Biostatistics 226, Winter 2018

Course Overview

Jan 9 – Design of RCTs: How to avoid bias and achieve valid results

Jan 16 – Analyses: Baseline, Efficacy, and Safety

Jan 23 – Monitoring trials for efficacy, futility, and patient safety

Jan 30 – Cluster Randomized Trials

Feb 6 – Precision medicine I: SMART (Adaptive) Trials

Feb 13 – Precision medicine II: N-of-1 Trials

Textbook References

Lawrence Friedman, Curt Furburg, David DeMets. *Fundamental of Clinical Trials, 5th edition*, 2015.
e-copy is available online to UCSF community

ADDED! DeMets, Furberg, Friedman. *Data Monitoring in Clinical Trials, A Case Studies Approach,* 2006.
https://doi-org.ucsf.idm.oclc.org/10.1007/0-387-30107-0

Vittinghoff, Glidden, Shiboski, McCulloch, *Regression Methods in Biostatistics*. 2012.
e-copy: http://www.springer.com/us/book/9781461413523

Review: Jan 16 – Analyses: Baseline, Efficacy, and Safety

HW Readings

- Austin 2010
  - Baseline Data –
    - Altman writes that "performing a significance test to compare baseline variables is to assess the probability of something having occurred by chance when we know that it did occur by chance."
    - Although randomization will, on average, balance covariates between treated and untreated subjects, it need not do so in any particular randomization.
  - Adjustment -- Should adjusted regression model be used when estimating the effect of the treatment on the outcome?
    - To adjust for chance imbalance in prognostically important baseline covariates (selected a priori).
    - To increase precision when estimating the treatment effect.
- Vickers Altman 2013 ChangeScores –
  - Continuous outcomes: Analyzing within-PPT changes over time does not control for baseline imbalance, but repeated measures (ANCOVA) model does.
- Lineberry Berlin 2016 Safety Data --

HW Exercise – Overview of a single RCT selected by each student

- 62% of interventions were therapeutic; 38% were behavioral

Jan 23 – Monitoring trials for efficacy, futility, & patient safety

Table: Outcomes of RCTs and corresponding analysis methods, with examples presented by students

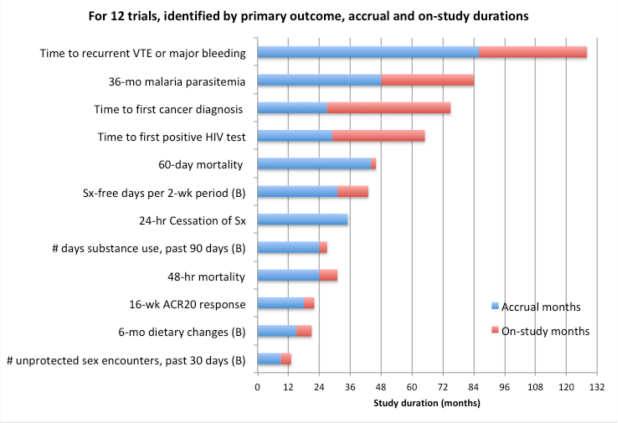
Outcome Type	Distribution	Within- & Between-arm parameters	*Outcomes from students (Behavioral, Therapeutic, Other)*
Continuous	Normal (or Normalizable)	$\delta = \mu_E - \mu_C$	• 6-mo changes in dietary intake • # minutes of physical activity
Binary	Binomial	$\delta = \pi_E - \pi_C$ $RR = \pi_E/\pi_C$ $OR = [\pi_E/(1-\pi_E)] / [\pi_C/(1-\pi_C)]$	• 24-hr cessation of symptoms • 48-dy mortality • 60-dy mortality • 16-wk ACR20 response • 36-mo malaria parasitemia (clustered-R)
(Ordered) Categorical	(Ordinal) Multinomial	$\{\pi_{E1},\pi_{E2},\pi_{E3},\pi_{E4}\}$, $\{\pi_{C1},\pi_{C2},\pi_{C3},\pi_{C4}\}$ Within arm: Categories are mutually exclusive and represent all outcomes, thus probabilities sum to 1.	• *Best clinical response levels = {progressed, stable, partial response, complete response}*
Counts *Can use offset to account for differing lengths of follow-up*	Poisson (or Negative Binomial)	$RR = \lambda_E/\lambda_C$	• # unprotected marital sex encounters, past 30 days • # days with substance use, past 90 days (clustered-R) • # symptom-free days, per 2-wk period (repeated measure)
Binary with censoring*	Exponential (or other)	$HR = \lambda_E/\lambda_C$	• Time to first positive HIV test • Incidence of cancer • Time to recurrent VTE / bleeding

