Homework 3 Review

Example 1:

- Outcome: Forced expiratory volume in one second (FEV_1), measured in liters
- Intervention: E = experimental asthma therapy; C = control asthma therapy.
- Design: Use inhaler for 7 days, as needed. Measure the outcome once per participant, at Day 7.
  - 1.1 From the list of outcome types on p.9, what is the likely distribution of this outcome?
  - 1.2 Express the study objective as a two-sided hypothesis using statistical notation (choose from parameters used in class, Lecture 3).
  - 1.3 Calculate the sample size needed to detect a difference between therapies of δ=0.25 L when the standard deviation is σ=0.75 L, the type-1 error rate is α=0.05, the type-2 error rate is β=0.10, and equal allocation to arms is planned ($\gamma = n_E/n_C = 1$). Use the following formula for the control-arm sample size:

$$n_C = \left(\frac{\gamma + 1}{\gamma}\right)\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\delta/\sigma)^2}$$

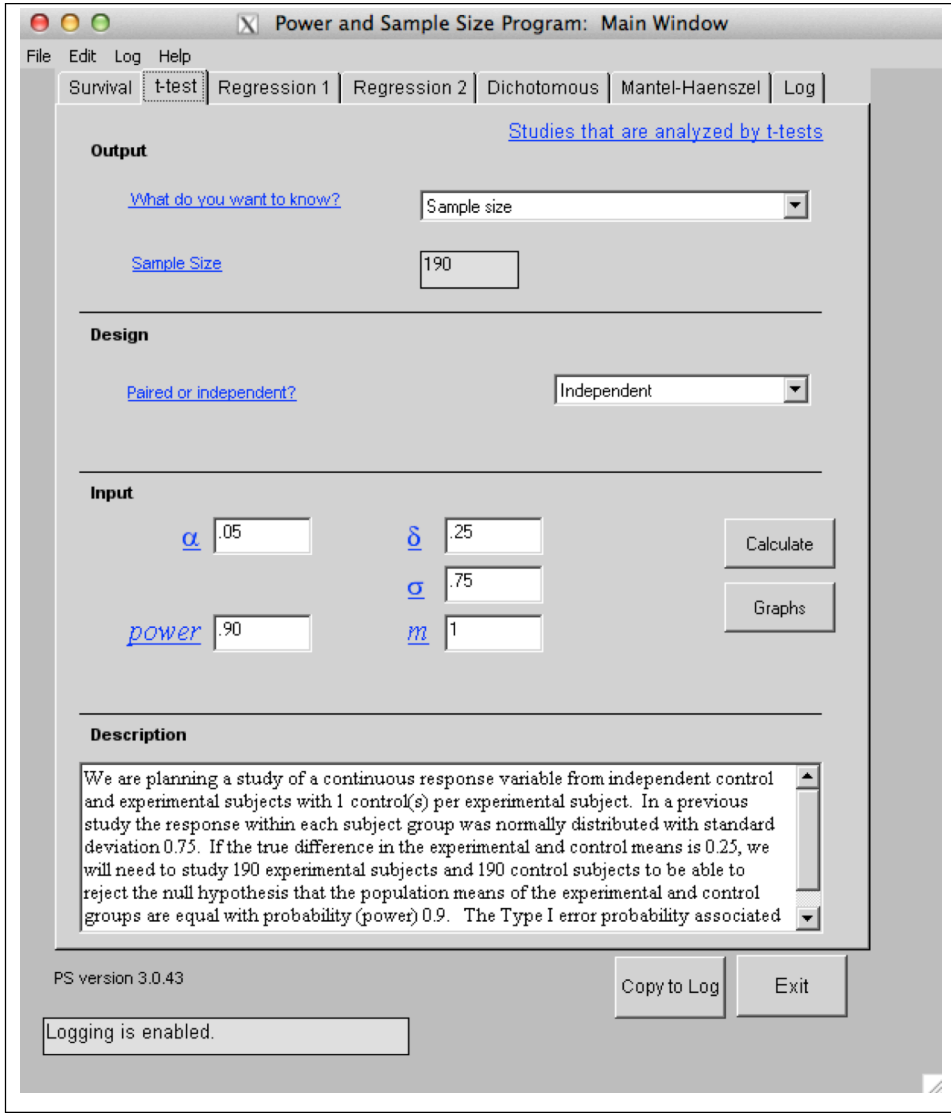
  - 1.4 What is the combined sample size across arms?
  - 1.5 Repeat this calculation using software, such as STATA (twomeans) or SAS (proc power), and provide a screenshot of your output.
  - 1.6 What statistical summaries of the intervention-outcome relationship are most likely of interest to report?
  - 1.7 Suppose you have funding to study $n_C + n_E = 300$ participants. Use algebra to convert the formula above to enable you to calculate the smallest value of δ that can be detected. Is this a clinically relevant value? Should you proceed with the trial?

