ $\gamma_0 + \gamma_1 T_i + \gamma_2 t_{ij} + \gamma_3 T_i t_{ij}$</p>
---	--



	CC	LOCF
MCAR	0	$(p_1 - p_0)\beta_2 - (1 - p_1)\beta_3$
MAR	$-\sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]$	$p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1) - p_0(\gamma_0 + \gamma_2) - (1 - p_0)\beta_0 - \gamma_1 - \gamma_3 - \sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]$

$T_i = 0, 1$: treatment assignment

$\sigma \propto$ covariance between both measurements

Direct Likelihood Maximization

MAR : $f(\mathbf{Y}_i^o | X_i, \theta) f(D_i | X_i, \mathbf{Y}_i^o, \psi)$

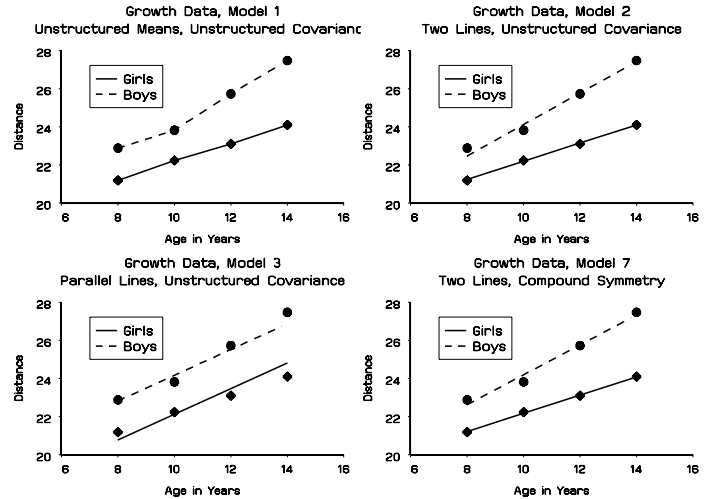
Mechanism is MAR
 θ and ψ distinct
 Interest in θ
 Use observed information matrix

} \implies Lik. inference valid

Outcome type	Modeling strategy	Software
Gaussian	Linear mixed model	MIXED
Non-Gaussian	Gen./Non-linear mixed model	NLMIXED, GLIMMIX

Original, Complete Orthodontic Growth Data

	Mean	Covar	# par
1	unstructured	unstructured	18
2	\neq slopes	unstructured	14
3	$=$ slopes	unstructured	13
7	\neq slopes	CS	6



Handling Incomplete Data in Longitudinal Studies – p. 150/178

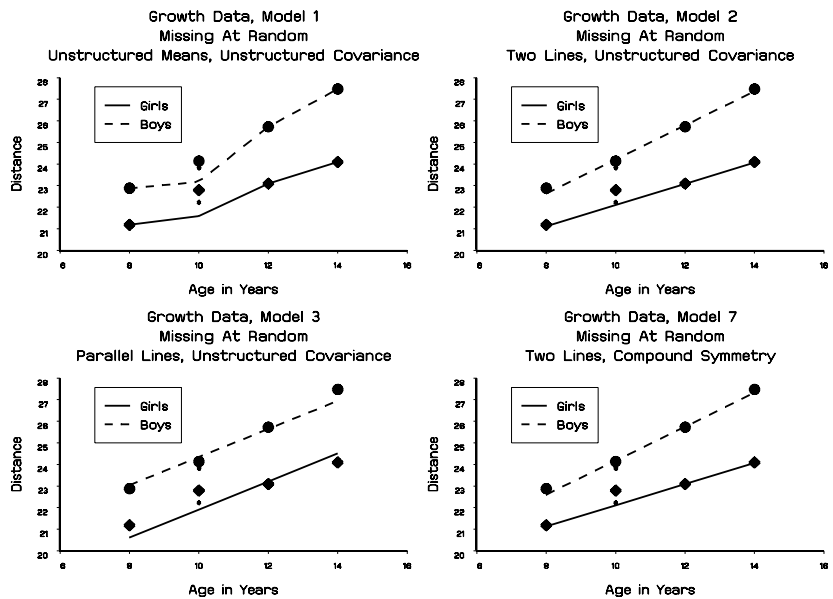
Trimmed Growth Data: Simple Methods

Method	Model	Mean	Covar	# par
CC	7a	$=$ slopes	CS	6
LOCF	2a	quadratic	unstructured	16
Uncond. mean	7a	$=$ slopes	CS	6
Cond. mean	1	unstructured	unstructured	18