* These binary events typically can only occur once per participant; is an "absorbing" state. Timing of event is key, hence design is longitudinal. Other outcome types are "transient" levels/states that could be assessed once or repeatedly during study follow-up, thus definition of primary outcome should include a time/deadline for occurrence.

Time-frames of RCTs, with examples presented by students

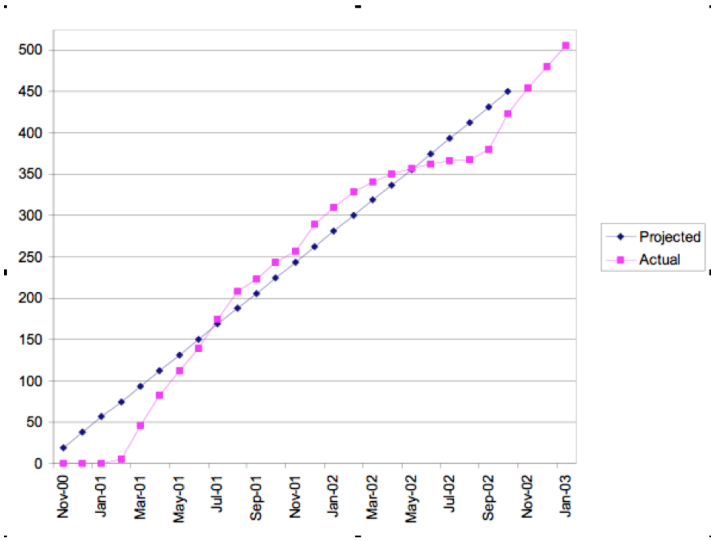
Enrollment / Accrual	*Outcome Assessment*	*Study Design & Outcomes (Behavioral, Therapeutic, Either)*	*Trial Duration*
Non-sequential RCT Simultaneous for all PPTs	*Y* = Single value per PPT, collected simultaneously for all. → No f'up period.	Cross-sectional	All accrual and follow-up on the same day
Sequential RCT Staggered over time	*Y* = Single value per PPT, collected after fixed follow-up period for all. → Fixed f'up duration for all.	Cross-sectional • 9 examples provided by students, excluding 4 listed below	From accrual of initial PPT to assessment of final PPT.
Sequential RCT Staggered over time	*{Y,date}* = Multiple values per PPT, collected over minimum intervals. → All PPTs should achieve minimum f'up; those accrued early could contribute longer f'up.	Longitudinal: Repeated measures • # symptom-free days, per 2-wk period Time-to-event • First positive HIV test • Cancer • Recurrent VTE or major bleeding	From accrual of initial PPT to final assessment of final PPT.

Time-frames of RCT examples presented by students; (B) denotes behavioral outcome



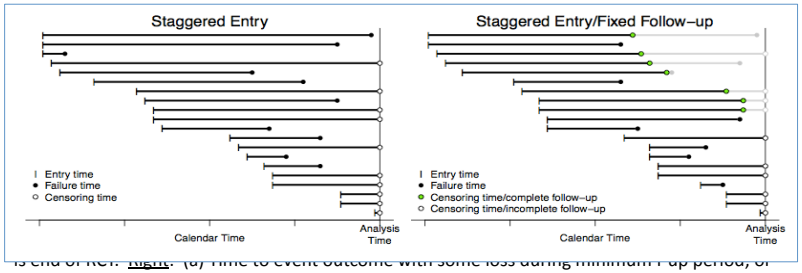
- Accrual period accounts for a large proportion of overall study time. *What are drivers of accrual duration?*
- "On-study" months: exposure and assessment periods can be distinct or overlap. *How decided?*
- Of 4 longest trials, 3 use time-to-event settings (when events are rare and longer follow-up time is used reduce # PPTs) and 1 targeted a long-term public health intervention (with a small intervention effect).
- Many RCTs fail to complete planned accrual. They then do not achieve planned power to detect hypothesized intervention effect.