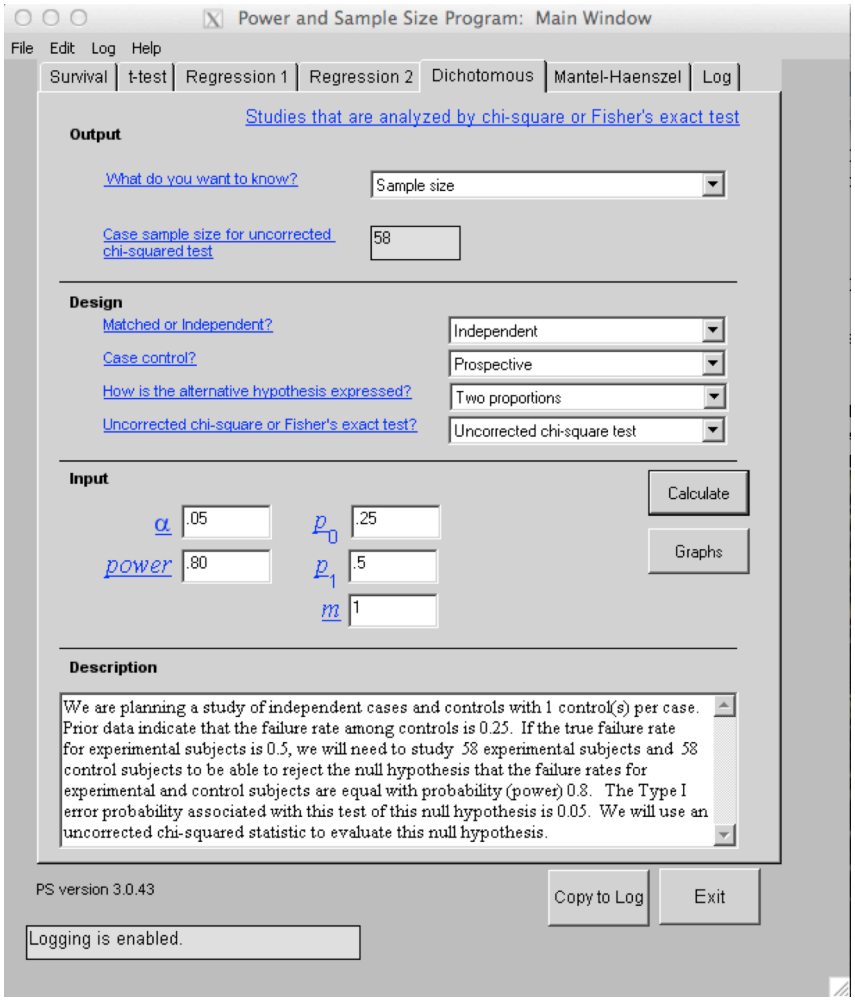
Example 2:

- Outcome: Oral candidiasis lesion present after 3 weeks of asthma inhaler use
- Intervention: E = experimental asthma therapy; C = control asthma therapy
- Design: Use inhaler for 3 weeks, as needed. Measure the outcome once per participant, at Day 21.
  - 2.1 From the list of outcome types on p.9, what is the likely distribution of this outcome?
  - 2.2 Express the study objective as a two-sided hypothesis using statistical notation (choose from parameters used in class, Lecture3).
  - 2.3 If there is no difference between arms, the unbiased estimate of (i) the proportion of participants with a lesion is

$$\bar{\pi} = \left(\frac{\gamma}{\gamma+1}\right)\pi_E + \left(\frac{1}{\gamma+1}\right)\pi_C$$

    and (ii) the standard deviation is $\sigma = \sqrt{\bar{\pi}(1-\bar{\pi})}$. Calculate the sample size needed to detect a difference between therapies of $\delta = \pi_E - \pi_C = 0.25$, when $\pi_C = 0.25$, the type-1 error rate is α=0.05, the type-2 error rate is β=0.20, and equal allocation to arms is planned (γ = 1). Use the sample-size formula above to calculate n_C.
  - 2.4 What is the combined sample size across arms?
  - 2.5 Repeat this calculation using software, such as STATA or SAS, and provide a screenshot of your output.
    - 2.5.1 http://www.ats.ucla.edu/stat/stata/dae/proportionpow.htm
  - 2.6 What statistical summaries of the intervention-outcome relationship are most likely of interest to report?


