Society for Clinical Trials 32nd Annual Meeting

Workshop P9
Challenges in Designing Small Clinical Trials

Sunday, May 15, 2011
1:00 PM - 5:00 PM
Plaza B

Small Clinical Trials: Statistical Aspects

Christopher S. Coffey, PhD
Department of Biostatistics College of Public Health
University of Iowa

Janet Wittes, PhD
Statistics Collaborative
+
Surprise Guest!

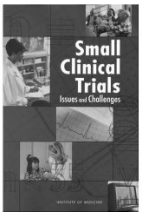
SCT Pre-Conference Workshop – May, 2011

What is "small"?

- A trial of a (super)-orphan disease
- A Phase 1 or Phase 2 trial
- An inadequately powered trial

Overview

- Executive Summary from National Academy of Sciences



Part 1:
Squeezing Precision from a Stone

- *Design and Analysis of Clinical Trials When the Target Population is Very Small*

#1: Define research question

Before undertaking a small clinical trial it is particularly important that the research question be well defined and that the outcomes and conditions to be evaluated be selected in a manner that will most likely help clinicians make therapeutic decisions.

VI-5

#2. Tailor the design

Careful consideration of alternative statistical design and analysis methods should occur at all stages in the multistep process of planning a clinical trial.

When designing a small clinical trial, it is particularly important that the statistical design and analysis methods be customized to address the clinical research question and study population.

VI-6

#3. Have clear methods of reporting

- In reporting the results of a small clinical trial, with its inherent limitations, it is particularly important to carefully describe all sample characteristics and methods of data collection and analysis for synthesis of the data from the research.

VI-7

#4. Do corroborative statistical analyses

Given the greater uncertainties inherent in small clinical trials, several alternative statistical analyses should be performed to evaluate the consistency and robustness of the results of a small clinical trial.

VI-8

#5. Exercise caution in interpretation.

One should exercise caution in the interpretation of the results of small clinical trials before attempting to extrapolate or generalize those results.

VI-9

#6. More research on alternative designs needed

Appropriate federal agencies should increase support for expanded theoretical and empirical research on the performances of alternative study designs and analysis methods that can be applied to small studies.

Areas worthy of more study may include theory development, simulated and actual testing including comparison of existing and newly developed or modified alternative designs and methods of analysis, simulation models, study of limitations of trials with different sample sizes, and modification of a trial during its conduct.

VI-10

Summary

1. Define your question
2. Tailor your design
3. Report clearly
4. Do corroborative analyses
5. Interpret cautiously
6. Do more research on other designs

VI-11

And…what else is new?

1. Define your question
2. Tailor your design
3. Report clearly
4. Do corroborative analyses
5. Interpret cautiously
6. Do more research on other designs

So why is this any different from other trials?

VI-12

What do we specify in ordinary trials?

- Patient population
- Question of interest / hypothesis to be tested
- Primary outcome
- Type I error rate
- Clinically meaningful difference and desired power
- Other nuisance parameters

VI-13

What do we do with these specs?

- Before the trial
  - Design the trial
  - Compute required sample size
- After the trial
  - Perform analyses
  - Interpret results

VI-14

Designs of interest in small clinical trials

- Longitudinal analysis
- Parallel group design
- Crossover design
- Add-on design
- N-of-1 design
- Sequential design
- Ranking/selection design
- Adaptive designs

VI-15

Theme

- There is no magic – small is not large
- We want the “right” answer
  - Perhaps less precision than in large trials
  - Perhaps we use a less definitive outcome
- We must know what we are sacrificing

Let No Charitable Hope

I was, being human, born alone;
I am, being woman, hard beset;
I live by squeezing from a stone
The little nourishment I get.

Elinor Wylie, 1932

Small study ≠ little version of big study

Basic Formula for Design

$$N=\frac{2\sigma^2(z_\alpha+z_\beta)^2}{\Delta^2}$$

N=sample size

σ^2=variance

Δ = difference clinically important to detect
(or smallest biologically credible difference)

Basic Formula for Design

Rewriting….

$$N \propto \sigma^2/\Delta^2$$

Where constant of proportionality is a function of:
Number of groups
Type 1 error rate
Type 2 error rate

$\Delta \propto \sigma/\sqrt{N}$

- If the population we are studying is big
  - E.g., cardiology, breast ca, Alzheimer's
- Just increase N to reduce Δ
- And (pace regulators and CROs)
  - A little sloppiness is not harmful

$\Delta \propto \sigma/\sqrt{N}$

- If the population we are studying is big
  - E.g., cardiology, breast cancer, Alzheimers
- Just increase N to reduce Δ

- BUT: If our population is small
  - E.g., genetic disease, rare cancers
- Cannot increase N
- Only solution: decrease variance

What methods can we use?

- Check that the sample size is truly small
- Use efficient outcome measures
- Use efficient designs when applicable

What is potential population?

- PiZZ emphysema
  - Genetics “known”
  - Pedigrees “known”
  - Therefore, can estimate population in US
- Our argument
  - Too few people to do RCT
  - Therefore, use historical control
- FDA: approved replacement therapy

What is potential population?

- Treatment became available
- Population increased…a lot!
- Consequence
  - We can't know effect of replacement therapy

Lesson #1:
Population may be larger than you think

- Don't rely on KOLs
  - "key opinion leaders"
- Use the literature, but be skeptical

Analytical Approaches

- Use efficient outcome measures
- Measure precisely
- Measure often (but not too often)
  - Rough rule: $\sigma^2=\sigma_{between}^2+\sigma_{within}^2/k$
  - So, if we take lots of measurements
    $\sigma^2=\sigma_{between}^2$
  - There is a point of diminishing returns
- Use efficient analytical methods

Use efficient outcome measures

- Continuous outcomes most efficient
  - Often clinically useful
  - Beware statistically, not clinically, significant
- Binary outcomes are least efficient
  - Sometimes the only outcome of real interest
  - Elimination of disease, restoration of function
- Time to event may be more efficient than binary

Some calculations for binary

The KOLs say:
"We NEVER see success in the controls
We expect all to succeed in treated."
Suppose all treated "succeed" and all controls "fail".
What are the one-sided p-values?

	6	8	9	10	12	14	15
1:1	.05	.015	-	.005	.0011	.0029	-
2:1	.067	-	.012	-	.0046	-	.00034

VI-29

But what if one control succeeds?

	8	9	10	12	14	15
1:1	.015	-	.005	.0011	.00029	-
2:1	-	.012	-	.0046	-	.00034

	8	9	10	12	14	15
1:1	.071	-	.024	.0076	.0023	-
2:1	-	.083	-	.0018	-	.0037

VI-30

And what if one treated fails?

	8	9	10	12	14	15
1:1	.015	-	.005	.0011	.00029	-
2:1	-	.012	-	.0046	-	.00034

	8	9	10	12	14	15
1:1	.071	-	.024	.0076	.0023	-
2:1	-	.083	-	.0018	-	.0037

	8	9	10	12	14	15
1:1	.071	-	.024	.0076	.0023	-
2:1	-	.048	-	.010	-	.0014

So we better have at least 10 participants.