Example: Trial recruitment, Observed vs. Expected



Sequential RCT: If accrual is staggered over time, outcomes become available over time too.

Plots: Calendar dates of {accrual, last follow-up} for selected participants and analysis date for all.



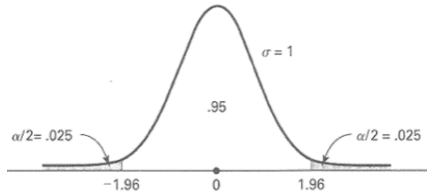
(b) fixed f'up times for all PPTs was planned, and was only partially achieved in some PPTs.

Data Monitoring: Plots show that at any date during the trial, investigators can capture

- the number of participants enrolled and
- the number of events that have occurred.

→ *These trial characteristics can be monitored over time, by arm.*

Review: Recall that a plot of the distribution of d shows a one-to-one relationship between data scale (x-axis) and probability scale (y-axis). *Here, d* represents the observed mean difference between active and control arms (data) and δ_0 represents the H_0 value the mean (parameter).



Standard normal probability density function (mean = 0, standard deviation = 1).
Horizontal axis: d-δ_0. **Vertical axis: Pr{**$|d-\delta_0|$**}. Shaded areas: Pr{**$|d-\delta_0| > c_n$**} =** α **.**

Correspondence between common benchmarks: When $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$ and $z_{1-\alpha} = 1.645$.

A statistical test is conducted via comparison with either the x-axis or an area of the probability distribution.

For 2-sided test:

- On data scale: Reject H_0 if $|z_{obs}| = |d-\delta_0| / (2\sigma/n) > z_{1-\alpha/2}$.
- Equivalent test on probability scale: Reject H_0 if p-value $< \alpha$.

Note that the test assumes H_0 is true; the results (data) reject or do not reject H_0.

At the end of the RCT, we want to be able to draw a firm conclusion: "reject H_0" or "do not reject H_0."

→ However, our conclusion will be based on data, and errors can occur. ←

Statistical test	*Two possible true states*	
	H_A: $\delta \neq 0$	H_0: $\delta = 0$
$\lvert z_{obs} \rvert \neq 0$: Reject H_0 … "positive trial"	*TP* = 1-β	*FP* = α
$\lvert z_{obs} \rvert = 0$: Do not reject H_0 … "negative trial"	*FN* = β	*TN* = 1-α

When H_0 is true: *Specificity* = 1-α = Pr{Do no reject H_0, when H_0 is true}

When H_A is true: *Sensitivity* = 1-β = power = Pr{Reject H_0, when H_0 is false}

Risk of false-positive finding is α.

→ Its consequence: Claim that an intervention is effective when it isn't.

→ Changes standard from Control mode to Experimental mode in broader population.

We work hard to minimize this important error.

Clinical considerations in support of interim monitoring:

- Ethics:
  - Stop early to avoid exposing patients to an inferior treatment (safety or efficacy) – considering both trial participants (cut accrual target / cut treatment period) and future patients
  - Stop early to provide patients with a superior treatment
- Rapidly disseminate results:
  - As early as possible if a treatment change is indicated.
  - Keep co-investigators abreast of progress; maintain motivation.
  - Planning subsequent trial.
  - Natural curiosity: Investigators like to look at the data to see if there is a treatment difference.

For large RCTs, to minimize bias an *Independent DATA MONITORING COMMITTEE* is assembled.