Power and Sample Size Program: Main Window
File Edit Log Help
Survival t-test Regression 1 Regression 2 Dichotomous Mantel-Haenszel Log
Studies that are analyzed by t-tests
Output
What do you want to know?
Sample size
Sample Size
190
Design
Paired or independent?
Independent
Input
α .05
δ .25
σ .75
power .90
m 1
Calculate
Graphs
Description
We are planning a study of a continuous response variable from independent control and experimental subjects with 1 control(s) per experimental subject. In a previous study the response within each subject group was normally distributed with standard deviation 0.75. If the true difference in the experimental and control means is 0.25, we will need to study 190 experimental subjects and 190 control subjects to be able to reject the null hypothesis that the population means of the experimental and control groups are equal with probability (power) 0.9. The Type I error probability associated
PS version 3.0.43
Copy to Log
Exit
Logging is enabled.


Power and Sample Size Program: Main Window
File Edit Log Help
Survival t-test Regression 1 Regression 2 Dichotomous Mantel-Haenszel Log
Studies that are analyzed by chi-square or Fisher's exact test
Output
What do you want to know?
Sample size
Case sample size for uncorrected chi-squared test
58
Design
Matched or Independent?
Independent
Case control?
Prospective
How is the alternative hypothesis expressed?
Two proportions
Uncorrected chi-square or Fisher's exact test?
Uncorrected chi-square test
Input
α .05
power .80
p0 .25
p1 .5
m 1
Calculate
Graphs
Description
We are planning a study of independent cases and controls with 1 control(s) per case. Prior data indicate that the failure rate among controls is 0.25. If the true failure rate for experimental subjects is 0.5, we will need to study 58 experimental subjects and 58 control subjects to be able to reject the null hypothesis that the failure rates for experimental and control subjects are equal with probability (power) 0.8. The Type I error probability associated with this test of this null hypothesis is 0.05. We will use an uncorrected chi-squared statistic to evaluate this null hypothesis.
PS version 3.0.43
Copy to Log
Exit
Logging is enabled.

Lecture 4

Learning Objectives:

- Identify the parameter of interest in a study design, including its values under H_0 and H_A.

- Understand the distinction between a parameter estimate at the end of the trial and the true parameter – both conceptually and via statistical notation.

- Understand how the statistical problem of inflation of the experiment-wise type-1 error rate is addressed in RCTs that undergo interim analyses and know what is meant by an α spending function.

- Understand 1-to-1 relationship between z-scores and α levels.

- Know why the design is based on the efficacy parameter rather than safety parameter

Lecture 4

Evaluating ongoing data is often the job of the Data and Safety Monitoring Board (DSMB) or Data Monitoring Committee (DMC)

- a committee composed of experts not affiliated with the trial
- advise the sponsor whether to stop the trial once experimental treatment is shown beneficial or harmful, or no longer likely to have clinically relevant effect.

Non-sequential scientific experiments

- To date we've discussed study designs for non-sequential scientific experiments –
  - Usually test a two-sided alternative test of the null hypothesis of no treatment effect

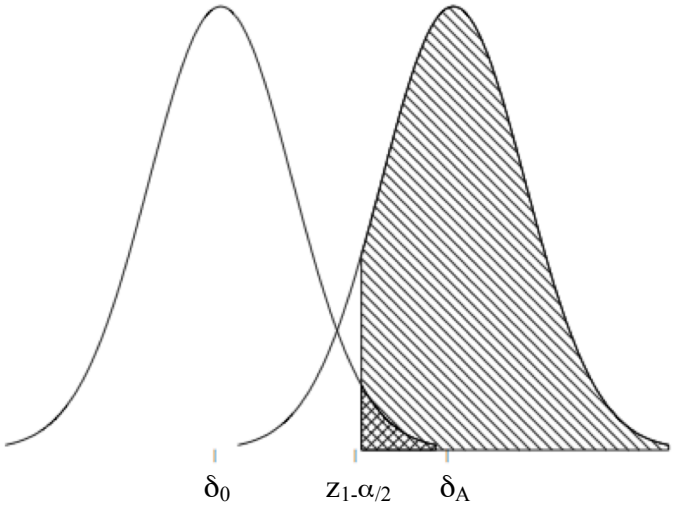
$$H_0: |\delta_0| = 0 \text{ versus } H_A: |\delta_0| \neq 0, \text{ where } \delta_0 = \mu_E - \mu_C$$

