What to do today (Jan 31, 2023)?

Part III.3.4C Statistical Monitoring: Group Sequential Tests

In particular, about how to determine b_j's at the jth interim reviews for $j = 1, \ldots, K$:

- Pocock (1977) gives clear guidelines for implementation of group sequential experimental designs, attaining type I error and power requirements.
- O'Brien and Fleming (1979) proposes a different class of group sequential tests based on an adaptation of a truncated SPRT.
- Lan and DeMets (1983) show that group sequential methods can be employed when group sizes are unequal and even unpredictable.

The three papers, building on foundation laid by others, together form the starting point for recent methodological research and the basis of current practice in clinical trial design.

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Consider $H_0 : \Delta = 0$ vs $H_1 : \Delta \neq 0$ with type I error rate of α.
Suppose $k = 1, \ldots, K$ interim analyses to be conducted at times $t_1, \ldots, t_K$ with the following procedures:

- Stop and reject H_0 at the first interim analysis if $|T(t_1)| \geq b(t_1)$;
- or stop and reject H_0 at the second interim analysis if $|T(t_1)| < b(t_1)$ but $|T(t_2)| \geq b(t_2)$;
- or . . .
- or stop and reject H_0 at the final analysis if $|T(t_1)| < b(t_1), \ldots, |T(t_{K-1})| < b(t_{K-1})$ and $|T(t_K)| \geq b(t_K)$; otherwise, accept H_0 if $|T(t_1)| < b(t_1), \ldots, |T(t_K)| < b(t_K)$.

To control the type I error (false positive rate), what $b(t_j)$ should be?

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Pocock's Test:

Consider treatment comparison between A and B in variable X: $X_{Ai} \sim N(\mu_A, \sigma^2)$, $X_{Bi} \sim N(\mu_B, \sigma^2)$.
At kth review, km subjects receive each treatment with group size m:

$$Z_k = \frac{1}{\sqrt{2km\sigma^2}}\Big[\sum_{i=1}^{km} X_{Ai} - \sum_{i=1}^{km} X_{Bi}\Big] \sim N(0,1)$$

under H_0 (or approximately). Reject H_0 at stage k if $|Z_k| \geq C_P(K, \alpha)$ for $k = 1, \ldots, K$; otherwise, continue if $k < K$ or accept H_0 if $k = K$.

Part III.3.4C Statistical Monitoring: Group Sequential Tests

The critical value $C_P(K, \alpha)$ is chosen such that

$$P_{\Delta=0}(\text{Reject } H_0 \text{ at analysis k=1,...,or k=K}) = \alpha.$$

Pocock's Test:

$C_P(K, \alpha)$ for two-sided tests

K	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1	2.576	1.960	1.645
2	2.772	2.178	1.875
		... ...	
6	3.023	2.453	2.164

Pocock (1977)

Part III.3.4C Statistical Monitoring: Group Sequential Tests

O'Brien and Fleming's Test:

Reject H_0 at stage k if $|Z_k| \geq C_B(K, \alpha)\sqrt{K/k}$ for $k = 1, \ldots, K$; otherwise, continue if $k < K$ or accept H_0 if $k = K$.
The critical values $c_k = C_B(K, \alpha)\sqrt{K/k}$ are not constant, and $C_B(K, \alpha)$ is chosen to ensure an overall type I error rate of α.

- The critical values are large at early stages than at later stages.
- O'Brien and Fleming's test requires in general a smaller group size m to achieve the same power.

Part III.3.4C Statistical Monitoring: Group Sequential Tests

O'Brien and Fleming's Test:
$C_B(K, \alpha)$ for two-sided tests

K	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1	2.576	1.960	1.645
2	2.580	1.977	1.678
		... ...	
6	2.631	2.053	1.765

O'Brien and Fleming (1979)

Part III.3.4C Statistical Monitoring: Group Sequential Tests

- What if the alternative is one-sided?
- What if the response is not normally distributed?
- What if we can't recruit subjects in groups?
- What if we'd like to choose a way to "spend" the type I error adaptively?
- ... ...

Let's study Lan and DeMets' approach (1983): the Error Spending Approach

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Recall that the group sequential tests of Pocock and O'Brien-Fleming are designed for a fixed number K, of equal sized groups of observations.
$\Longrightarrow$ equally spaced information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of the data at the reviews
e.g. $\mathcal{I}_k = \left[Var(\hat{\Delta}^{(k)})\right]^{-1} = \frac{km}{2\sigma^2}$

- Can we have a flexibility to choose how to "spend" the type I error?
- Can we choose how much to spend the type I error according to the amount of "information" available?

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Spending Type I Error:
Given the maximum number of interim analyses K,

- partition the nominal level α into $\pi_1, \ldots, \pi_K$ such that $\sum_k \pi_k = \alpha$;
- critical values c_k for the standardized statistics Z_k are calculated such that, conditionally on $\mathcal{I}_1, \ldots, \mathcal{I}_k$,

$$P_{H_0}\left(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k\right) = \pi_k$$

for $k = 1, \ldots, K$

The test proceeds according to the familiar stopping rule: rejecting H_0 at review k if $|Z_k| \geq c_k$ for $k \leq K$, or stopping to accept H_0 if it has not been rejected by review K.