distorting

Trimmed Growth Data: Direct Likelihood

Mean	Covar	# par
7 \neq slopes	CS	6



Handling Incomplete Data in Longitudinal Studies – p. 152/178

Comparison of Analyses

Principle	Method	Boys at Age 8	Boys at Age 10
Original	Direct likelihood, ML	22.88 (0.56)	23.81 (0.49)
	Direct likelihood, REML \equiv MANOVA	22.88 (0.58)	23.81 (0.51)
	ANOVA per time point	22.88 (0.61)	23.81 (0.53)
Direct Lik.	Direct likelihood, ML	22.88 (0.56)	23.17 (0.68)
	Direct likelihood, REML	22.88 (0.58)	23.17 (0.71)
	MANOVA	24.00 (0.48)	24.14 (0.66)
	ANOVA per time point	22.88 (0.61)	24.14 (0.74)
CC	Direct likelihood, ML	24.00 (0.45)	24.14 (0.62)
	Direct likelihood, REML \equiv MANOVA	24.00 (0.48)	24.14 (0.66)
	ANOVA per time point	24.00 (0.51)	24.14 (0.74)
LOCF	Direct likelihood, ML	22.88 (0.56)	22.97 (0.65)
	Direct likelihood, REML \equiv MANOVA	22.88 (0.58)	22.97 (0.68)
	ANOVA per time point	22.88 (0.61)	22.97 (0.72)

Handling Incomplete Data in Longitudinal Studies – p. 153/178

Behind the Scenes

- R completers $\leftrightarrow N - R$ “incompleters”

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{pmatrix} \right)$$

- Conditional density

$$Y_{i2}|y_{i1} \sim N(\beta_0 + \beta_1 y_{i1}, \sigma_{22.1})$$

Frequentist versus Likelihood

μ_1	freq. & lik.	$\hat{\mu}_1 = \frac{1}{N} \sum_{i=1}^N y_{i1}$
μ_2	frequentist	$\tilde{\mu}_2 = \frac{1}{R} \sum_{i=1}^R y_{i2}$
μ_2	likelihood	$\hat{\mu}_2 = \frac{1}{N} \left\{ \sum_{i=1}^R y_{i2} + \sum_{i=R+1}^N \left[\bar{y}_2 + \hat{\beta}_1 (y_{i1} - \bar{y}_1) \right] \right\}$

(Weighted) Generalized Estimating Equations

MAR and non-ignorable !

- Standard GEE inference correct only under MCAR

-

MAR: *weighted* GEE

Robins, Rotnitzky, Zhao (JASA 1995)

Fitzmaurice, Molenberghs, Lipsitz (JRSSB 1995)

- Weigh a contribution by inverse dropout probability
- Adjust estimating equations

