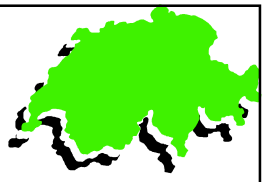
Interpreting treatment effects from a number of trials with application to designing future trials

Steven A. Julious
University of Sheffield



European Statistical Meeting on Meta Analysis

EFSPI

26th June 2009

European Statistical Meeting on Meta Analysis

9.00	Welcome *Norbert Neumann (President BBS)*
9.10	Interpreting Treatment Effects from a Number of Trials with Application to Designing Future Trials *Steven Julious (University of Sheffield)*
9:50	Sequential Methods for Random Effects Meta - Analysis *Anne Whitehead (University of Lancaster)*
10.30	Coffee
10:50	Meta-Analysis and Validation of Surrogate Endpoints *Tomasz Burzykowski (IDDI Belgium)*
11:30	Meta-Analysis to Compare Combination Therapies A Case Study in Kidney Transplantation *Amy Racine et al. (Novartis Basel)*
12.10	Subgroup Analysis using Bayesian Hierarchical Models: A Case Study *Beat Neuenschwander et al. (Novartis Basel)*
14.00	Work of the Cochrane Collaboration Cochrane collaboration: Systematic Reviews of the Effects of Healthcare Interventions *Sally Hollis (Astra Zeneca, UK)* Meta-Analysis on Erythropoiesis-Stimulating Agents *Guido Schwarzer (University of Freiburg)*
15.20	Coffee
15.40	Panel Discussion: Meta-Analyses and Implication for the Industry *Stephen Senn (University of Glasgow), David Wright (MHRA)*
16.30	Meeting closes

Steven A. Julious

European Federation of Statisticians in the Pharmaceutical Industry

Adaptive Randomisation

EFSPI Scientific Meeting on

Adaptive Randomisation - Today and Tomorrow

December 7th, 2006

Adaptive Randomisation – Today and Tomorrow

Morning: Adaptive Randomisation – A Personal and Collective Perspective

- Why I hate Minimisation **Stephen Senn (University of Glasgow)**
- Minimization, permuted blocks and simple randomization: **Marc Buyse (IDDI)**
- Dynamic allocation : How to stratify if you must: **Kit Roes (Organon and EFSPI)**
- Regulatory perspective: **P. Volkers (Paul-Ehrlich-Institute)**

- Discussion : **Andy Grieve (Pfizer)**

Afternoon session 1

- Use of the Minimization Technique at the EORTC: History and Experience: **C.Coens (EORTC)**
- Minimisation – Reducing Predictability whilst Retaining Balance within Centres: **S. Brown (CTRU, University of Leeds)**
- Case Studies Using Dynamic Randomisation Techniques other than Minimisation: **D. McEntegart (ClinPhone)**

Afternoon session 2

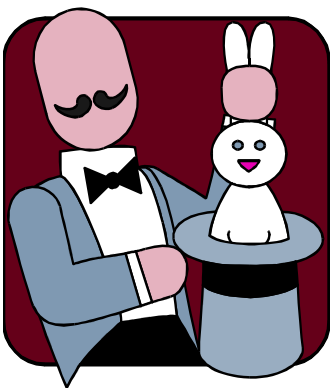
- Use of Simulation to Compare the Performance of Minimisation with Conventional Stratified Randomisation: **R. Toorawa (Covance)**
- Inference in Covariate-Adaptive Allocation: **E. Valdés Márquez (MPS, University of Reading)**
- A European Project for the Development and Deployment of a Validated Randomisation System: **E. van der Donk (NKI)**

Introduction

- Sample size calculations are one of the most important aspects of a trial
- To design a clinical trial we need to have a good estimates of
  - Effect size of interest
  - Population variance
  - Non-inferiority margin
- This presentation will discuss issues in their estimation

Sample Size Estimation Basics

Not as simple as……



Size matters….