Part III.3.4C Statistical Monitoring: Group Sequential Tests

- Slud and Wei (1982): choose the desired π_k's to satisfy the constraint and then determine c_k's.
- the Error Spending Function (Lan and DeMets, 1983): given $\mathcal{I}_{max}$, the target information level,

$$\pi_1 = f(\mathcal{I}_1/\mathcal{I}_{max}),\ \ \pi_k = f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max})\ \ k = 2, 3, ...$$

  - Lan and DeMets (1983): $f(t) = \min(2 - 2\Phi(z_{\alpha/2}/\sqrt{t}), \alpha)$
  - Kim and DeMets (1987): $f(t) = \min(\alpha t^{\rho}, \alpha)$ with $\rho = 1, 1.5$ and 2
  - Jennison and Turnbull (1989, 1990) show with some ρ the corresponding bounaries similar to Pocock's and O'Brien-Fleming's.

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Analysis following a group sequential test:
The stopping occurs at $T = \min\{k : Z_k \notin \mathcal{C}_k\}$. The joint distribution of (T, Z_T) is

$$p(k, z; \theta) = \begin{cases} g_k(z; \theta) & z \notin \mathcal{C}_k \\ 0 & z \in \mathcal{C}_k \end{cases}$$

with $g_k(z; \theta)$ to be obtained recursively.

- Point Estimation:
  e.g. the MLE (sample mean) $\hat{\theta} = Z_T / \sqrt{\mathcal{I}_T}$ is a biased estimator of θ

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Analysis following a group sequential test:

- P-Value: given observed $(T, Z_T) = (k^*, z^*)$,

  $$P_{H_0}\big(\text{obtain (k,z) as extreme or more extrem than } (k^*, z^*)\big)$$

  - the P-value$< \alpha$ if and only if H_0 is rejected
  - the P-value doesn't depend on information levels or group size beyound the observed stopping stage $T = k^*$.

- Confidence Interval: $\{\theta : (T, Z_T) \in A(\theta)\}$

  $$A(\theta) = \big\{(k, z) : (k_l(\theta), z_l(\theta)) \preccurlyeq (k, z) \preccurlyeq (k_u(\theta), z_u(\theta))\big\}$$

Part III.3.4C Statistical Monitoring: Group Sequential Tests

- Commonly used in practice
- **However,** it depends on strict adherence to a precisely specified stopping rule

What if the stopping rule is not followed closely?

In medical setting (*subjective and complex!*), "Statistical tools are ... at best red flags ... and can never be used as hard and fast decision rules." – Coronary Drug Project Research Group (1980)

Part III.3.4C Statistical Monitoring: Group Sequential Tests

Alternative Procedures:

- Bayesian approach (e.g., Berger and Berry, 1988)
  surprising frequentist properties?

- Stochastic curtailment ("conditional power function", e.g., Lan, Simon and Halperin, 1982)
  if the reference test is irrelevant?

- **Repeated Confidence Intervals Approach** (Jennison and Turnbull, 1989)
  see the following ...

Part III.3.4D Statistical Monitoring: Repeated Confidence Intervals

Repeated Confidence Intervals $\{I_k\}$: Jennison and Turnbull (1989)

$$P_\theta(\theta \in I_k, 1 \le k \le K) \ge 1 - \alpha, \quad \theta \in \Theta$$

For example, $k = 1, \ldots, K$,

$$I_k = \left[\bar{X}_{n(k)} - \frac{c_k \sigma_0}{\sqrt{n(k)}}, \bar{X}_{n(k)} + \frac{c_k \sigma_0}{\sqrt{n(k)}} \right],$$

and c_k's are chosen recursively.

The "derived" test: to terminate with rejection of H_0 at kth stage, if I_k fails to contain $\theta = \theta_0$; otherwise, the study continues until stage K.

Part III.3.4D Statistical Monitoring: Repeated Confidence Intervals

Repeated Confidence Intervals Approach

- permits analyses independent of pre-specified stopping rules;
- is able to be used as a guideline for early termination;
- provides an interval estimate at each interim review, "adjusted" for multiple looks. (*Bonus!*)

However, it is on a metric ...

Part III.4.1 Some General Issues in Data Analysis

- Careful analysis requires a major investment in time, effort and expense: it must be done with as much care and concern as any of the design or data-gathering aspects.

- Inappropriate statistical anlayses can introduce bias, result in misleading conclusions, and impair the credibility of the trial.

- The analytic approaches for late phase (III and IV) trials, or the vaious exploratory analysis approaches for early phase (I and II) studies have become quite conventional.

Excluding randomized participants or observed outcomes from analysis and subgrouping on the basis of outcome or other response variables can lead to biased results. Those biases can be of unknown magnitude or direction. – FFDM (2010)

Part III.4.1 Some General Issues in Data Analysis

A. Which Participants Should Be Analyzed?

- to remove from the analysis the participants who didn't fit the eligibility criteria or didn't follow the protocol perfectly, or
- once a participant is randomized, the participant should always be followed and included in the analysis?