VI-31

If 1 control succeeds & 1 Rx fails

	10	12	14	15
1:1	.11	.04	.015	-
2:1	-	.067	-	.017

So we better have more than 12!!!

VI-32

Some examples (some “clinical”; some “surrogate”)

Outcome	Continuous	Binary	Time to event
MPS VI	Urinary GAG	Urinary GAG>150	Not relevant
	12min walk	Walk>150 meters	Not relevant
Hypertension	SBP	SBP<70mmHg	Not relevant?
Pain	Pain score	Pain>4 (7pt scale)	Time to pain relief
Vision	# letters lost	Loss of 15 letters	Time to 15 letter loss

Some examples (some “clinical”; some “surrogate”)

Outcome	Continuous	Binary	Time to event
MPS VI	Urinary GAG	Urinary GAG>150	Not relevant
	12min walk	Walk>150 meters	Not relevant
Hypertension	SBP	SBP<70mmHg	Not relevant?
Pain	Pain score	Pain>4 (7pt scale)	Time to pain relief
Vision	# letters lost	Loss of 15 letters	Time to 15 letter loss

So, if we can convince ourselves that the continuous measure is clinically interesting, we gain efficiency going with it.

My view: I am less fussy about surrogate/clinical in orphans than in common diseases.

Measure precisely and often **(but not too often or too precisely)**

- Rough rule: $\sigma^2=\sigma_{between}^2+\sigma_{within}^2/k$
- So, if we take lots of measurements

$$\sigma^2=\sigma_{between}^2$$

- There is a point of diminishing returns

What can we gain in Δ?

Ratio/#measures	1	2	3	4	10	∞
5	1	1.04	1.06	1.07	1.08	1.10
3	1	1.07	1.10	1.11	1.14	1.15
1	1	1.15	1.23	1.26	1.35	1.41

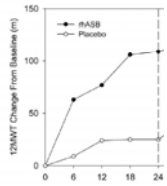
Ratio: Between Variance/Within Variance

Measures: number of measurements at a single time

Use efficient statistical methods

- Suppose you are measuring over time
- Standard: Final value – baseline
- Better: Final value; baseline as covariate

Mucopolysaccharidosis (MPS) VI



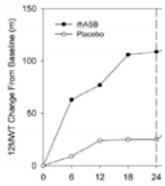
Use efficient statistical methods

- Suppose you are measuring over time
- Standard: Final value – baseline
- Better: Final value; baseline as covariate
- Still better: Longitudinal
  - Differentiate “through” vs. “at”
  - Think about variance/covariance structure
  - Think how you want to model time

MPS VI

Use of t-test
mean Δ: 82 m
p-value: 0.10

Use of longitudinal analysis
mean Δ: 94 m
p-value: 0.025



How to increase power

- Usual RCT: as model-free as possible
  - Have large sample sizes
  - Do ITT
  - Simple outcome: “Count the corpses”
  - Don’t worry about noise

How to increase power

- Usual RCT: as model-free as possible
- Small populations
  - Use models (but prespecify)
  - Check EACH observation before you unblind
  - Think of other designs

Designs

- Parallel design
  - Randomize experimental vs. control
  - Compare the two with respect to outcome
    - Preselected, of course!

Design: historical control

- When are they ok?
  - Outcomes with standard or no treatment
    - Well studied
    - Data from the studies are relevant to today
  - New treatment expected to be much better
  - Outcome can be objectively measured
  - Randomized controls not feasible

Historical controls

- Advantages
  - Inexpensive (….not always!)
  - Patients don't get nothing or SOC
  - You often find a BIG difference
- Disadvantage
  - Current & historical populations may differ
  - Current treatment may be different
    - (even if there is no "therapy")

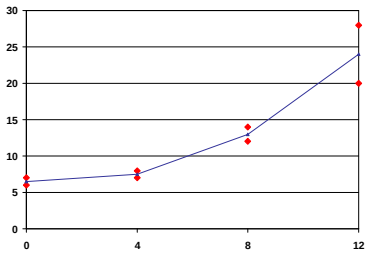
Lessons

- Lesson #1:
  - Use literature, but be skeptical
- Lesson #2:
  - Small population no excuse for poor study
    - Choice
      - Controlled study with low power
      - Historical control with higher power

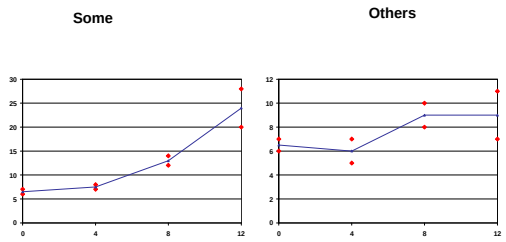
Lesson #2

- Small population no excuse for poor study
- Choice
  - Controlled study with low power
  - Historical control with higher power

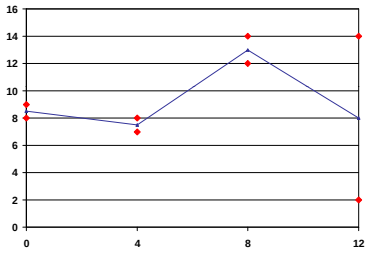
Example: A tiny study (n=40) Some people looked like this



Example: A tiny study and pop-



And one looked like this:



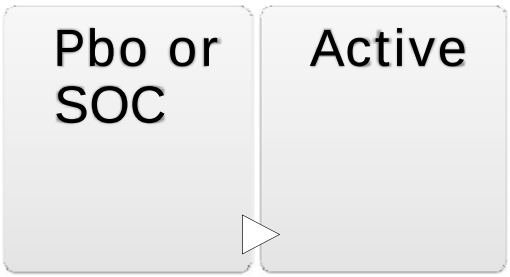
What we did

- Asked why the low point
  - Answer: person was sick
  - She asked to reschedule; investigator “no”
  - We deleted that point
- But, we reviewed EVERY point for illness
  - No other comment about illness
  - We deleted no other point

Lessons

#1: Use literature, but be skeptical
#2: Small pop no excuse for poor study
#3: Think of each observation as precious

The two-stage crossover design

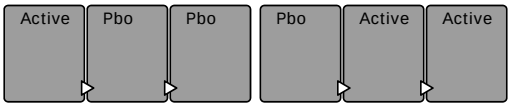


Design: Crossover

- Each subject exposed to all treatments
  - Order of treatments randomized
  - First may show better (or worse) effect)
- Prognostic factors balanced–self vs. self
- N reduced considerably (self to self)
- All participants receive some active Rx

A way to deal with carryover

- Half look like this Half look like this



Problems

- Treatment must be taken regularly over time
- Relevant outcomes
  - must occur and
  - be measured over time
- Not relevant to acute treatments
  - antibiotics for infections
  - treatment for myocardial infarction

When would cross-over trial work?