  - Conduct a single analysis at the end of a trial – often using a test statistic with a standard normal distribution

$$|z_{obs}| = \frac{|d-\delta_0|}{\sqrt{2\sigma/n}}, \quad \textit{where } d = \bar{x}_E - \bar{x}_C \text{ and } \bar{x}_E = \frac{1}{n_E}\sum x_i.$$

Reject H_0 if $|z_{obs}| > z_{1-\alpha/2}$. Estimate mean (95% CI) for δ.

y



- We've noticed that for trials with staggered accrual, outcomes become available steadily over time
→ *Interim analyses could be performed!*

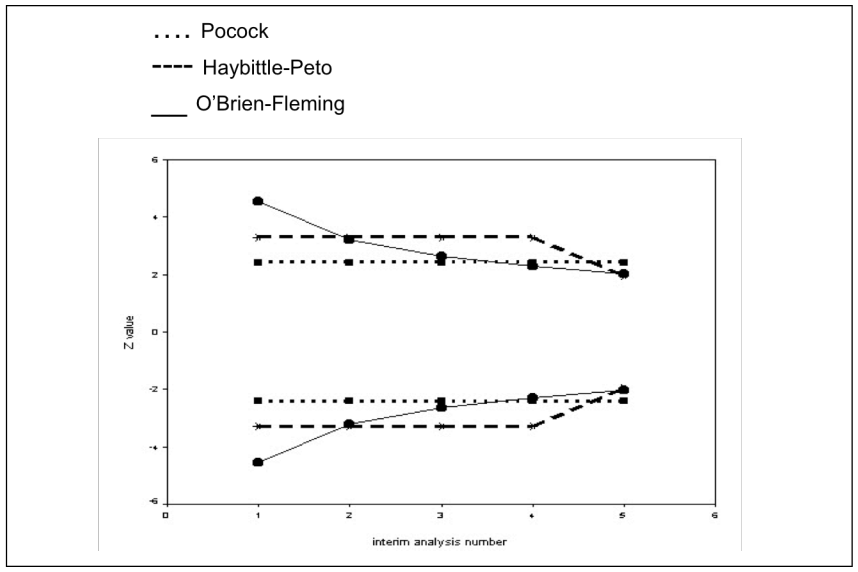
Sequential scientific experiments

- Trade-off arises: Want to stop as early as possible (participant's interests) *without* sacrificing the validity and credibility we worked so hard to ensure – via the trial design, protocol specification (reading for Assignment 3), and trial execution (collective interests). We also want to maintain ability to examine secondary hypotheses.

- Statistical concern: Repeated analyses of accumulating data risks inflation of type-1 error rate, α.
  - *Concept*: A dart near bull's-eye is impressive if based on a single throw, but not if based on many.
  - *Math*: If we divide the accumulating data into k components and test after each set becomes available, with α=0.05, we find that the experiment-wise error rate is larger than α=0.05:

Number of analyses, *k*	1	2	3	4	5	10
Cumulative α	0.05	0.083	0.107	0.126	0.142	0.19

- Boundaries have been defined that allow multiple analyses but protect the cumulative α level and maintain close-to-chosen level (eg, α=0.05) at the end-of-trial analysis.



Some commonly used choices of boundaries:

Boundaries for 5 equally-spaced analyses, at α=0.05

		Pocock		O'Brien-Fleming	
Analysis	Unadjusted α	$Z_{1-\alpha/2}$	Cumulative α	$Z_{1-\alpha/2}$	Cumulative α
1	.05	2.41	.0158	4.56	$5x10^{-6}$
2	.083	2.41	.0275	3.22	.00126
3	.107	2.41	.0365	2.63	.00891
4	.126	2.41	.0439	2.28	.0256
5	.142	2.41	.0500	2.04	.0500

NB: $Z_{1-\alpha/2}$ is sometimes referred to as a *z*-score

α spending function: A plot of the cumulative α levels – "experiment-wise error rate" – in relation to the information fraction in the sample or analysis time governed by a boundary definition.