The *intention-to-treat* principle: all participants randomized and all events as defined in the protocol should be accounted for in the primary analysis.

- "modified intention-to-treat", or "per protocol", or "on treatment" analyses

  Any deviations from pure intention-to-treat offer the potential for bias and should be avoided or at a minimum presented along with a strict intention-to-treat analysis.

Part III.4.1 Some General Issues in Data Analysis

- *exclusions:* people who are screened as potential participants for a randomized trial but don't meet all of the entry criteria, and thus aren't randomzied

- *withdrawals:* participants who have been randomized but are deliberately excluded from the analysis.
  - reasons for withdrawing: ineligibility, nonadherence, ...
  - omitting participants from analyses can bias the results of the study, and stimulate criticisms of the study

Part III.4.1 Some General Issues in Data Analysis

B. Missing or Poor Quality Data

- data missing for a variety of reasons:
  - subjects were not able to keep their scheduled clinical visits, or to perform/undergo the particular procedures/assessments
  - follow-up was not completed as outlined in the protocol

- approaches to deal with mssing/poor quality data
  - to withdraw participants who have poort data completely: guarantee the validity of the randomization?
  - with missing at random assumption: single/multiple imputation, when an intention-to-treat analysis is not feasible
  - for missclassification, measurement errors, censored data ...

Part III.4.1 Some General Issues in Data Analysis

C. What to Do at the Right Beginning?

To have clear answers to the 5W questions:

- **W**ho are the study subjects?
- **W**hat variables are the collected data (observations) on?
- **W**hen are the data collected?
- **W**here are the data collected?
- **W**hy are the data collected?

Then, to draft a careful analysis plan, which is to be followed, perhaps with updates, in the whole data analysis.

Part III.4.1 Some General Issues in Data Analysis

D. Reporting and Interpreting Results:
The investigators have an obligation to review their study and its findings critically and to present sufficient information so that readers can properly evaluate the trial. –FFDM2010

- Reporting Guidelines: authorship, disclosure of conflict of interest, presentation of data
- Publication Bias:
  timely preparation and submission of the trial results - whether positive, neutral, or negative: negative trials are more likely to remain unpublised than positive trials
- What to Cover: did the trial work as planned? how do the findings compare with results from other studies? what are the clinical implications of the findings?

Part III.4.2 Baseline Assessment

In clinical trials, *baseline* refers to the status of a participant before the start of intervention: *Relevant baseline data should be measured in all study participants.* – FFDM2010.

- Baseline data are measured by interview, questionnarie, physical examination, laboratory tests, etc.
- Uses of baseline data
  - to determine the eligibility of participants,
  - to stratify participants in the treatment allocation,
  - to evaluate baseline comparability,
  - to subgroup participants in the final analysis,
  - to identify subjects who are more likely to develop serious adverse events (pharmacogenetics, a rapidly emerging field in medicine)
  - ...

Part III.4.2 Baseline Assessment

- Analysis of baseline data
  - descriptive analysis of characterisics/demographics of study subjects:
    - natural history analyses
    - numerical description of the study participants
    - preparation for regression analyses with baseline measures as predictors
  - testing for baseline imbalance
    - justification for comparability of groups
    - missing mechanism

Part III.4.3 Assessment of Safety and HR Qof L

A. Analyzing Adverse Events

Adequate attention needs to be paid to the assessment, analysis, and reporting of adverse events to permit valid assessment of potential risks of interventions. – FFDM2010

- Strengths/limitations with clinical trial data:
  the dual goals of a randomized clinical trial: to determine the efficacy and safety of an intervention
  - allow proper hypothesis test, have a proper and balanced control group, reduced potential biases in reporting safety data
  - identification

Part III.4.3 Assessment of Safety and HR Qof L

A. Analyzing Adverse Events (cont'd)

- Types of analyses
  - the presence vs absence of an adverse event:
    cross-sectional, longitudinal
  - the proportion of participants withdrawn from treament due to adverse events
  - quantitative degree (score) of an adverse event
    ordinal categorical data
  - ... ...

Part III.4.3 Assessment of Safety and HR Qof L

B. Analyzing Health-Related QofL

Assessments of the effects of interventions on participants' health-related quality of life is a critical component of many clinical trials, especially ones which involve interventions directed to the primary or secondary prevention of chronic disease. – FFDM2010

- Defining HRQL:
  - primary dimensions: physical functioning, social functiong, psychological functioning, perception of overall QofL, perceptions of health status
  - Additional dimensions: neuro-psychological functioning, personal productivity, sleep disturbance, intimacy and sexual functioning, pain, symptoms
- Quantifying HRQL measures
- Interpretation

What to study next?

Part III. Clinical Trials

- *Part III.1 Introduction*
- *Part III.2 Important Aspects in Study Design*
- *Part III.3 Clinical Trial Conduct*
- **Part III.4 Data Analysis**
  - *III.4.1 Some General Issues*
  - *III.4.2 Baseline Assessment*
  - *III.4.3 Assessment of Safety and Health-Related QoL*
  - **III.4.4 Ecacy Assessment**
- **Example for Clinical Trial: ACTG359**