- Experts not affiliated with the trial with no conflicts of interest. Typically unblinded.
  - Clinicians
  - Biostatisticians
  - Other experts

- Role: Advise the sponsor (funder):
  - to ensure the safety of trial participants
  - to recommend stopping the trial once experimental treatment is shown beneficial or harmful, or no longer likely to have clinically relevant efficacy
  - if accrual target is being achieved
  - if design modifications should be made

Resources

- DeMets, Furberg, Friedman textbook, especially Chapter 2 (HW3) – *overview of 29 case studies (therapeutic trials; prevention and treatment)*
- NIH : Overall policy and guidance by Institute/Center on data and safety monitoring.

  https://humansubjects.nih.gov/data_safety

  – Code of Federal Regulations (CFR) Title 45 Part 46

  – FDA Guidance for Industry E6 Good Clinical Practice: Consolidated Guidance :

  > 5.5.2 The sponsor may consider establishing an independent data monitoring committee (IDMC) to assess the progress of a clinical trial, including the safety data and the critical efficacy endpoints at intervals, and to recommend to the sponsor whether to continue, modify, or stop a trial. The IDMC should have written operating procedures and maintain written records of all its meetings.

  ○ NIDDK – addresses statistical issues related to efficacy monitoring → *My role on independent DSMBs, including for NIAID, reflects this description.*

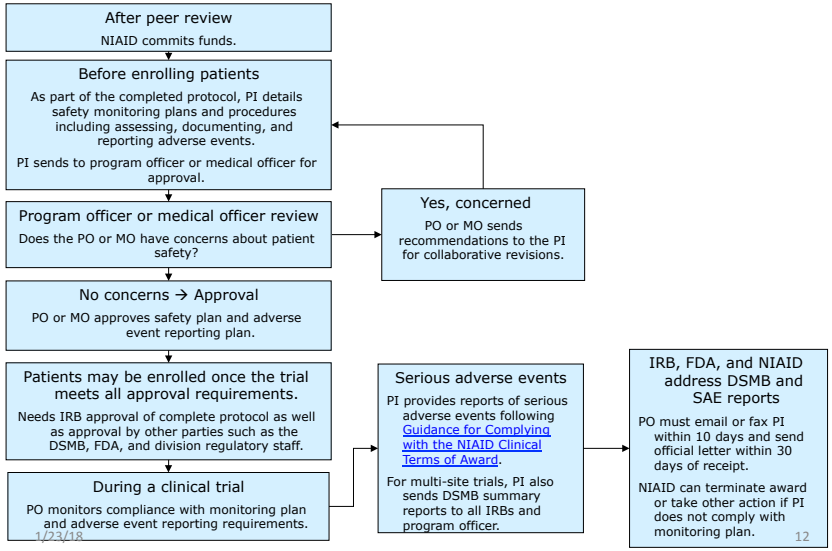
  ○ NCI -- discusses site monitoring for protocol compliance; emphasizes safety over efficacy monitoring → *My role on UCSF HDFCCC DSMB reflects this description.*

  ○ NIAID – see flowchart (next slide) on reporting requirements for safety

Last reviewed October 5, 2010

Relevant Extramural SOPs: Human Subject Resources

NIAID: Clinical Trial Safety Monitoring and Reporting Requirements



Statistical considerations regarding interim monitoring of a sequential RCT:

Objectives:

- Minimize the number of patients on the inferior treatment through early termination of the trial if a difference in treatments is observed.
- Minimize the overall number of patients randomized if no difference in treatments is observed.
- Revise the trial sample size if needed.

Competing objective: Continue to trial end in order to: minimize bias in treatment effect, maximize precision, optimize power to study prognostic factors and secondary outcomes.

Problem: Repeated analyses of accumulating data risks inflation of false-positive (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

Solution: ***Group Sequential Monitoring***

- Analyze results periodically, after each set of 2n patient responses, up to a maximum of K "looks".
- Controls "spending" of the type I error rate across analyses to avoid inflation.
- Extends to "alpha-spending approach."
- *Power:* Sequential monitoring design can be as powerful as the fixed-sample design and feasible to code, communicate, and conduct
- *Flexible:* The timing of analyses and number of responses between analyses should not be restrictive in practice.