A study that is **too small** or **too large** poses ethical problems

Too few

You will not be able to answer the question posed

Too many

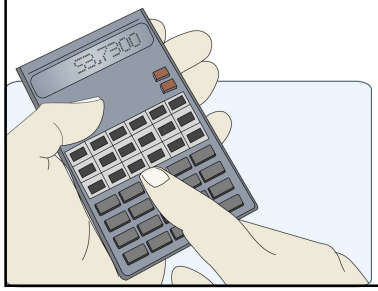
You will waste resources, and possibly give patients a treatment proven to be inferior

The Three Most Important Components of any Study Are

- Design
- Design
- Design

The sample size is just one component of the design

At least as much time should be spent designing a study as analysing it



Designing a Superiority Trial

Normal Approximation

- For the sample size for treatment group A is

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

where r is the allocation ratio, σ^2 is the known population variance, d is the treatment difference of interest and n_A is the sample size on A

- Note $n_B = rn_A$ and $n = n_B + n_A$ is minimised when r=1

Most of the Equation are Design Features

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

•The significance level would be set by precedent or other requirements

•The power would also be predetermined at 80-90%

•The allocation ratio would be set according to the study objective

Most of the Equation are Design Features

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

- The significance level would be set by precedent or other requirements
- **The power would also be predetermined at 80-90%**
- The allocation ratio would be set according to the study objective

Most of the Equation are Design Features

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

- The significance level would be set by precedent or other requirements
- The power would also be predetermined at 80-90%
- **The allocation ratio would be set according to the study objective**

We need to have an estimate of the variance

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

- The variance is often obtained from previously undertaken trials

We also need to have an estimate of the effect size

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

Where does an estimate of effect size come from?

What is an Effect Size?

- The clinically important effect such that a difference between treatments could be declared
- The minimum value worth detecting below which no difference can be declared
- Statistical significance v clinical significance
- What effects have been observed...

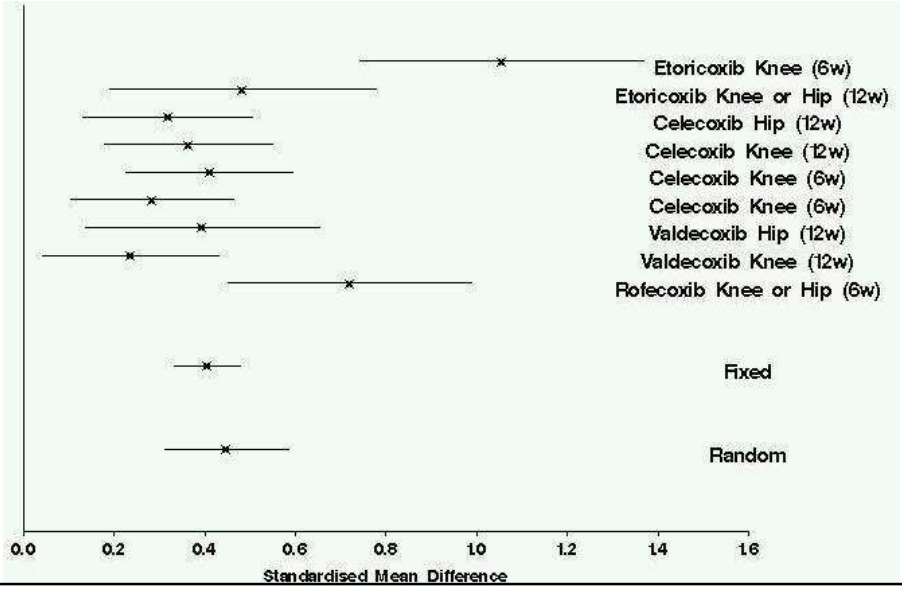
Meta analysis

- A meta analysis of 9 Cox-2 studies was undertaken for the indication of osteo-arthritis
  - Celecoxib, Eterocoxib, Rofecoxib and Valdecoxib.
- SBAs and publications were used as the data source.
- WOMAC Pain is taken as the endpoints of investigation.