- Examples
  - Epilepsy
  - Diabetes
  - Pain relief
  - Asthma

Example: MPS VI

ENZYME REPLACEMENT THERAPY FOR MUCOPOLYSACCHARIDOSIS VI: A PHASE 3, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, MULTINATIONAL STUDY OF RECOMBINANT HUMAN N-ACETYLGALACTOSAMINE 4-SULFATASE (RECOMBINANT HUMAN ARYLSULFATASE B OR RHASB) AND FOLLOW-ON, OPEN-LABEL EXTENSION STUDY

Paul Harmatz, MD, Roberto Giugliani, MD, PhD, Ida Schwartz, MD, Nathalie Guffon, MD, Elsa Leão Teles, MD, M. Clara Sá Miranda, PhD, J. Edmond Wraith, MD, Michael Beck, MD, Lala Arash, MD, Maurizio Scarpa, MD, Zi-Fan Yu, ScD, Janet Wittes, PhD, Kenneth I. Berger, MD, Mary S. Newman, MS, Ann M. Lowe, MD, Ehl Kakkis, MD, PhD, and Stuart J. Swiedler, MD, PhD, for the MPS VI Phase 3 Study Group*

J Pediatrics, 2006; 533-539

Study

- 39 patients with MPS VI
- Rare, fatal, lysosomal storage disease
- 24 week study
- Primary outcome: distance walked in 12 min
- Treatment groups:
  - recombinant human arylsulfatase B(rhASB)
  - Matching placebo

Epi study

- 121 patients
- High urinary GAG (>200μg/mg creatinine)
  - Short stature
  - Low body weight
  - Impaired endurance
  - Compromised pulmonary function
  - Reduced joint range of motion

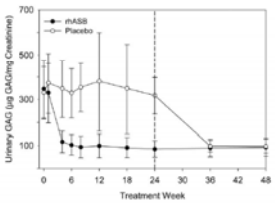
Phase 1 and Phase 2

- Phase 1
  - rhASB tolerated and rapidly reduced GAG
- Phase 2
  - 11 patients showed improved 12MWT
  - No controls

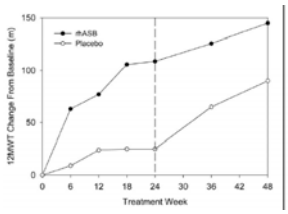
Phase 3

- 12 MWT measured at baseline and
  - Weeks 6, 12, 18, 24, 36, 48
- At week 24, placebo patients got product
- A number of secondaries
  - Stair climb
  - ROM
  - Pulmonary function
  - Safety
  - Immunogenicity

Results: urinary GAG



Results: 12MWT



The warts

- Change in primary outcome
  - Original outcome: Week 24 – baseline
  - Change: Longitudinal analysis
  - Okayed change with FDA
- We removed one observation
- Baseline 12MWT (“p”=0.014)
  - rhASB: 227±170
  - Placebo: 381±202

Lessons

- Every number is important
- People don’t understand randomization
- Think hard about reducing variance
- Don’t be too afraid of making changes
  - Make sure you are totally blind when you do

Summary

- Small studies: hard to do and unforgiving
- They must be data-intensive
- Think hard about the analysis
  - Use outcomes that are meaningful
  - Use statistical methods that are efficient
- You have to deal with the sample you get
  - (pace Rumsfeld)
  - You will not have a second chance!

The Mystery Guest

N-of-1 RCTs

Example: An asthmatic man, taking 2 sorts of pills and 2 sorts of puffers, still had “attacks” of severe shortness of breath on bathing or dressing, and even in the middle of the night.

Neither he nor his physician was convinced that one of his pills was helping him (although both of them thought that it probably was helpful).

N-of-1 trial of the pill

By mutual agreement, and in collaboration with statistician-colleagues, a double-blind N-of-1 trial of that pill was set up:

Designed to determine its effect on his treatment targets of:

- shortness of breath,
- need for his rescue inhaler, and
- sleep disturbance,

over treatment periods lasting 10 days.

Design

These 10 day periods were paired and then randomized to provide him either active drug or an identical-appearing placebo during the first period - with the alternative treatment during the second period of each pair.

At the end of each period he reported his symptom severity on a 7-point Likert scale:

- the higher the score, the better he felt.

Results

After 2 pairs of treatment periods it was clear to both of them that he was very much better during some periods than others. They stopped the trial and looked at his reports.

Results

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing				
Breathlessness - bathing				
Breathlessness – at night				
Need for rescue inhaler				
Sleep disturbance				

Period 1

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3			
Breathlessness - bathing	3			
Breathlessness – at night	4			
Need for rescue inhaler	3			
Sleep disturbance	5			

Periods 1 and 2

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3	6		
Breathlessness - bathing	3	5		
Breathlessness – at night	4	7		
Need for rescue inhaler	3	5.5		
Sleep disturbance	5	5.5		

Periods 1 and 2

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3	6		
Breathlessness - bathing	3	5		
Breathlessness – at night	4	7		
Need for rescue inhaler	3			
Sleep disturbance	5			

Second pair of periods

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3	6	3	
Breathlessness - bathing	3	5	3	
Breathlessness – at night	4	7	4	
Need for rescue inhaler	3	5.5	3	
Sleep disturbance	5	5.5	3	

Second pair of periods

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3	6	3	6
Breathlessness - bathing	3	5	3	5
Breathlessness – at night	4	7	4	5
Need for rescue inhaler	3	5.5	3	5
Sleep disturbance	5	5.5	3	5

Second pair of periods

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3	6	3	6
Breathlessness - bathing	3	5	3	5
Breathlessness – at night	4	7	4	5
Need for rescue inhaler	3	5.5	3	
Sleep disturbance	5	5.5	3	

Period 2 vs Period 1

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Period 1	Period 2	Period 1	Period 2
Breathlessness - dressing	3	6	3	6
Breathlessness - bathing	3	5	3	5
Breathlessness – at night	4	7	4	5
Need for rescue inhaler	3		3	
Sleep disturbance	5		3	

Which was better?

Symptom Lower score = worse	First pair of periods		Second pair of periods	
	Active	Placebo !	Active	Placebo !
Breathlessness - dressing	3	6	3	6
Breathlessness - bathing	3	5	3	5
Breathlessness – at night	4		4	5
Need for rescue inhaler	3		3	
Sleep disturbance	5		3	

So they stopped that pill forever!

They then did a second N-of-1 trial of his inhaler.
- he again quit after 2 pairs of treatment periods.
- breaking the second code revealed that he and his clinician were wrong again!

Six months later he was much less symptomatic
- on one less drug!

Data outputs from N-of-1 Trials

Global impressions and preferences

Ordinal scales (Awful, Lousy, Tolerable, Okay, Great)

Likert Scales

0 – 100

Days in each treatment

Bayesian Inferences

Prerequisites for N-of-1 RCTs

Is an N-of-1 RCT <u>really indicated</u> for your patient?

1. Is the effectiveness of the treatment really in doubt?
2. Will the treatment, if effective, be long-term?
3. Is your patient eager to collaborate in designing and carrying out an N-of-1 RCT?