Handling Incomplete Data in Longitudinal Studies – p. 156/178

Multiple Imputation

MAR

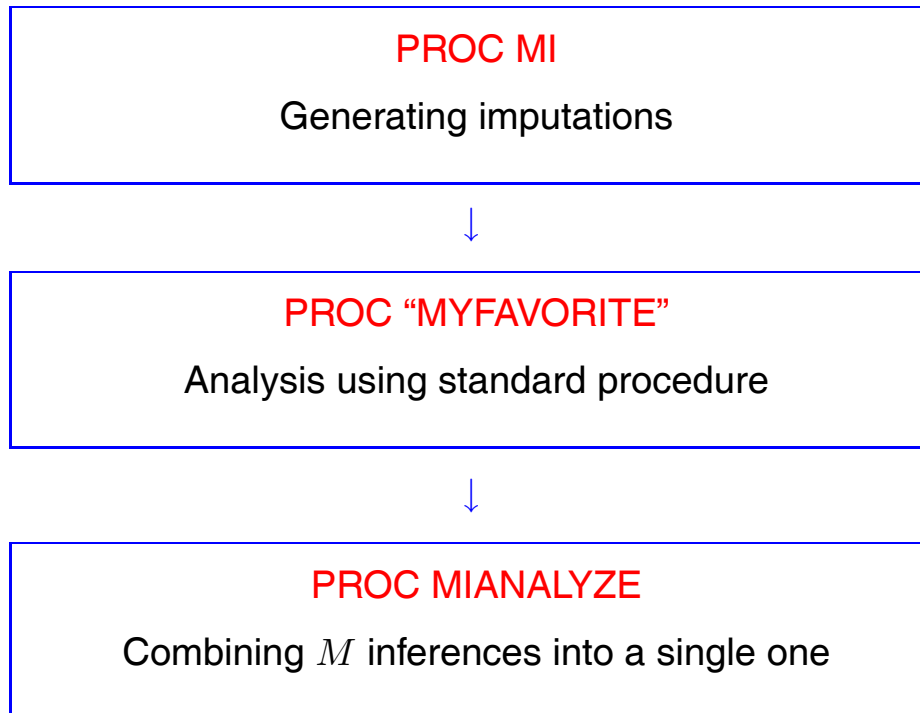
- Draw M times from θ and $f(Y_i^m | Y_i^o, \theta)$
- Analyze each of the so-completed datasets
- Combine M inferences into a single one:
 - ▷ Within-imputation variability
 - ▷ Between-imputation variability

Rubin (1987)

Rubin and Schenker (1987)

Handling Incomplete Data in Longitudinal Studies – p. 157/178

SAS for Multiple Imputation



Handling Incomplete Data in Longitudinal Studies – p. 158/178

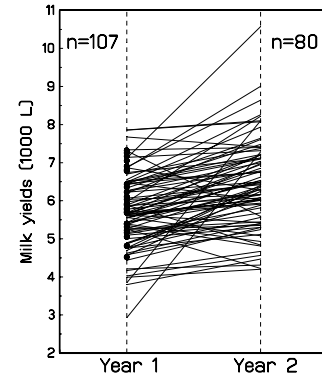
Use of Multiple Imputation

- Incomplete longitudinal outcomes: **direct-likelihood** simplest choice
- **MI**: missing covariates along with missing outcomes
- **MI**: many analyses on the same incomplete set of data
- **MI-GEE** as alternative to **WGEE**

Handling Incomplete Data in Longitudinal Studies – p. 159/178

Mastitis in Dairy Cattle

- Infectious disease of the udder
- Leads to a reduction in milk yield
- High yielding cows more susceptible?
- **But** this cannot be measured directly because of the effect of the disease: *evidence is missing*



Diggle and Kenward (JRSCC 1994)

Handling Incomplete Data in Longitudinal Studies – p. 160/178

A Full Selection Model

$$\boxed{\text{MNAR}} : \int f(\mathbf{Y}_i | \boldsymbol{\theta}) f(D_i | \mathbf{Y}_i, \boldsymbol{\psi}) d\mathbf{Y}_i^m$$

Diggle and Kenward (JRSCC 1994)

$$\boxed{f(\mathbf{Y}_i | \boldsymbol{\theta})}$$

Linear mixed model

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i$$

&

$$\boxed{f(D_i | \mathbf{Y}_i, \boldsymbol{\psi})}$$

Logistic regressions for dropout

$$\text{logit} [P(D_i = j | D_i \geq j, Y_{i,j-1}, Y_{ij})]$$

$$= \psi_0 + \psi_1 Y_{i,j-1} + \psi_2 Y_{ij}$$

A Version for the Mastitis Data

- **Model for milk yield:**

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu \\ \mu + \Delta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \right]$$

- **Model for mastitis:**

$$\begin{aligned} \text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] &= \psi_0 + \psi_1 Y_{i1} + \psi_2 Y_{i2} \\ &= 0.37 + 2.25 Y_{i1} - 2.54 Y_{i2} \\ &= 0.37 - 0.29 Y_{i1} - 2.54 (Y_{i2} - Y_{i1}) \end{aligned}$$