Note

- The differences observed have had to be standardised to make them scale independent
  - This is done by dividing the observed mean difference by the observed standard deviation.
- This has a by product of removing variability as a consideration in the sample size calculation.

Meta analysis for WOMAC Pain Across the Cox-2 Compounds



Results

- The overall result was
  - 0.45 of an SD ≈ 10mm
- The minimum differences observed were
  - 0.28 of an SD ≈ 6mm
- SD is taken as 22mm

Steven A. Julious

The Calculations

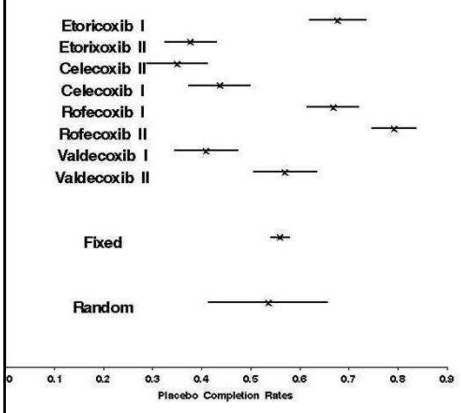
- If we powered on effect size 0.45 of an SD we'd required 105 patients per arm
  - If the true difference was nearly to 0.28 of an SD we'd have 52% power
- If we powered on an effect size of 0.36 (about 8mm) of an SD we'd required 164 patients per arm
  - If the true difference was nearer to 0.28 then we'd have 72% power

Other Important Design Factors

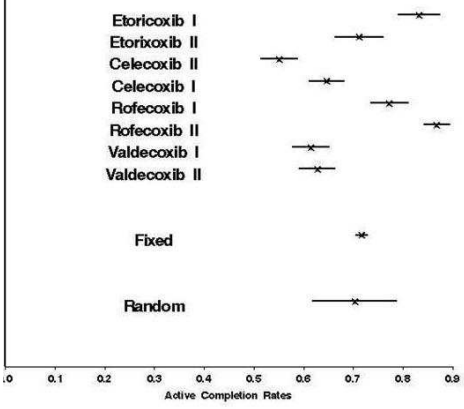
- The sample size calculation is number of evaluable subjects required for analysis
  - What is the total sample size required to ensure evaluable number of subjects for analysis?
  - What proportion of subjects will make it into the per protocol population
    - Important for non-inferiority studies
- Factors such as these could be assessed from retrospective studies
  - Through a meta analysis..

Completion Rates in Different Trials

Placebo



Active



In RA

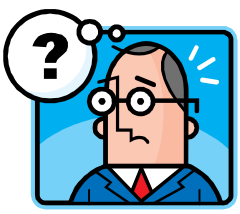
Should a Meta Analysis be Used to Help Design a Future Study?

Should a Meta Analysis be Used to Help Design a Future Study?

Yes

Should a we Simply Use a Point Estimate from a Study to Design a Future Study?

Should a we Simply Use a Point Estimate from a Study to Design a Future Study?



With Caution

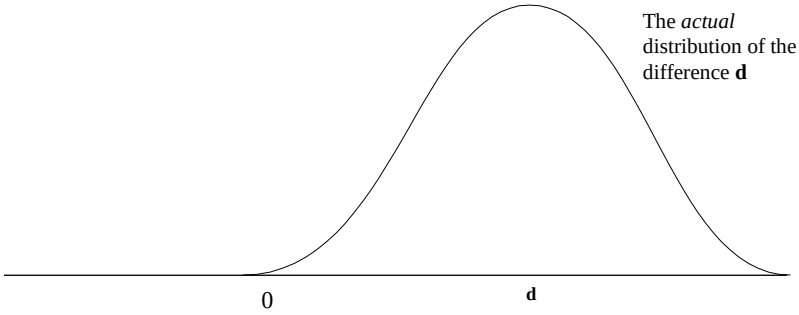
Designing a Phase III Trial

- You design a Phase III trial
  - The standard deviation is assumed to be s
  - The effect size is d
  - You calculate a sample size n with 90% power and two side significance level of 5%
- The trial is run and you see exactly the same effect (d) as you designed upon. Your two sided P-value is:
  - P=0.002

Why is this so?