Prerequisites for N-of-1 RCTs

Is an N-of-1 RCT feasible in your patient?

1. Does the treatment have a rapid onset?
2. Does the treatment stop acting soon after it is discontinued?
3. Is an optimal duration for a treatment period feasible?
4. Can outcomes that are relevant and important to your patient be measured?
5. Can sensible criteria for stopping the trial be established?
6. Is an unblinded run-in period necessary?

Prerequisites for N-of-1 RCTs

Is this N-of-1 RCT feasible in your practice setting?

1. Is there a pharmacist who could help you?
2. Is help available for interpreting the data?
   Is this N-of-1 RCT ethical?
3. Is there free, informed consent?
4. Can your patient withdraw from the trial without loss of care?
5. Will the same degree of confidentiality apply as in other clinical situations?

And now for something entirely different

Phase-2 Futility Trials

- Retreat from the current practice of conducting Phase 3 trials as the first rigorous test of efficacy for a treatment with a (reasonably) early putative benefit
- eg, preventing peri-operative complications,
- eg, preventing early deterioration,
- eg, speeding early stages of healing/recovery

- Instead, conduct Phase 2 “futility” trials

Setting up a Phase-2 Futility Trial

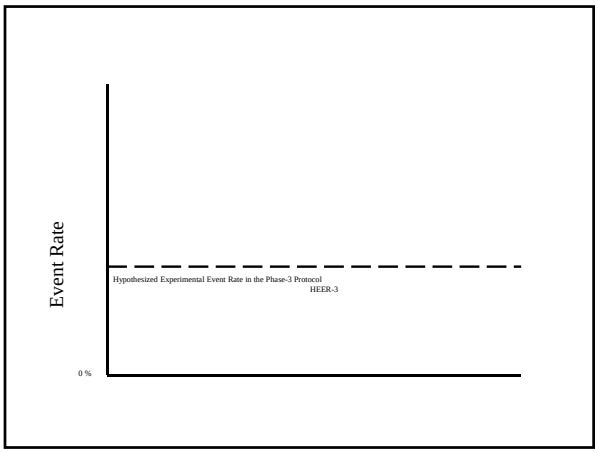
- Take the Hypothesized Experimental Event Rate you were going to use in calculating your sample-size for your Phase- 3 trial of new vs. old Rx:
  (HEER-3).

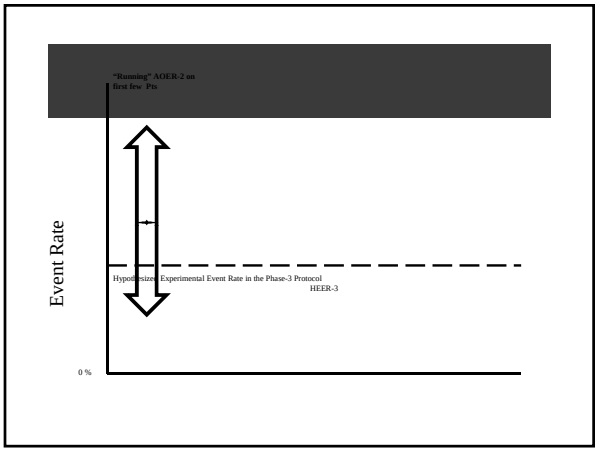
- But instead of actually starting your Phase-3 trial, treat everybody with your new Rx.

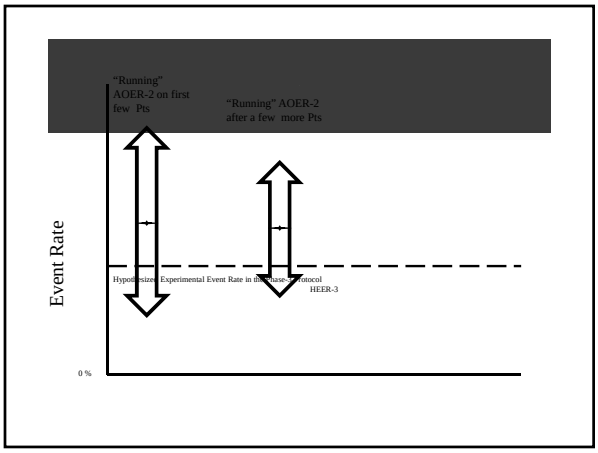
Running a Phase-2 Futility Trial

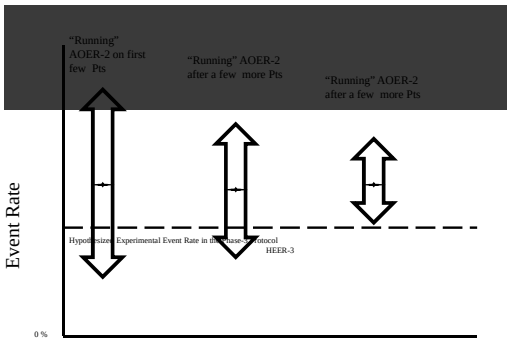
- Keep a “running” comparison between HEER-3 and the confidence interval for the event rate you are Actually Observing in your Phase-2 Futility Trial
  (AOER-2)

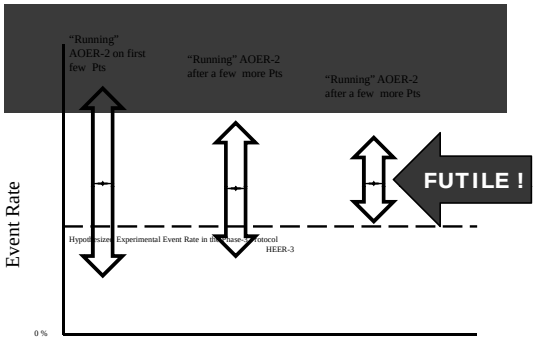
- Stop for futility if the lower bound of the confidence interval for AOER-2 exceeds HEER-3


Event Rate
Hypothesized Experimental Event Rate in the Phase-3 Protocol
HEER-3
0 %


"Running" AOER-2 on first few Pts
Event Rate
Hypothesized Experimental Event Rate in the Phase-3 Protocol
HEER-3
0 %


"Running" AOER-2 on first few Pts
"Running" AOER-2 after a few more Pts
Event Rate
Hypothesized Experimental Event Rate in the Phase-3 Protocol
HEER-3
0 %





But will AOER-2 >>> HEER-3 ever happen?

- NINDS funded 88 Phase-3 trials of "neuroprotective" drugs for lessening the early damage of strokes.
- Tonnes of blood, toil, sweat, tears and $$$ later, NONE of them found a benefit from any of these drugs (only 1 was marginally positive).
- Yuko Palesch and Barbara Tilley (with DLS as scribe) got their hands on 6 of these Phase-3 trials.

But will AOER-2 >>> HEER-3 ever happen?

1. We used the Hypothesized Experimental Event Rates (HEER-3)s from their Phase-3 protocols
2. We then took their Experimental Patients, one-by-one, as they entered their Phase-3 trials
3. And followed their progress as if they were consecutive patients in a Phase-2 trial, thereby generating a "running" Actually Observed Event Rate (AOER-2)

But will AOER-2 >>> HEER-3 ever happen?