Criticism → Sensitivity Analysis

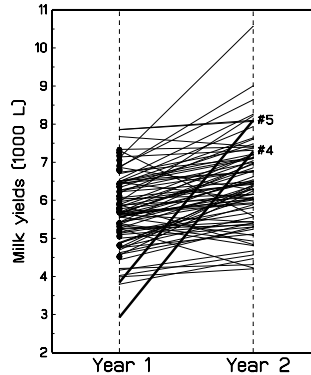
“... , estimating the ‘unestimable’ can be accomplished only by making modelling **assumptions**,.... The consequences of model **misspecification** will (...) be more severe in the non-random case.” (Laird 1994)



- **Several plausible models** or **ranges of inferences**
- **Change distributional assumptions** (Kenward 1998)
- **Local and global influence methods**
- **Semi-parametric framework** (Scharfstein *et al* 1999)
- **Pattern-mixture models**

Kenward's Sensitivity Analysis

- Deletion of #4 and #5 $\Rightarrow G^2$ for ψ_2 : 5.11 \longrightarrow 0.08



- Cows #4 and #5 have unusually large increments
- Kenward conjectures: #4 and #5 ill during the first year

Kenward (SiM 1998)

Local Influence

Verbeke, Thijs, Lesaffre, Kenward (Bcs 2001)

- Perturbed MAR dropout model:

$$\text{logit} [P(D_i = 1 | Y_{i1}, Y_{i2})]$$

$$= \psi_0 + \psi_1 Y_{i1} + \omega_i Y_{i2}$$

- Likelihood displacement:

$$LD(\omega) = 2 \left[L(\hat{\theta}, \hat{\psi}) - L(\hat{\theta}_\omega, \hat{\psi}_\omega) \right] \geq 0$$

Local Influence

Verbeke, Thijs, Lesaffre, Kenward (Bcs 2001)

- Perturbed MAR dropout model:

$$\text{logit} [P(D_i = 1 | Y_{i1}, Y_{i2})]$$

$$= \psi_0 + \psi_1 Y_{i1} + \omega_i Y_{i2}$$

$$\text{or } \psi_0 + \psi_1 Y_{i1} + \omega_i (Y_{i2} - Y_{i1})$$

- Likelihood displacement:

$$LD(\omega) = 2 \left[L(\hat{\theta}, \hat{\psi}) - L(\hat{\theta}_\omega, \hat{\psi}_\omega) \right] \geq 0$$

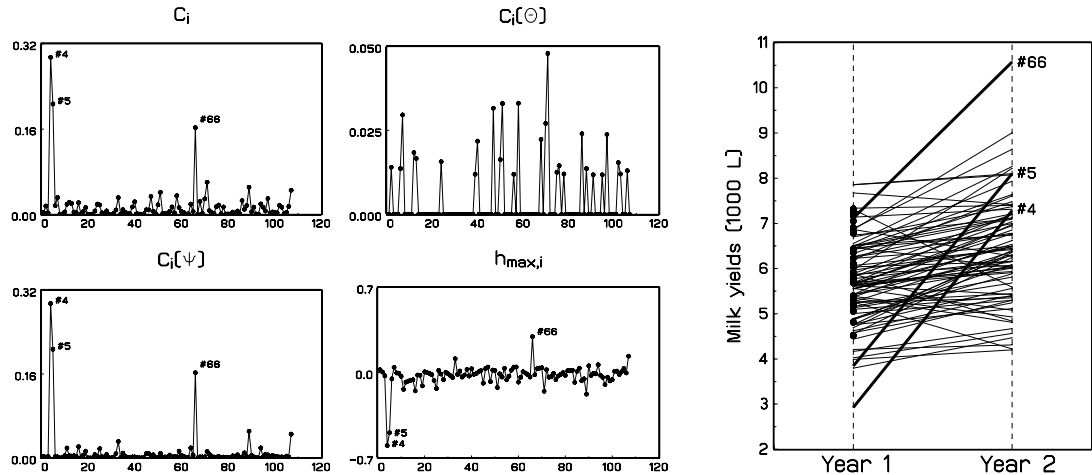
Handling Incomplete Data in Longitudinal Studies – p. 166/178