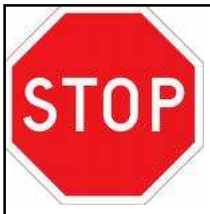
- This is all down to the distribution under the alternative hypothesis
- Suppose the alternative hypothesis is true…

Distribution Under the Alternative Hypothesis



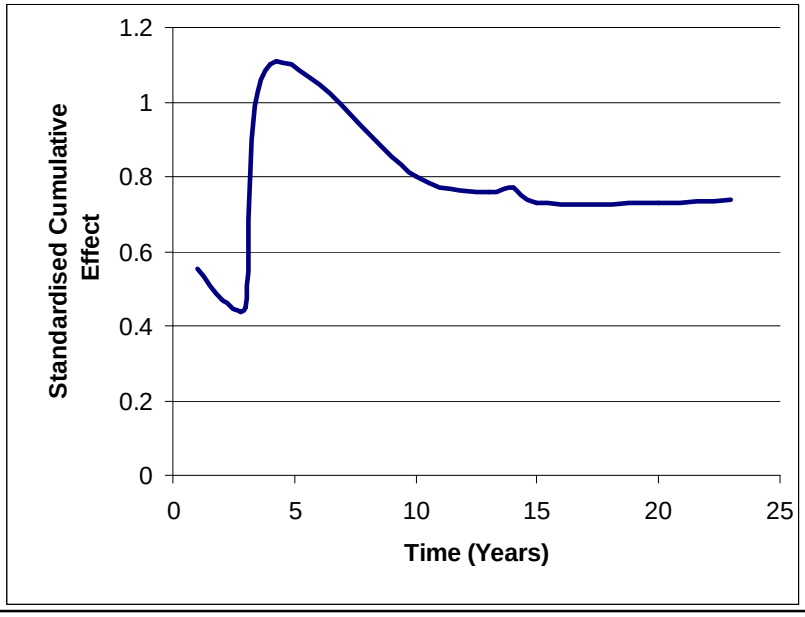
What does this mean?

- Under the alternative hypothesis there is a distribution of responses
  - And the chance to make a Type II error
- For P-value of P<0.05 if d is the true effect then we could be observing 0.6 of d
  - If an effect of d is still chosen as the effect size for Phase III given that we have seen P<0.05 then we may only have power of 56%
  - If we are to use the observed response from IIb we'd need a P-value of 0.002 to have 90% power



What about Going from Phase to Phase

Cumulative Meta Analysis of Paroxetine Data



Variability

Normal Approximation

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{rd^2}$$

This variance estimate is usually estimated from retrospective data sometimes from a number of studies.

Population Considerations

- Is the study population similar to your own?
- The most obvious consideration is to ask whether the demographics were the same - for example if the trial conducted was a multi centre one was it conducted in similar countries?
  - Different countries may have different types of care (e.g. different concomitant medication) and so may have different trial populations.
- Was the same type of patient enrolled?
  - The same mix of mild, moderate and severe?
- Was it conducted over the same seasons?
  - Relevant for conditions such as asthma?