4. Kept track of the Confidence Interval around this running Actually Observed (Phase 2) Event Rate (AOER-2)
5. And compared this emerging confidence interval with the Hypothesized Experimental Event Rate (HEER-2) they had hoped to observe for an effective treatment, as stated in their Phase-3 protocol

But will AOER-2 >>> HEER-3 ever happen?

6. If the confidence interval for the Actually Observed (Phase-2) Event Rate (AOER-2) excluded the Hypothesized Experimental Event Rate HEER-3, we declared it <u>futile</u> to proceed to Phase-3

Example

Tirilazad mesylate for better function	# of Patients

Example

Tirilazad mesylate for better function	# of Patients
From their Phase 3 Trial Sample Size calculation	

Example

Tirilazad mesylate for better function	# of Patients
From their Phase 3 Trial Sample Size calculation	1130

Example

Tirilazad mesylate for better function	# of Patients
From their Phase-3 Trial Sample Size calculation	1130
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	

Example

Tirilazad mesylate for better function	# of Patients
From their Phase-3 Trial Sample Size calculation	1130
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	660

Example

Tirilazad mesylate for better function	# of Patients
From their Phase-3 Trial Sample Size calculation	1130
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	660
Number we needed in our Phase-2 Trial to show that their Phase-3 Trial was Futile	

Example

Tirilazad mesylate for better function	# of Patients
From their Phase-3 Trial Sample Size calculation	1130
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	660
Number we needed in our Phase-2 Trial to show that their Phase-3 Trial was Futile	189

3 month data

Phosphenytoin for better function at 3 months	# of Patients

3 month data

Phosphenytoin for better function at 3 months	# of Patients
From their Phase-3 Trial Sample Size calculation	

3 month data

Phosphenytoin for better function at 3 months	# of Patients
From their Phase-3 Trial Sample Size calculation	600

3 month data

Phosphenytoin for better function at 3 months	# of Patients
From their Phase-3 Trial Sample Size calculation	600
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	

3 month data

Phosphenytoin for better function at 3 months	# of Patients
From their Phase-3 Trial Sample Size calculation	600
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	462

3 month data

Phosphenytoin for better function at 3 months	# of Patients
From their Phase-3 Trial Sample Size calculation	600
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	462
Number we needed in our Phase-2 Trial to show that their Phase-3 Trial was Futile	

3 month data

Phosphenytoin for better function at 3 months	# of Patients
From their Phase-3 Trial Sample Size calculation	600
Number they enrolled before they gave up their Phase-3 Trial (P = 0.87)	462
Number we needed in our Phase-2 Trial to show that their Phase-3 Trial was Futile	19

But will AOER-2 >>> HEER-3 ever happen?

- Of the 6 Phase-3 neuroprotective trials we studied:
  - 3 of them flunked their “Phase-2 Futility” trial, and all 3 also flunked their Phase-3 trials.
  - 3 of them passed their “Phase-2 Futility” trials, and
- they included the only trial that was marginally positive in Phase-3

Yes, AOER-2 >>> HEER-3 will happen!

- And you can learn this at a fraction of the blood, tears, toil, sweat, and $$$ of a "negative" Phase-3 trial

Challenges in Designing Small Clinical Trials

Part III

Christopher S. Coffey, PhD

Department of Biostatistics
College of Public Health
University of Iowa

SCT Pre-Conference Workshop
Essentials of Randomized Clinical Trials

VI-1

OVERVIEW

The wonderful land of Asymptopia:

2

OVERVIEW

Small clinical trials present unique challenges.

Small clinical trials are more prone to variability and may only be adequately powered to detect large intervention effects.

As a consequence, the importance of adequate study planning is magnified in small clinical trials.

It is often the case that more time is required during the study planning for small clinical trials.

Critically important to have a true collaboration between clinicians and statisticians.

VI-3

OVERVIEW

Designs of interest in small clinical trials:

- Longitudinal analysis
- Parallel group design
- Crossover design
- Add-On design
- N-of-1 design
- Futility design
- Sequential design
- Ranking/Selection design
- Adaptive designs

Covered by Dr. Wittes

Covered by Dr. Sackett

4

SEQUENTIAL DESIGNS

In a study with a sequential design, participants are sequentially enrolled in the study and randomly assigned a treatment.

The probability of assignment to a given treatment is allowed to change over the course of the study.

Sequential designs are particularly useful for early phase dose-response designs.

5

SEQUENTIAL DESIGNS

Phase I clinical trials typically want to determine some maximum tolerated dose (MTD).

Accurate determination of the MTD is very important since the dose established as the MTD will be used for further testing in later phases.

- Passing on too low of a dose may jeopardize a potentially useful drug
- Passing on too high of a dose puts patients in later phase trials at risk
- Adaptive dose finding methods are well suited for this situation

6

SEQUENTIAL DESIGNS

Random Walk Rule (RWR, biased coin) designs:

- Non-parametric model-based approaches to MTD estimation.
- MTD is treated as a quantile of dose-response distribution, but no underlying paramteric distribution is assumed.
- Instead of applying a deterministic rule, a "biased coin" is flipped after observing each response.
- Algorithm escalate to next dose with probability p where p depends on targeted level of response, θ.

SEQUENTIAL DESIGNS

Strengths of RWR Design:

- Non-parametric
- Allow chosen design points to be distributed unimodally around MTD (or other percentile).
- Provides unified approach targeting any quantile of interest.
- Exact finite distribution theory
- Simple and intuitive to implement
- Simple software in MATLAB has been developed which gives the finite properties of the design (Durham, Flournoy, & Rosenberger, 1997)

8

SEQUENTIAL DESIGNS

Continual Reassessment Method (CRM):

- Originated as a Bayesian method for phase I cancer trials of cytotoxic agents.
- Assumes that probabilities of both efficacy and toxicity increase with increasing dose
- Assumes a particular model (such as logistic function)
- Assignment of doses converges to the MTD.
- See Garrett-Moyer (2006) for an excellent tutorial.

9

SEQUENTIAL DESIGNS

Steps for Implementing CRM:

1) Begin with assumed a priori dose-toxicity curve and a chosen target toxicity rate – assign first patient dose most likely to be associated with target toxicity level.
2) Updated dose-toxicity curve is shifted slightly up or down depending on whether or not first patient experienced a dose-limiting toxicity.
3) Next patient assigned dose closest to target toxicity level based on updated curve.
4) Continue this process until some pre-defined level of certainty achieved or pre-defined stopping criteria met.
5) Once stopping criteria achieved, final dose is selected as the MTD.