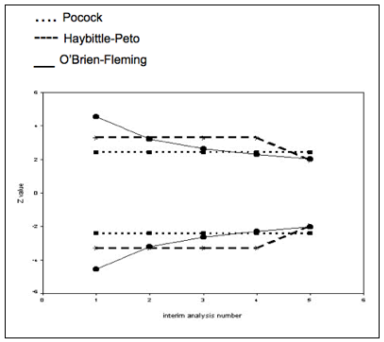
Early approaches:

Boundaries were defined to allow multiple analyses while protecting the cumulative α level and maintaining close-to-chosen α-level at the end-of-trial analysis.

*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	5×10^{-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

* Haybittle-Peto, a possible compromise, is also shown in the figure



Procedure: As the trial progresses, evaluate where it stands relative to these boundaries.

Calculate evolving test statistic at each of the *j=1, ….,k* analyses at compare with the adjusted *z*-scores (see table above for selected values):

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

Two measures were used to compare performance of stopping rules

Example: Suppose α = 0.05, 1– β = 0.90, n_E/n_C =1, and {δ_0 = 0, δ_A = σ}.

1. Maximum Sample size: How many subjects are needed for fixed-sample design (i.e., K = 1):

   $n = (z_{1-\alpha/2} + z_{1-\beta})^2\, 2\sigma^2/(\delta_A - \delta_0)^2 = (1.96 + 1.2816)^2(2) \rightarrow$ 22 patients per arm

   N = 2n = 44 patients in the trial, if n_E/n_C =1 (i.e., equal allocation to arms)

2. Group Sequential Design, K = 5:

   Inflation Factor: IF = 1.23, if using Pocock boundaries,

   IF = 1.03, if using O'Brien-Fleming boundaries

   → *Group sequential designs inflate the maximum sample size relative to fixed sample designs.*
   (Note: The trial is more likely to reach the maximum sample size if H_0 is true.)

   Average Sample Number: ASN = 16.8 → (22-16.8)/22 = -24% , if using Pocock boundaries,

   ASN = 17.2→ (22-17.2)/22 = -22%, if using O'Brien-Fleming boundaries

   → *Group sequential designs decrease the average sample size relative to fixed sample designs.*
   (Note: They do so by stopping earlier than the fixed-sample design. This is more likely if H_A true.)

Critiques:
* *Pocock* – not conservative enough late in trial; increases total N
* *O'Brien-Fleming* – little increase in N but too conservative early in trial
* *Haybittle-Peto*, a possible compromise, is also shown in the figure

Ref: Wang and Tsiatis (Biometrics 1987; 43:193-199), Table 2, provides inflation factors.

Improved approach: Re-express boundaries as smooth α spending functions

The amount of information in the trial is a function of the variability in the outcome variable, the number of patients, and the amount of follow-up time.

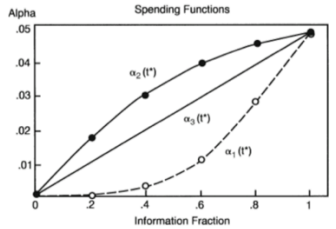
For a given boundary definition, express the cumulative α level – "experiment-wise error rate" – as a function of the *'information fraction'* in the sample.

Lan-DeMets* boundaries and cumulative α for $n_K = 5n$ and $\alpha = .05$

		Pocock		O'Brien-Fleming		Uniform	
Analysis	t^*	$z_{\alpha/2}$	α_2	$z_{\alpha/2}$	α_1	$z_{\alpha/2}$	α_3
1	0.2	±2.438	.0148	±4.877	10^{-6}	±2.576	.0100
2	0.4	±2.427	.026	±3.357	.00079	±2.492	.0200
3	0.6	±2.410	.035	±2.680	.0076	±2.411	.0300
4	0.8	±2.397	.043	±2.290	.024	±2.339	.0400
5	1.0	±2.386	.050	±2.031	.050	±2.275	.0500