Illustrative Example

Study	HAMD Entry Criteria	Number of Centres	Duration	Year	Pop	Region	Phase	DF	Var
1	18	1	6	1984	A	N. Amer	II	22	41.59
2	18	1	6	1985	A/G	N. Amer	II	160	59.72
3	18	6	6	1985	A/G	N. Amer	III	232	57.11
4	21	3	6	1986	A/G	N. Amer	III	9	62.97
5	18	10	6	1985	A/G	N. Amer	III	49	58.32
6	18	28	12	1991	A/G	N. Amer	III	109	42.51
7	18	23	12	1991	A/G	N. Amer	III	133	68.98
8	18	12	8	1992	A/G	N. Amer	III	121	51.81
9	18	1	6	1982	A	Europe	III	19	62.44
10	15	1	6	1983	A/G	Europe	III	8	44.71
11	15	12	12	1994	A/G	N. Amer	III	80	38.81
12	13-18	12	8	1994	P	N. Amer	III	85	46.09
13	15	18	10	1994	A	N. Amer	IV	41	60.01
14	15	20	12	1996	A	N. Amer	III	99	61.42
15	20	20	12	1996	A	N. Amer	III	108	61.65
16	18	29	12	1996	G	N. Amer	III	105	45.54
17	20	40	8	2001	A/G	N. Amer	III	140	58.36
18	18	1	4	1983	A	Europe	III	20	43.64
19	18	1	4	1983	A	Europe	III	1	19.32
20	18	1	4	1989	A	Europe	II	2	43.90

Variances Broken Down by Population and Region

Population	Overall		Europe		North America	
	Var	df	Var	df	Var	df
All	55.03	1543	50.48	50	55.19	1493
Adult	58.59	312	51.58	42	59.70	430
Adult/Geriatric	55.66	1041	44.71	8	55.74	1033
Paediatric	46.09	85	.	.	46.09	85
Geriatric	45.54	105	.	.	45.54	105

What about European Paediatric and Geriatric Populations?

Population	Overall		Europe		North America	
	Var	df	Var	df	Var	df
All	55.03	1543	50.48	50	55.19	1493
Adult	58.59	312	51.58	42	59.70	430
Adult/Geriatric	55.66	1041	44.71	8	55.74	1033
Paediatric	46.09	85	.	.	46.09	85
Geriatric	45.54	105	.	.	45.54	105

Designing a Non-Inferiority Study

Non-inferiority Trials

- A non-inferiority study is usually planned to detect if the effect of the investigative treatment is not much worse than the control treatment
  - Defined by a non-inferiority margin, *d*.
- An assessment of non-inferiority of a new treatment is usually performed by comparing the worst tail of 95% confidence interval with the non-inferiority margin to rule out the inferiority of a new treatment.
- The threshold setting of *d* is not straightforward
- Often the margin is defined as some fraction of the active control effect over placebo to be retained
  - Known as the preservation fraction

Background to Non-inferiority

- A non-inferiority assessment could be considered indirect comparison requires Trial 1 and Trial 2 be conducted in sequence with the following setups
  - Trial 1: Placebo and Treatment A,
  - Trial 2: Treatment A and Treatment B,
- Here treatment A should have been shown effective in trial 1 (a placebo-controlled trial) in order to launch Trial 2 (an active-controlled trial).

Background to Non-inferiority

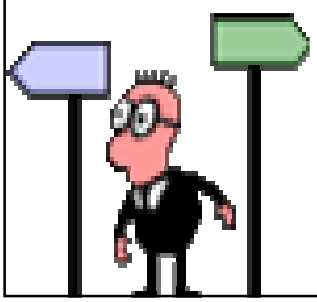
- In practice for ‘Trial 1’ a meta analysis my be undertaken to assess the benefit of A over placebo
- The non-inferiority margin is then set as some fraction of this observed effect
- By excluding this margin we can indirectly conclude superiority over placebo?

Assumptions

Three General Assumptions for Indirect Comparisons

- Assay sensitivity of the studies.
- Constancy of the effect of the common comparator in both scenarios.
- The patient population and the primary efficacy endpoint studied between the two trials are essentially the same in both scenarios.