10

SEQUENTIAL DESIGNS

For example, consider the following curve:

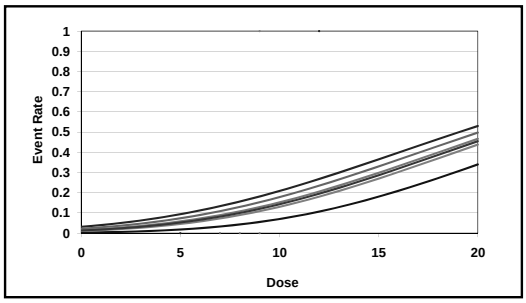
If target level of toxicity is 10%, then dose level 5 would be the optimal starting dose.

Thanks to George Howard (University of Alabama at Birmingham) for Example

11

SEQUENTIAL DESIGNS

An example of how the CRM might work:

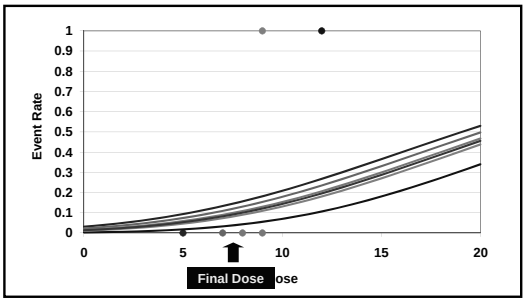


Thanks to George Howard (University of Alabama at Birmingham) for Example

12

SEQUENTIAL DESIGNS

An example of how the CRM might work:



Thanks to George Howard (University of Alabama at Birmingham) for Example

13

SEQUENTIAL DESIGNS

Drawbacks of CRM:

- Mathematical and statistical complexities make it difficult for many clinical investigators to understand.
- Properties must be assessed via simulation.
- Early on, large dose escalations can occur based on little information which may cause more patients to be treated at unsafe doses.
- Dosing first patients at level deemed appropriate by the a priori curve may be worrisome due to uncertainty surrounding this curve.

14

SEQUENTIAL DESIGNS

To address some of the concerns with the original CRM, Several modified CRM approaches have been developed and implemented:

- Always start at the lowest dose level under consideration
- Enroll 2-3 patients in each cohort
- Proceed as a standard 3+3 dose escalation design in the absence of dose-limiting toxicities.
- Any given dose escalation cannot increase by more than one level.

15

BAYESIAN DESIGNS

The CRM is an example of a Bayesian design.

The Bayesian paradigm provides a method for incorporating external information into the design process.

Bayesian methods:

- Quantify prior information and uncertainty into a probability distribution
- Update information as the study progresses

16

BAYESIAN DESIGNS

The parameter of interest is considered as a *random variable* with a *prior* distribution.

Bayesian inference requires a joint distribution of the unknown parameter(s) and the data.

This is usually accomplished by specifying:

- A prior distribution over the parameter space
- A likelihood – the conditional distribution of the data, given the parameters

17

BAYESIAN DESIGNS

Bayesian inference about the parameters uses the posterior distribution – the conditional distribution of the parameter(s) given the data.

As data accumulate, the prior is updated from this posterior distribution.

Modern computational methods can be used to calculate these prior distributions (e.g. R, WinBUGS)

18

BAYESIAN DESIGNS

At any stage, the posterior distribution can be used to draw inferences concerning the parameter.

A credible set may be constructed by:

$$Pr(\rho_L < \rho < \rho_U \mid data) = 0.95$$

Bayesian interpretation can justify statements like:

- "the probability the null hypothesis is true is 0.5"
- "the probability that ρ is between 0.3 and 0.9 is 95%"

19

SELECTION DESIGNS

Ranking and selection procedures are statistical techniques for comparing parameters of multiple (*k*) study populations.

Generally require smaller sample sizes than trials designed to estimate and test treatment effects.

20

SELECTION DESIGNS

Selection designs can be used to:

- Select the treatment with the best response out of *k* potential treatments
- Rank treatments in order of preference
- Rule out poor treatments for further study (Helpful with 'pipeline' problem)

21

CASE STUDY

The QALS study was an adaptive, two-stage, bias-adjusted, randomized, placebo-controlled, double-blind phase II trial (n = 185).

The primary endpoint was the decline over 9 months in ALS Functional Rating Scale-revised (ALSFRSr).

Stage 1 (dose selection, n = 35 per group) compared CoQ10 doses of 1,800 and 2,700 mg/day.

Stage 2 (futility test, n = 75 per group) compared the dose selected in the first stage against placebo.

22

CASE STUDY

The QALS study was a seamless phase I-II study.

The study involved a two-stage, randomized, placebo-controlled, double-blind trial with dose selection and non-superiority (futility) testing.

The primary endpoint was the decline over 9 months of follow-up in the subjects ALS Functional Rating Scale-revised score (ALSFRSr).

The trial was funded through the National Institute of Neurological Disorders and Stroke (NINDS) and conducted at Columbia University.

23

CASE STUDY

QALS Example (Stage One):

- Wish to select the better of two CoQ10 dosages (1800 mg/day or 2700 mg/day) in terms of ALSFRSr change from baseline to nine months
- Assumptions for sample size calculation
  - Standard Deviation = 8.4
  - True Difference in Mean Response = 1.7
  - Desired Probability of Correct Selection = 80%

24

CASE STUDY

QALS Example:

- A sample size of 35 subjects per group will yield 80% probability of a correct dosage selection if the group difference in mean response is at least 1.7 points.

Dose with smaller mean 9 month decline will be selected.

What is type I error (probability of choosing treatment A if it is no better than treatment B)?

25

ADAPTIVE DESIGNS

There may be limited information to guide initial choices for the design of a study.

Since more knowledge will accrue as the study progresses, adaptive designs allow these elements to be reviewed during the trial.

An *adaptive design* allows for changing or modifying the characteristics of a trial based on cumulative information.

26

ADAPTIVE DESIGNS

Recently, there has been considerable research on adaptive designs (ADs).

The rapid proliferation of interest in adaptive designs and inconsistent use of terminology has created confusion about similarities and differences among the various techniques.

For example, the definition of an “adaptive design” itself is a common source of confusion.

27

ADAPTIVE DESIGNS

PhRMA Working Group on Adaptive Designs (2006):

"By adaptive design we refer to a clinical study design that uses accumulating data to modify aspects of the study as it continues, without undermining the validity and integrity of the trial."

"…changes are made by design, and not on an ad hoc basis"

"…not a remedy for inadequate planning."

28

ADAPTIVE DESIGNS

Infinite number of adaptive design possibilities:

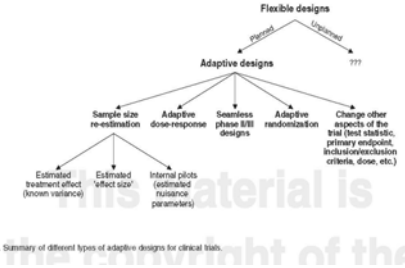


Fig. 1. Summary of different types of adaptive designs for clinical trials.

Source: Coffey CS and Kairalla JA (2008). Adaptive Designs: Progress and Challenges. Drugs in R&D, 9(4): 229-242.

29

ADAPTIVE DESIGNS

Researcher's perspective:

- "I'd like to conduct an adaptive design, but the FDA is unlikely to accept anything out of the ordinary."