* Lan-DeMets notation: Pocock, α_2; O'Brien-Fleming, α_1; Uniform, α_3

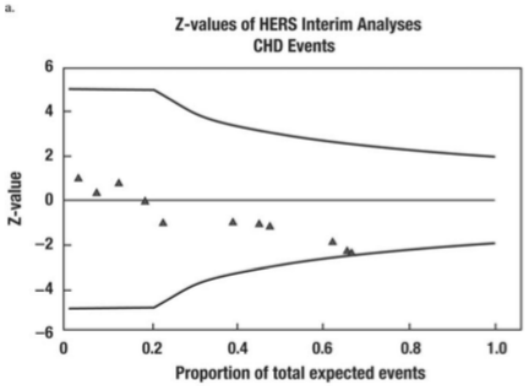
→ Pocock and O-F designs are special cases of this unifying approach.



→ N.B. Never change the boundary function once the trial is underway.

Example, from DFF textbook, Case 17 – HERS Trial:

Figure 1. Lan and DeMets plots of Z-values over the first two-thirds of the study for interim looks at the relative hazard comparing hormone treatment to placebo for CHD deaths. Positive values represent lower event rates in the hormone treatment group.

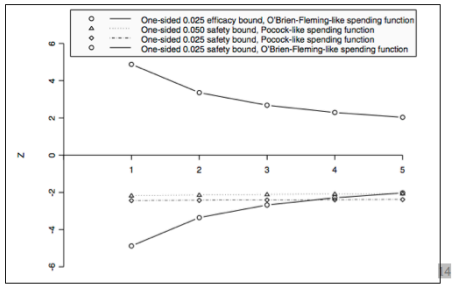


Summary of Group Sequential Design Procedure:

1. Calculate N, given α and β, to detect $\delta_A - \delta_0$ using the fixed-N procedure.
2. Choose the maximum number of looks, K, and the boundary function. Inflate N to reflect the larger maximum number of subjects required by the group sequential design.

Summary of Group Sequential Analysis Procedure:

1. Break the sample size up into K looks. At each look, use the boundaries and decision rules specified by the chosen group sequential design:
   - Calculate the evolving intervention effect estimate and plot it against the boundaries. Stop the trial if the z-score at any look crosses a boundary.
   - Translate the K boundaries into the corresponding significance levels, α_j, j = 1, …, K. Stop the trial if the p-value at any look is smaller than α_j.
2. The confidence interval reported must also reflect the multiple analyses.
   - *Chance* fluctuations in the estimates, z_j, 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.
3. Boundaries for identifying *benefit* and *harm* may differ → be asymmetric; harm may be tighter.



- Vertical axis: Plot outcomes such that benefits are positive and harm are negative. *Note usefulness of conversion to standard normal scale!*
- Horizontal axis can be expressed as percentages of *Information Time*: One can visually see the proportion of planned trial information that's been collected to date.

More comments on efficacy monitoring:

- A formal significance test is only one factor in the decision to continue or stop a trial. DSMB may recommend stopping an arm or the whole trial or may choose to watch and wait.
- Estimated efficacy is just that – an estimate, that is subject to variation. At a particular interim analysis, it may be on a high or low "bounce."
- Stopping for evidence in favor of H_A (efficacy or harm): controls spending of α.
  Stopping for evidence in favor of H_0 (futility): controls spending of β.
- Evidence in support of benefit and harm needn't be symmetrical.
  - The trial seeks to "prove" benefit: direction of interest is one-sided, using one primary endpoint.
  - The trial seek to "detect" harm: direction of concern is one-sided, based on multiple measures of harm (eg, organ/system classes of adverse events).

Monitoring Safety is Statistically Challenging

- Adverse event (AE) definitions must be consistent
  - 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)
- AEs may be disease-related or treatment-related
  - Adverse events that are rare and unexpected may signal harm → relies heavily on medical judgment.
  - AEs vary in frequency and in severity (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.
  - AEs may arise earlier or later than efficacy outcomes → thus f'up periods for efficacy and safety outcomes can differ.
    - CHD versus breast cancer (WHI Trial).
    - Diabetes versus death (ACCORD Trial).

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

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