Issues with Indirect Comparisons

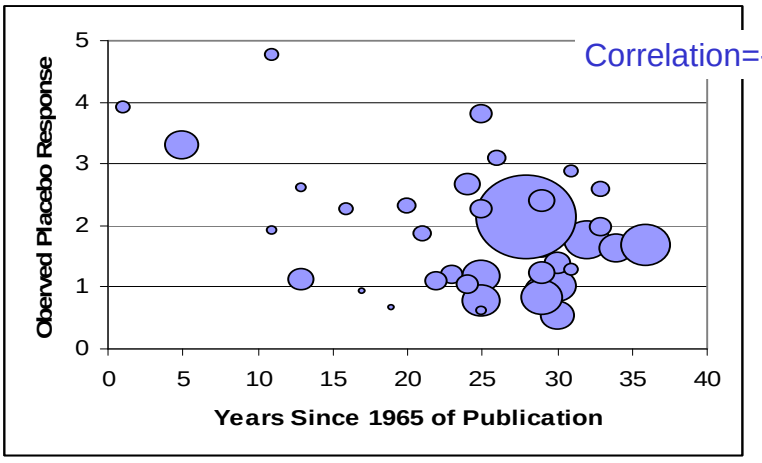


Placebo Improving with Time

Year Published	Sample Size	Placebo Occlusion Rate	Placebo Event Rate	Comparison To
1984	81	0.35	0.63	Low Dose
1989	345	0.23	0.44	High Dose
1990	750	0.18	0.33	Low Dose
1991	328	0.11	0.30	High Dose
1993	145	0.09	0.17	Low Dose

Taken from Lim et al 2003 BMJ 327;1309-13

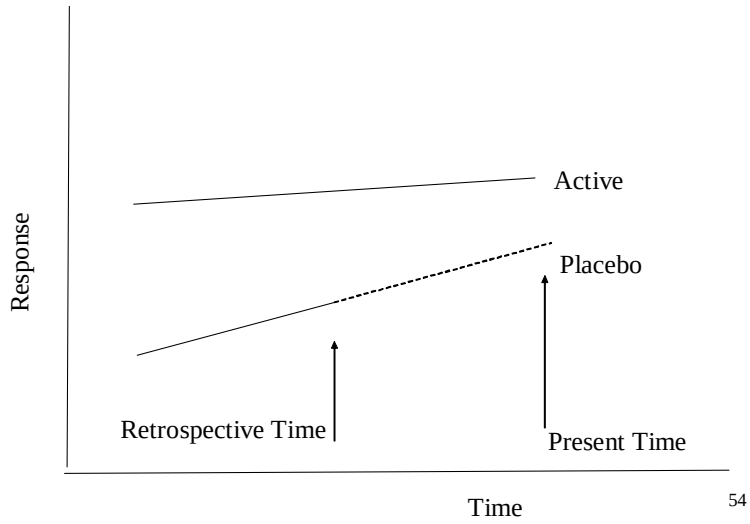
Placebo Response against Year since Publication in Depression



Impact on Non-inferiority

- If there is placebo response is improving with time in the population in which the trial is to be conducted, the assessment of effect over placebo could be biased.
- The bias issue would occur when we wish to choose between two treatments that are not investigated in the same trial, but have each been compared with placebo.

Graphical Representation of Placebo Creep



Issues in Setting Non-inferiority Margins

- There is an issue in that the estimate of effect over placebo in 'Trial 1' may possibly be overestimated for comparison in Trial 2
  - Due to the placebo responses improving over time i.e. placebo creep.
- There is an issue in that the lack of constancy of control effect prescribed by the placebo creep cannot be formally tested.

