What to do today (Jan 26, 2023)?

Part I. Introduction

Part II. Epidemiologic Concepts and Designs

Part III. Clinical Trials

Part III.1 Introduction

Part III.2 Important Aspects in Study Design

Part III.3 Clinical Trial Conduct

- **Part III.3.1 Recruitment of Study Participants**
- **Part III.3.2 Data Collection and Quality Control**
- **Part III.3.3 Assessing and Reporting Adverse Events**
- **Part III.3.4 Statistical Monitoring**

Part III.4 Data Analysis

Part IV. Modern Biostatistical Methods

Discussion on Homework 1.

Part III.3.1 Recruitment of Study Participants

Successful recruitment depends on developing a careful plan with multiple strategies, maintaining flexibility, establishing interim goals, preparing to devote the necessary effort and obtaining the same size in a timely fashion. – FFDM2010

- Planning: selecting study sample (recruitment sources); realistic goal
- Conducting and monitoring

Part III.3.2 Data Collection and Quality Control

During all phases of a study, sufficient effort should be spent to ensure that all data critical to the interpretation of the trial, i.e., those relevant to the main questions posed in the protocol, are high quality. – FFDM2010

- minimizing poor quality data
- development of forms, training and certification,...
- quality monitoring, audits

Part III.3.2 Data Collection and Quality Control

Data and Safety Monitoring Board (DSMB): an indpendent group of experts that advises the sponsors and the study investigators.
Members represent the disciplines of

- clinical, laboratory, epidemiology, biostatistics, data management, ethics

Charges of the DSMB include

- Protocol review
- Interim review
- Manuscript review

Part III.3.2 Data Collection and Quality Control

Essential data include the following

- baseline information
- measures of adherence to the study intervention
- concomitant interventions
- primary response variable(s)
- secondary response variables
- other prespecified variables
- adverse events with emphasis on serious events
- signs and symptoms, toxicity (lab) information

Part III.3.3 Assessing and Reporting 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

- clinical trials in the assessment of adverse events: strengths vs limitations
- determinants of adverse events: definitions, classification, ...
- safety monitoring
- analyzing adverse events
- reporting adverse events

Part III.3.4A Statistical Monitoring: Introduction

(Monitoring Response Variables)
After a clinical trial is open, it's required to closely monitor

- its recruitment,
- its data collection,
- its safety, and
- its response (especially later, as the data matured)

During the trial, response variables need to be monitored for early dramatic benefits or potential harmful effects. Preferably, monitoring should be done by a person or group independent of the investigator. Although many techniques are available to assist in monitoring, none of them should be used as the sole basis in the decision to stop or continue the trial. – FFDM2010

Part III.3.4A Statistical Monitoring: Introduction

The study data are monitored (analyzed?) periodically during the course of the trial for ethical and practical considerations.

Early Stopping of Clinical Trials: some reasons

- Serious toxicity or adverse events
- Established benefit
- No trend of interest
- Design of logistical difficulties too serious to fix

Part III.3.4A Statistical Monitoring: Introduction

A common practice: most large scale clinical trials are monitored by an independent DSMB.

- Since there is a lot invested (scientifically, emotionally, financially, etc) in a trial by the investigators who designed and are conducting the trial, they may not be the best suited for deciding whether the clinical trial should be stopped.
- The primary responsibility of DSMB is to ensure the safety and well being of the patients that have enrolled into the trial.
- Statistical issues in the design and analysis of clinical trials which allow the possibility of early stopping.

An important issue in deciding whether a study should be stopped early: a treatment difference during an interim analysis is sufficiently large or small to warrant early termination?
$\Longrightarrow$ **Group Sequential Methods**

Part III.3.4A Statistical Monitoring: Introduction

- Group-sequential methods give rules for early stopping a study based on treatment differences that are observed during interim analyses.
- The term group-sequential refers to the fact that the data are monitored sequentially at a finite number of times (calendar) where a group of new data are collected between the interim monitoring times.
  The new data may come from new patients entering the study or additional information from patients already in the study or a combination of both.