Computational Approach

- **Continuous outcomes:**
 - **Fit MAR model:**
 - linear mixed model for outcomes
 - logistic regression for dropout
 - **Evaluate closed-form expressions for local influence**

Handling Incomplete Data in Longitudinal Studies – p. 167/178

Application to Mastitis Data



- Removing #4, #5 and #66 $\Rightarrow G^2 = 0.005$
- The components in h_{\max} highlight the same cows, but different signs for (#4,#5) and #66

Handling Incomplete Data in Longitudinal Studies – p. 168/178

The Slovenian Plebiscite

Rubin, Stern, and Vehovar (JASA 1995)

- Slovenian Public Opinion (SPO) Survey
- Four weeks prior to decisive plebiscite
- Three questions:
 1. Are you in favor of Slovenian independence ?
 2. Are you in favor of Slovenia's secession from Yugoslavia ?
 3. Will you attend the plebiscite ?
- **Political decision: ABSENCE \equiv NO**
- Primary Estimand:

θ : Proportion in favor of independence

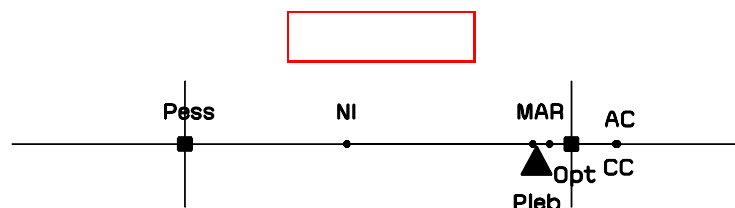
Handling Incomplete Data in Longitudinal Studies – p. 169/178

Slovenian Public Opinion Survey

		<i>Independence</i>		
(Secession)	<i>Attendance</i>	Yes	No	*
Yes	Yes	1191	8	21
	No	8	0	4
	*	107	3	9
No	Yes	158	68	29
	No	7	14	3
	*	18	43	31
*	Yes	90	2	109
	No	1	2	25
	*	19	8	96

Handling Incomplete Data in Longitudinal Studies – p. 170/178

Slovenian Public Opinion Survey

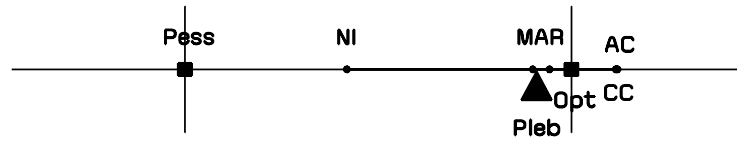


Estimator	$\hat{\theta}$
Pessimistic bound	0.694
Optimistic bound	0.904
Complete cases	0.928 ?
Available cases	0.929 ?
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782

Handling Incomplete Data in Longitudinal Studies – p. 171/178

Slovenian Plebiscite: The Truth ?

$$\theta = 0.885$$



Estimator	$\hat{\theta}$
Pessimistic bound	0.694
Optimistic bound	0.904
Complete cases	0.928 ?
Available cases	0.929 ?
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782

Handling Incomplete Data in Longitudinal Studies – p. 172/178

Did “the” MNAR model behave badly ?

Consider a family of MNAR models

Baker, Rosenberger, and DerSimonian (SiM 1992)

$$E(Y_{11jk}) = m_{jk},$$

$$E(Y_{10jk}) = m_{jk}\beta_{jk}, \quad \text{attendance}$$

$$E(Y_{01jk}) = m_{jk}\alpha_{jk}, \quad \text{independence}$$

$$E(Y_{00jk}) = m_{jk}\alpha_{jk}\beta_{jk}\gamma,$$

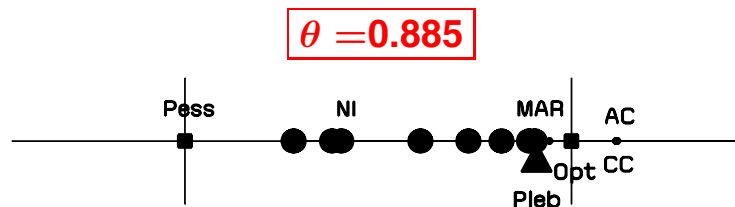
Handling Incomplete Data in Longitudinal Studies – p. 173/178