Considerations in Defining the Margin

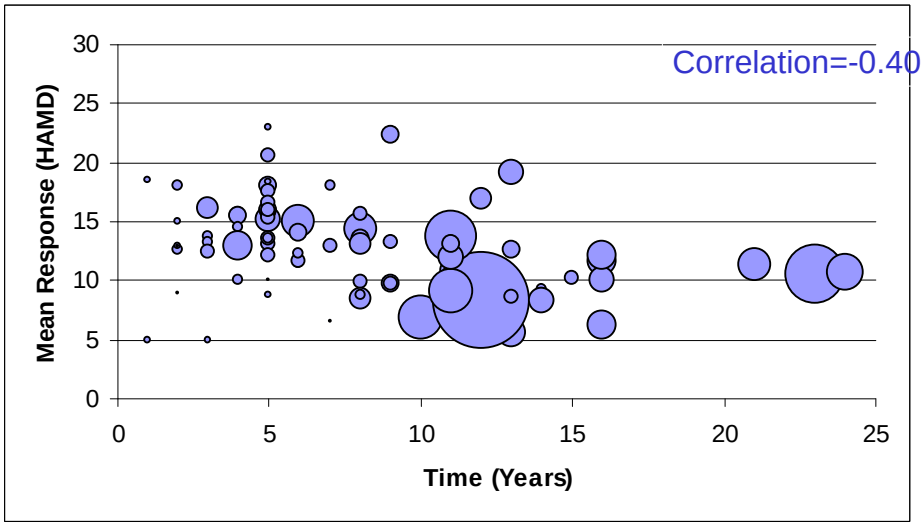
- How should the heterogeneity of the control effect and its variability across completed placebo-controlled trials, vis-a-vis Trial 1, be incorporated?
- Should differential weight be given to the response from the most recent studies and/or from the studies with smaller effects?
- Should when setting the preservation fraction some discount be set for the active control effect due to placebo creep?

Illustrative Example

The Data

- The data used are taken from the GlaxoSmithKline Clinical Trial Register that are publicly available (ctr.gsk.co.uk).
- Paroxetine trials indicated for depression treatment were chosen where Hamilton Depression Rating Scale (HAMD) was the primary efficacy outcome.
- 81 trials dating from 1981 through to 2004 were identified with the primary time point (varied from 5 to 12 weeks) absolute mean value taken for comparison purposes.
- Of these trials 17 were from Phase II, 54 from Phase III and 10 from Phase IV.

Mean response and variance of the response for Paroxetine by time



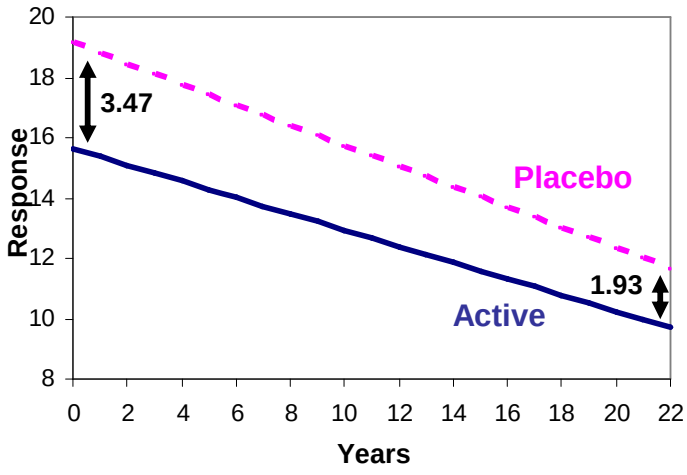
Analysis of Individual Treatments

	Paroxetine			Placebo		
	Correlation	Intercept	Slope	Correlation	Intercept	Slope
81 Trials	-0.40	14.29	-0.25	na	na	na
23 Trials	-0.60	15.64	-0.27	-0.57	19.11	-0.34
16 III/IV Trials	-0.51	15.05	-0.23	-0.49	18.56	-0.30

Analysis of Treatment Differences

	Correlation	Intercept	Slope
23 Trials	-0.278	3.472	-0.070
16 III/IV Trials	-0.279	3.504	-0.072

Predicted responses from regression analysis for Placebo and Active



Discussions

- It was highlighted how when designing a clinical trial there may be issues in quantifying
  - Effect size of interest
  - Population variance
  - Non-inferiority margin