Common boundary types, with critiques:

** Pocock – not conservative enough late in trial*

** O-F – too conservative early in trial*

** compromise shown*

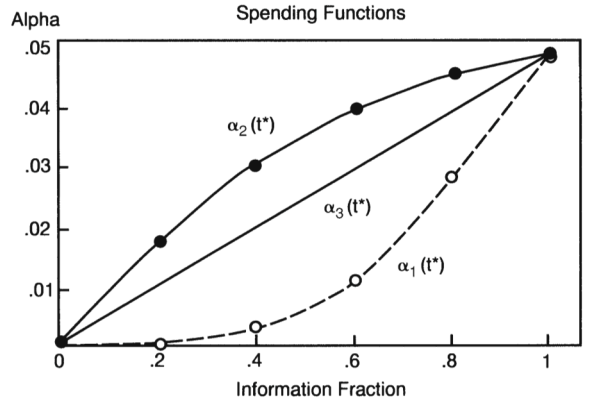
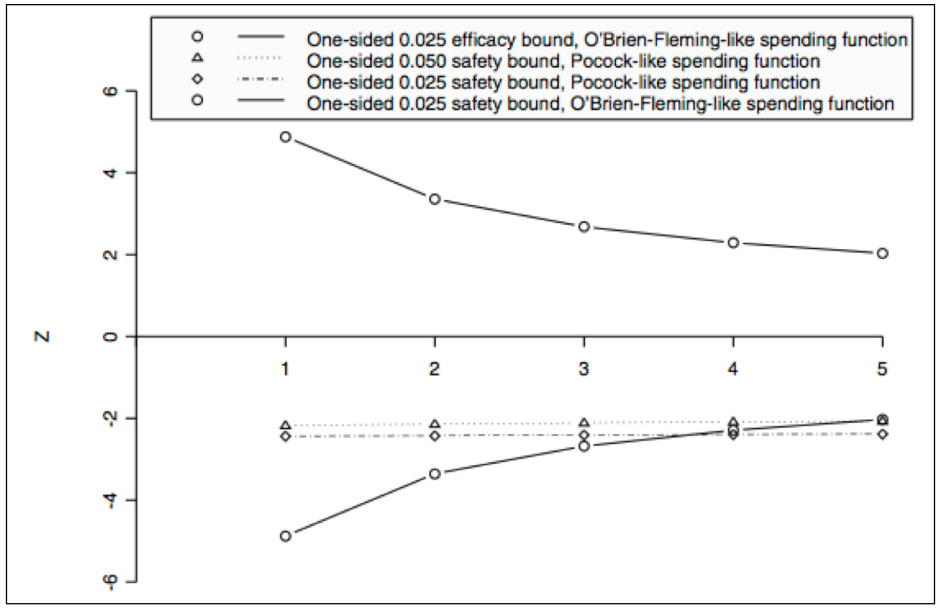


Fig. 16.9 Alpha-spending functions for $K=5$, two-sided $\alpha=0.05$ at information fractions $t^*=0.2$, 0.4, 0.6, 0.8, and 1.0 where $\alpha_1(t^*)\sim$O'Brien–Fleming; $\alpha_2(t^*)\sim$Pocock; $\alpha_3(t^*)\sim$uniform alpha spending functions [211]

Efficacy monitoring: Boundaries for identifying *benefit* and *harm* may differ → "*asymmetric*" …

- The trial seeks to "prove" efficacy: direction of interest is one-sided, using <u>one</u> primary endpoint.
- The trial doesn't seek to "prove" harm: direction of concern is one-sided, based on <u>multiple</u> measures of harm (eg, organ/system classes of adverse events).