FDA Reviewer's perspective:

- "I'd like to encourage use of adaptive designs, but researchers must provide sufficient evidence that the design preserves the operating characteristics of the study (type I error, bias, etc.)."

30

ADAPTIVE DESIGNS

Justifying properties of adaptive designs often requires conducting extensive simulation studies.

The scope of the required simulations is generally non-trivial, particularly for academic researchers relying on NIH-funding.

Researchers must walk tightrope between need to perform extensive simulations and the substantial amount of effort required to perform simulations.

Burton et al. (2006, Stat Med) provide an excellent description of how a protocol for a simulation study should be developed.

31

ADAPTIVE DESIGNS

This issue is exacerbated by fact that this is generally required prior to submission of a grant application for funding.

Many pharmaceutical companies are developing in-house teams primarily responsible for assisting with such simulations.

Greater barriers exist for implementing the same type of infrastructure within the NIH-funded environment.

Need for discussions on how to remove these barriers to increase the use of adaptive designs.

32

GROUP SEQUENTIAL DESIGNS

Group sequential designs are useful for monitoring ongoing study data.

Such designs allow stopping a trial early if it becomes clear that a treatment is superior/inferior.

This minimizes the number of subjects treated with an inferior treatment – minimizes harm.

33

GROUP SEQUENTIAL DESIGNS

Given the definition of an AD, a group sequential design is an AD that allows for premature termination of a trial due to efficacy or futility – based on the results of an interim analysis.

Hence, group sequential designs are one of the most commonly used ADs in clinical trials.

34

SEQUENTIAL DESIGNS

Several approaches have been proposed to adjust for the repeated testing and control the overall study type I error rate:

- Pocock (1977): Same critical value at each interim look
- O'Brien-Fleming (1979): Nominal significance level increases as study progresses
- Haybittle-Peto (1971,1976): Reject H_0 if $|Z|>3$ for each interim look
- Lan-DeMets (1983,1989): Alpha spending function approach

35

SEQUENTIAL DESIGNS

It is important to recognize that ***many of these techniques are based on large sample theory.***

Accordingly, variance estimates are often treated as known quantities during analysis and asymptotically correct Z-distributions (i.e, Z) are used for critical value calculation.

For small clinical trials, these distributions become increasingly inaccurate and may lead to an inflation in the type I error rate.

36

SEQUENTIAL DESIGNS

Pocock (17) suggested a simple, easy approach for small clinical trials.

Use the significance level of Gaussian derived critical values using α-spending functions and sample size to compute critical values from a t-distribution.

For example, if the O-F stopping rule is to stop if $p < 0.002$ ($Z_{crit} = 2.88$) with $n_1 = 100$, then define the critical value as:

$$t_{crit} = t^{-1}(1\text{-}0.002,99) = 2.95$$

37

ADAPTIVE RANDOMIZATION

An *adaptive randomization* design allows modification of randomization schedules during conduct of trial.

With *response adaptive randomization*, the allocation probability for assigning patients to treatment groups is based on responses observed in previous patients.

With *covariate adaptive randomization*, the allocation probability for assigning patients to treatment groups is chosen to reduce covariate imbalance between groups.

38

BIOMARKER ADAPTIVE DESIGN

A *biomarker adaptive design* allows for adaptations based on short-term biomarkers that are thought to inform the treatment effect on a clinical endpoint.

Compared to a gold standard endpoint (i.e., survival), a biomarker can be measured earlier, easier, and more frequently.

In such designs, biomarkers may be used at an interim analysis to assist in decision making.

The final analysis may still be based on the gold standard endpoint.

39

ENRICHMENT DESIGNS

With an *enrichment design*, an additional screening process is included.

A trial is initially planned to consider the entire population.

An initial screening period is used to determine those candidates most likely to benefit from the test agent.

This subgroup of patients are then randomized to receive either the active agent or control.

40

ENRICHMENT DESIGNS

Thus, “enrichment” attempts to choose from all possible patients those that are must suitable for showing a treatment effect.

This can decrease noise and eliminate non treatment-related effects.

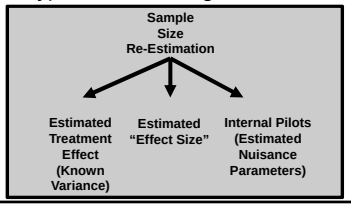
In general, enrichment is about improving efficiency, which should also increase the power of the study (or reduce the required sample size).

41

SAMPLE SIZE RE-ESTIMATION

A *sample size re-estimation* (SSR) design refers to an adaptive design that allows for an adjustment of sample size based on a review of the interim data.

Historically, a great deal of controversy surrounding adaptive designs has been focused around a particular type of SSR design:



42

SAMPLE SIZE RE-ESTIMATION

With *internal pilot* (IP) designs, modifications are based only on re-estimated nuisance parameters.

- nuisance parameter: Parameter not directly related to 'effect' of interest, but must be specified in order to compute sample size or power (e.g. variance, control group event rate, etc.)

IP designs can be used in large RCT's without affecting the type I error rate.

This is not true with small clinical trials and methods are required to adjust for possible bias.
(Stein, 1945; Zucker et al, 1999, Proschan & Wittes, 2000, Kieser & Friede, 2000; Coffey & Muller, 2001)

43

ADAPTIVE DESIGNS

Group Sequential Methods vs. Internal Pilots:

Group Sequential Methods

- Protect against observing a larger or smaller than expected effect during the course of a study

Internal Pilots

- Protect against mis-specification of nuisance parameters at the design stage

An information based monitoring approach allows an internal pilot to be used within a group sequential framework.

44

ADAPTIVE DESIGNS

Mehta & Tsiatis (2001) and Tsiatis (2006) proposed an asymptotically correct method for simultaneous use of group sequential and internal pilot designs in RCT's.

However, asymptotic theory may not control type I error rate in small clinical trials.

Kairalla et al (2010) provided exact theory and methods for single degree of freedom tests.

Current research seeks to utilize this theory to control potential error inflation in small clinical trials.

45

ADAPTIVE SEAMLESS DESIGNS

A *seamless design* combines objectives traditionally addressed in separate trials into a single trial.

This eliminates the time that would have occurred between trials if they had been conducted separately.

Seamless designs also have the potential to provide additional efficiencies in terms of the total required # of patients or long-term follow-up.

46

ADAPTIVE SEAMLESS DESIGNS

An adaptive seamless design combines trials and uses data from patients enrolled before and after the adaptation for the final analysis.

Most interest to date has been with a seamless transition between phase IIb (learning) and phase III (confirming).

However, there are also opportunities for seamless designs in early development (phase I/IIa).

47

ADAPTIVE SEAMLESS DESIGNS

Adaptive seamless designs have the potential to improve the drug development process by reducing the timelines for approval.

Statistical methods must account for the fact that data from the second stage are combined with relevant data from stage one in final analysis.

Hence, extra planning is necessary when implementing an adaptive seamless design protocol.

The potential benefits should be carefully weighed against the challenges of such designs.