Identifiable Models

Model	Structure	d.f.	loglik	θ	C.I.
BRD1	(α, β)	6	-2503.06	0.891	[0.877;0.906]
BRD2	(α, β_j)	7	-2476.38	0.884	[0.868;0.899]
BRD3	(α_k, β)	7	-2471.59	0.881	[0.865;0.896]
BRD4	(α, β_k)	7	-2476.38	0.779	[0.702;0.857]
BRD5	(α_j, β)	7	-2471.59	0.848	[0.814;0.882]
BRD6	(α_j, β_j)	8	-2440.67	0.822	[0.792;0.850]
BRD7	(α_k, β_k)	8	-2440.67	0.774	[0.719;0.828]
BRD8	(α_j, β_k)	8	-2440.67	0.753	[0.691;0.815]
BRD9	(α_k, β_j)	8	-2440.67	0.866	[0.849;0.884]

Handling Incomplete Data in Longitudinal Studies – p. 174/178

An “Interval” of MNAR Estimates



Estimator	$\hat{\theta}$
[Pessimistic; optimistic]	[0.694;0.904]
Complete cases	0.928
Available cases	0.929
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782
MNAR “interval”	[0.753;0.891]

A more formal look at intervals of estimates

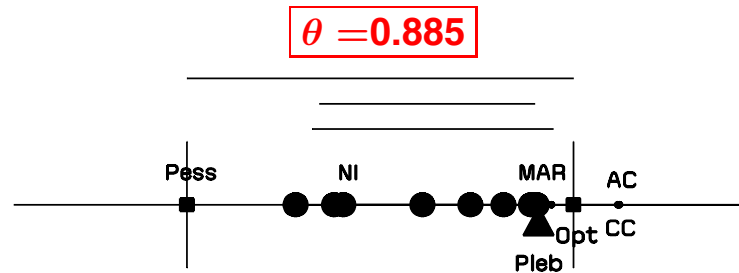
Handling Incomplete Data in Longitudinal Studies – p. 175/178

Non-identifiable Models Added

Model	Structure	d.f.	loglik	θ	C.I.
BRD1	(α, β)	6	-2503.06	0.891	[0.877;0.906]
BRD2	(α, β_j)	7	-2476.38	0.884	[0.868;0.899]
BRD3	(α_k, β)	7	-2471.59	0.881	[0.865;0.896]
BRD4	(α, β_k)	7	-2476.38	0.779	[0.702;0.857]
BRD5	(α_j, β)	7	-2471.59	0.848	[0.814;0.882]
BRD6	(α_j, β_j)	8	-2440.67	0.822	[0.792;0.850]
BRD7	(α_k, β_k)	8	-2440.67	0.774	[0.719;0.828]
BRD8	(α_j, β_k)	8	-2440.67	0.753	[0.691;0.815]
BRD9	(α_k, β_j)	8	-2440.67	0.866	[0.849;0.884]
Model 10	(α_k, β_{jk})	9	-2440.67	[0.762;0.893]	[0.744;0.907]
Model 11	(α_{jk}, β_j)	9	-2440.67	[0.766;0.883]	[0.715;0.920]
Model 12	$(\alpha_{jk}, \beta_{jk})$	10	-2440.67	[0.694;0.904]	

Handling Incomplete Data in Longitudinal Studies – p. 176/178

Intervals of Ignorance



Estimator	$\hat{\theta}$
[Pessimistic; optimistic]	[0.694;0.904]
MAR (3 questions)	0.883
MNAR	0.782
MNAR "interval"	[0.753;0.891]
Model 10	[0.762;0.893]
Model 11	[0.766;0.883]
Model 12	[0.694;0.904]

Handling Incomplete Data in Longitudinal Studies – p. 177/178

Conclusions

MCAR simple	CC LOCF	biased inefficient not simpler than MAR
MAR	direct likelihood multiple imputation weighted GEE	easy to conduct Gaussian & non-Gaussian
MNAR	variety of methods	untestable assumptions useful in sensitivity analysis