* Efficacy & safety outcomes can be the same (eg, mortality) or different. Either way, one can set up one-sided tests so that

- positive values support greater efficacy
- negative values support greater harm

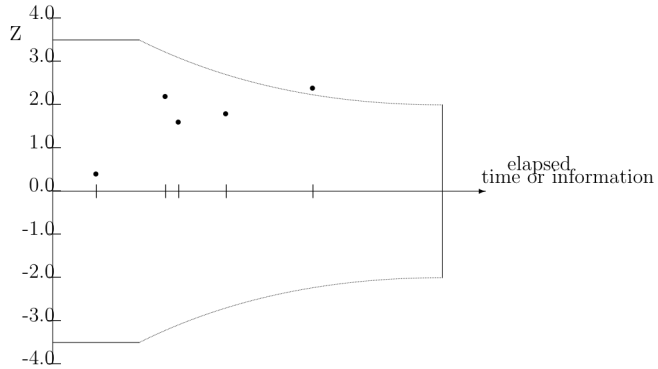
… to make them consistent with concept that benefit and harm are opposites.

* Often use tighter boundaries for safety.
* Note usefulness of conversion to standard normal!

As the trial progresses, calculate evolving test statistic at each of the *j=1, …,k* analyses at compare with the adjusted *z*-scores (see table on page 5 of notes for selected values):

$$z_j = \frac{d_1 - \delta_0}{\sqrt{2\sigma/n_1}} + \frac{d_2 - \delta_0}{\sqrt{2\sigma/n_2}} + \cdots + \frac{d_j - \delta_0}{\sqrt{2\sigma/n_j}}$$

* Equivalently, plot the evolving test statistics against the boundaries:



Sequential outcomes and boundaries for interim standardized test statistics from a clinical trial.

The DSMB uses this information as a guide. ‡

It may recommend stopping an arm or the whole trial or may choose to watch and wait.

Results may demonstrate a difference between arms, or that a relevant difference is unlikely.

In practice, design methods have evolved so that

- equal and/or pre-specified spacing of analysis is unnecessary
- "time" axis can be expressed as calendar time or as "information fraction" (eg, logrank test depends on # events observed)
- standardized test statistics can be truncated to avoid extreme outcomes (early and late in trial)
- software is used for these design and analyses

‡ Note: *Chance* fluctuations in the estimate, *d*, can occur. Large values are more likely to result in termination of the trial than small fluctuations. If only one analysis at the end of the trial, these average out. With multiple analyses, they don't.

Monitoring Safety is Statistically Challenging

- Drug studies (*see ELITE example, next page):*
  - MedDRA = Medical Dictionary for Regulatory Activities is an adverse event classification dictionary used in the USA, European Union, and Japan, endorsed by the International Conference on Harmonisation (ICH)

- Probabilities of events are ill-defined, given multiple outcomes, correlated outcomes, and multiple interim analyses.

- AEs may arise earlier or later than efficacy outcomes.
  - CHD versus breast cancer (WHI Trial).
  - Diabetes versus death (ACCORD Trial).

- AEs vary in frequency and in severity (eg, mild to life-threatening). The threshold for harm depends on the context.
  - The threshold in a prevention study is lower than in a treatment trial.
  - Healthy people may be less willing to tolerate drowsiness than patients with metastatic cancer.

- Adverse events that are rare and unexpected may signal harm → relies heavily on medical judgment.

- AEs may be disease-related or treatment-related → follow-up period may differ from efficacy f'up period.

- Estimates of mean differences (95% CIs) are more useful than tests of differences between arms; reference is expected prevalence in study population.

… Thus A Structured Approach Helps Identify Harm

- A pre-specified safety boundary helps a DSMB make judgments about the endpoint of interest when the data are showing harm.

- It allows earlier termination of the study with a statistically based conclusion of harm.

(next page) → Example: ELITE Trial – reporting of AEs …

Date modified [redacted]

E. CLINICAL/SAFETY/ADVERSE EVENT DATA

Table E.1 Summary of Serious Adverse Events (SAE's)