48

ADAPTIVE SEAMLESS DESIGNS

Furthermore, at the end of each phase in a traditional approach, the analysis and interpretations are mainly performed by the sponsor.

As a consequence, the “go” or “no go” decision is fully made by the sponsor unless there is a safety concern.

However, with seamless designs, there is a concern that the decisions made have the potential to convey knowledge to observers based on actions taken based on the interim results.

49

ADAPTIVE SEAMLESS DESIGNS

To alleviate the concern, the DSMB may play an important role in the decision-making process between phases.

As a consequence, the roles and responsibilities of the DSMB are beginning to become more complex.

During the case study discussion, I will give the perspective of the DSMB during the conduct of the clinical trial of high dose Coenzyme Q10 in Amyotrophic Lateral Sclerosis (QALS).

50

CASE STUDY

QALS DSMB

Robert Holloway, Jr., M.D. (DSMB Chair)
Associate Professor of Neurology, University of Rochester School of Medicine and Dentistry

Christina Clark, M.B.A.
Foundation for Interdisciplinary Motor Neuron Medicine

Christopher Coffey, Ph.D.
Associate Professor of Biostatistics, University of Alabama at Birmingham

Laurie Gutmann, M.D.
Professor of Neurology, West Virginia University

Theodore R. Holford, Ph.D.
Professor of Biostatistics, Yale University

51

CASE STUDY

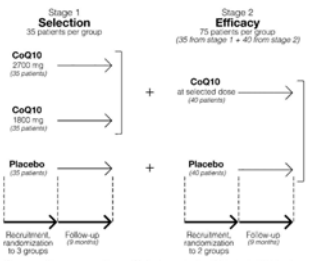


Figure. Two-stage phase II design of the Clinical Trial of High Dose Coenzyme Q10 in ALS (QALS study; total: 185 patients, 105 in stage 1 and 80 in stage 2).

Source: Levy et al, 2006

52

CASE STUDY

The objective of stage one was to use a selection procedure to identify which of two doses of CoQ10 (1800 or 2700 mg/day) was preferred.

For the QALS study, 35 patients per group were enrolled in the first stage.

This provided at least an 80% probability of a correct selection if the true absolute difference in mean ALSFRSr decline over 9 months between the two doses is 1.7 points or more.

53

CASE STUDY

The use of the selection procedure is a bit more complex in a two-stage design.

One cannot simply “pick the winner” since safety issues must be taken into account as well.

If the review is to be done in a blinded manner, there is also a need to develop a plan for dealing with the unexpected situation where the placebo group has the smallest decline.

54

CASE STUDY

As much as possible, a DSMB would need to incorporate these safety issues when creating the DSMB charter.

The selection strategy was described in a written document that was agreed upon both by the DSMB and the investigators.

To completely remove any uncertainty, the DSMB met in-person to approve a final selection strategy prior to viewing any of the summary data by treatment group.

55

CASE STUDY

The DSMB was blinded to the two doses of coQ10 during the DSMB deliberations – informed which group was placebo.

After the deliberations were completed, the DSMB learned which dose had been selected.

	Treatment A (N=35)	Treatment B (N=35)	Treatment C (N=35)
Primary analysis			
Mean (SD) total ALSFRS-R score decline over 9 months *(worst-scenario imputation)*	10.9 (9.3)	9.0 (8.2)	9.8 (8.3)

Selected Dose since no safety concerns existed in this group

56

CASE STUDY

If no adjustment is made, this approach increases the risk of type I error because the final test statistic does not account for the selection of dose in stage one.

A bias correction was developed by the investigators.

The final test statistic incorporated the bias correction to preserve the overall type I error rate.

The performance of this approach was validated via simulation.

57

CASE STUDY

In stage two, a fixed number of additional placebo and CoQ10 subjects at the selected dose were enrolled.

The objective of stage two was to conduct a futility test (superiority null vs. a non-superiority alternative) to compare the preferred dose from stage one against placebo.

58

CASE STUDY

In stage 2, the following futility hypothesis was tested

H_0: CoQ10 reduces mean ALSFRSr decline by at least 20% compared to placebo

$(\mu_\star - 0.8\mu_0 \leq 0)$

versus

H_1: CoQ10 does not reduce mean ALSFRSr by at least 20% compared to placebo – futile to consider in a phase III trial

$(\mu_\star - 0.8\mu_0 > 0)$

where $\mu_\star$ denotes the mean response of the selected dose and μ_0 denotes the mean response of placebo.

59

CASE STUDY

The use of the futility (non-superiority) test meant that the burden of proof was on the data to demonstrate futility.

A non-significant result would mean only that the investigators could not rule out superiority – justifying bringing CoQ10 to phase III for confirmatory testing.

This is appropriate for phase II trials where one deems error of failing to go forward to phase III with a truly superior treatment as arguably more serious than error of going forward with an ineffective treatment.

60

CASE STUDY

Before the viewing the data for the futility analysis, the DSMB again met to ensure that all key parties agreed on the plan.

The protocol for the primary analysis specified that a score of 0 at nine months should be imputed for deceased patients.

There was concern that this might be too extreme, particularly if deaths were imbalanced across groups.

The DSMB requested a secondary analysis that imputes the ALSFRS-R score that corresponds to the 10th percentile for deceased patients.

61

CASE STUDY

This proved extremely important in the interpretation of study results.

In the primary analysis, the null hypothesis was not rejected and non-futility was declared.

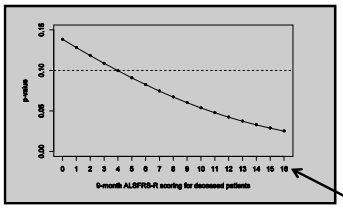
	CoQ10 2,700 mg* (N = 75)	Placebo (N = 75)	t statistic	p-value
Mean (SD) ALSFRS-R score decline over 9 months ***(worst-scenario imputation for LTF)***	8.8 (7.3)	9.4 (8.8)	1.09	0.14

62

CASE STUDY

However, a series of secondary and post-hoc analyses all concluded futility.

As a result, the investigators decided not to pursue a phase III clinical trial of coQ10 in this population.



10th percentile

63

CASE STUDY

The QALS trial was a success – required only 185 participants to:

- Select a preferred dose
- Establish that the cost and effort of undertaking a phase III trial would not be promising

Furthermore, the QALS experience demonstrates that an adaptive design can be implemented within the existing monitoring structures within the NIH setting.

64

CASE STUDY

The complexity of the design underscores the need to maintain a stable DSMB through the entire study.

The adaptive nature of the QALS study required:

- An increased number of face-to-face meetings
- Increased committee memory and collaboration when proceeding from one stage to the next

It would have been very difficult and disruptive to change members in the middle of this trial.

65

SUMMARY

An appropriate study design has sufficient sample size, adequate power, and proper control of bias to allow a meaningful interpretation of the results.

Although small clinical trials pose important limitations, the above issues cannot be ignored.

The majority of methods research for clinical trials is based on large sample theory.

Additional research into innovative designs for small clinical trials is needed.

66