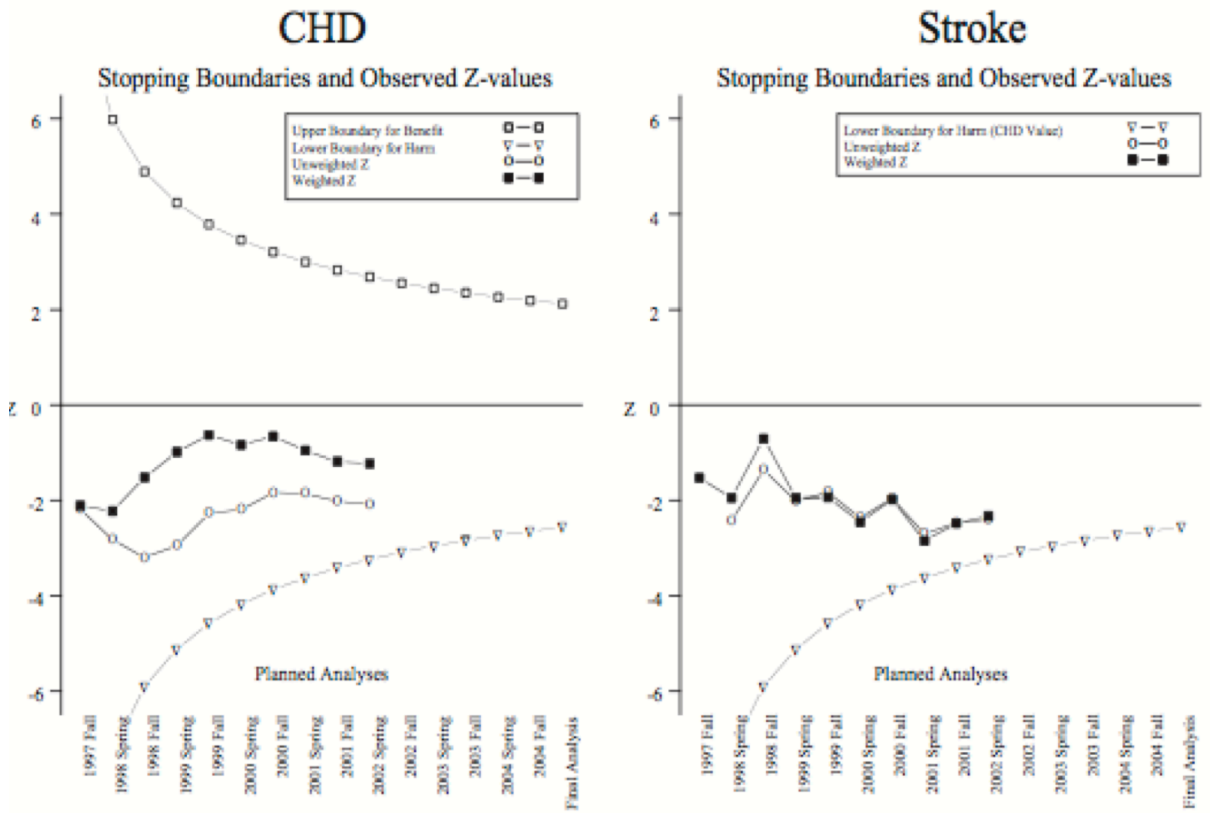
[redacted] Total

	< 6 years # Events/ # Subjects Treatment A	< 6 years # Events/ # Subjects Treatment B	>=10 years # Events/ # Subjects Treatment A	>=10 years # Events/ # Subjects Treatment B	# Events/ #Subjects Treatment A	# Events/ #Subjects Treatment B
ALL SAE'S (# EVENTS / # SUBJECTS)	15/12	8 /8	27/24	29/28	42/36	37/36
BLOOD AND LYMPHATIC SYSTEM DISORDERS	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Aplastic anaemia	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
CARDIAC DISORDERS	1 /1	0 /0	3 /3	4 /4	4 /4	4 /4
Angina unstable	0 /0	0 /0	0 /0	2 /2	0 /0	2 /2
Atrial fibrillation	0 /0	0 /0	1 /1	1 /1	1 /1	1 /1
Myocardial infarction	1 /1	0 /0	1 /1	1 /1	2 /2	1 /1
Pleuropericarditis	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0
EAR AND LABYRINTH DISORDERS	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Vertigo	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
GASTROINTESTINAL DISORDERS	0 /0	0 /0	1 /1	4 /4	1 /1	4 /4
Colitis ischaemic	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Gastrooesophageal reflux disease	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Ileitis	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Lower gastrointestinal haemorrhage	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Upper gastrointestinal haemorrhage	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0
GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS	2 /2	1 /1	0 /0	2 /2	2 /2	3 /3
Death	1 /1	0 /0	0 /0	1 /1	1 /1	1 /1
Non-cardiac chest pain	1 /1	1 /1	0 /0	1 /1	1 /1	2 /2
IMMUNE SYSTEM DISORDERS	0 /0	1 /1	0 /0	0 /0	0 /0	1 /1
Drug hypersensitivity	0 /0	1 /1	0 /0	0 /0	0 /0	1 /1
INFECTIONS AND INFESTATIONS	3 /2	0 /0	2 /2	2 /2	5 /4	2 /2
Bacteraemia	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
Cellulitis	1 /1	0 /0	1 /1	0 /0	2 /2	0 /0
Lobar pneumonia	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0
Pelvic abscess	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Pneumonia	1 /1	0 /0	0 /0	1 /1	1 /1	1 /1
INJURY, POISONING AND PROCEDURAL COMPLICATIONS	3 /3	0 /0	2 /2	1 /1	5 /5	1 /1
Cervical vertebral fracture	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
Femoral neck fracture	0 /0	0 /0	2 /2	1 /1	2 /2	1 /1
Foot fracture	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
Fracture	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
MUSCULOSKELETAL AND CONNECTIVE TISSUE DISORDERS	0 /0	1 /1	0 /0	0 /0	0 /0	1 /1
Systemic lupus erythematosus	0 /0	1 /1	0 /0	0 /0	0 /0	1 /1
NEOPLASMS BENIGN, MALIGNANT AND UNSPECIFIED (INCL CYSTS AND POLYPS)	6 /6	2 /2	13 /11	12 /12	19 /17	14 /14
B-cell lymphoma	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0
Breast cancer	1 /1	1 /1	3 /3	3 /3	4 /4	4 /4
Breast cancer in situ	2 /2	1 /1	2 /2	4 /4	4 /4	5 /5
Colon adenoma	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0
Colorectal cancer	0 /0	0 /0	2 /2	2 /2	2 /2	2 /2
Endometrial cancer	0 /0	0 /0	2 /2	1 /1	2 /2	1 /1
Gastric cancer	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
Glioma	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Malignant peritoneal neoplasm	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
Mycosis fungoides	0 /0	0 /0	0 /0	1 /1	0 /0	1 /1
Ovarian epithelial cancer	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0
Pancreatic carcinoma	1 /1	0 /0	0 /0	0 /0	1 /1	0 /0
Polycythemia Vera	0 /0	0 /0	1 /1	0 /0	1 /1	0 /0

Example: WHI Trial

In a stratum of women with an intact uterus, the WHI Trial compared progesterone/estrogen replacement therapy (PERT) to placebo, hypothesizing that PERT would decrease rates of heart attacks, hip fractures, and colorectal cancer but increase rates of pulmonary embolism, invasive breast cancer, and endometrial cancer. During monitoring, rates of heart attack and stroke (not pre-specified) *were higher in PERT than placebo. The DSMB recommended stopping the study 3 years early because risk outweighed benefits. (Anderson GL, et al (2007) Clinical Trials; Fig 2)*

Monitoring and reporting in the WHI



Unexpected AE (eg, stroke, in WHI Trial) is frequent or serious enough to outweigh intervention benefit.

- Early in the trial, neither benefit nor risk is clear.
- Is the signal real? Does the possible benefit outweigh the risks?

Actions of the DSMB are driven by the perceived risk-benefit ratio.

- It may ask that additional lab tests be performed or data be collected.
- It may recommend stopping.

Actions of the DSMB also are driven by the need to preserve the integrity of the trial

- It may recommend more frequent meetings, possibly by phone, without specifying why. (DSMBs typically meet face-to-face annually.)
- Rather than stop the trial prematurely, it may ask for rapid collection of relevant data: Is the unexpected result real or due to chance?

Ways a DSMB can act that preserve the integrity of a trial include

- Being careful about what information they reveal to investigators.
- Recommending modifications of the Informed Consent document.
- Recommending protocol amendments.
- Requesting different layout or information in DSMB Reports that would be more helpful.

References for this lecture:

- Anderson GL, Kooperberg C, Geller N, Rossouw JE, Pettinger M, Prentice RL. Monitoring and reporting of the Women's Health Initiative randomized hormone therapy trials. Clin Trials. 2007;4(3):207-17.

- Proschan, Michael A, K.K. Gordan Lan, Janet Turk Wittes**.** Statistical monitoring of clinical trials [electronic resource]: a unified approach. New York, NY: Springer, c2006. Chapters 1 and 8.

- Friedman LM, Furburg CD, DeMets DL. Fundamental of Clinical Trials (4th edition). eBook is available online to UCSF community through the UC library.