What are group-sequential methods? Anything new to us?

Part III.3.4A Statistical Monitoring: Introduction

- Suppose the study goal is to test $H_0 : \Delta = 0$ vs $H_1 : \Delta \neq 0$
- The test statistic at time t is

$$T(t) = \frac{\hat{\Delta}(t)}{SE(\hat{\Delta}(t))} \sim N(0,1)$$

  under H_0 exactly (or approximately).
- Reject H_0 at time t, if $|T(t)| \geq b(t)$
- What should be the boundary $b(t)$?

For example, $\Delta = \mu_A - \mu_B$ for the treatment difference between A and B in response Y, and $\hat{\Delta}(t) = \bar{Y}_{n_A(t)} - \bar{Y}_{n_B(t)}$. If t=the end of the study, $b(t) = 1.96$ so that $P_{H_0}(|T(t)| \geq 1.96) = 0.05$.

Part III.3.4A Statistical Monitoring: Introduction

What if the data were monitored at K different times, say, $t_1, \ldots, t_K$, and we would want to reject H_0 at the first time t_j such that $|T(t_j)| \geq b(t_j)$?

If choose $b(t_1) = \ldots = b(t_K) = 1.96$?

Effect of multiple looks on tyep I error

K	1	2	3	5	10	20	50	1000	∞
False Positive	.050	.083	.107	.142	.193	.246	.320	.530	1.00

Why?

Part III.3.4A Statistical Monitoring: Introduction

- The event of rejecting H_0 is $\bigcup_{j=1}^{K} \left\{ |T(t_j)| \geq b(t_j) \right\}$.
- The event of accepting H_0 is $\bigcap_{j=1}^{K} \left\{ |T(t_j)| < b(t_j) \right\}$.
- type I error rate:

$$P_{H_0}\left(\bigcup_{j=1}^{K} \left\{ |T(t_j)| \geq 1.96 \right\} \right) > P_{H_0}\left(|T(t_j)| \geq 1.96 \right) = 0.05$$

How to choose the boundaries, $b(t_j)$?

Part III.3.4B Statistical Monitoring: A Short History

The sequential approach has been a natural way to proceed throughout the history of experimentation.

- The formal application started in late 1920s in statistical quality control in manufacturing production.

e.g. Shewhart (1931) introduced control charts for process control.
e.g. Dodge and Romig (1929) defined a two-stage acceptance sampling plan for components which could be tested and classified as effective or defective.

Part III.3.4B Statistical Monitoring: A Short History

The idea of the two-stage sampling was easily generalized to that of multi-stage or multiple sampling plan.

- $\Longrightarrow$ the multi-stage plans developed by the Columbia University Research Group in the World War II
- $\Longrightarrow$ form the basis of the US military standard for acceptance sampling, MIL-STD-105E (1989)

Modern theory of sequential analysis stemmed from the work by Arbram Wald (1947) in US and George Barnard (1946) in Great Britain, who were participating in industrial advisory groups for war production and development from 1943.

Part III.3.4B Statistical Monitoring: A Short History

Consider $X \sim f(x;\theta)$ and test on $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$
Recall that, if the data are iid observations $X_1, \ldots, X_n$, the LRT statistic is

$$T_n = -2 \log \left[\frac{L(\theta_0; x_1, \ldots, x_n)}{L(\hat{\theta}; x_1, \ldots, x_n))} \right] \sim \chi^2(1)$$

approximately under H_0. So that the rejection region is $\left\{(x_1, \ldots, x_n) : T_n > \chi^2_{\alpha/2}(1) or T_n < \chi^2_{1-\alpha/2}(1)\right\}$ to control the type I error at α.

The type II error of the test when $\theta = \theta_1$ is $\beta = P_{\theta=\theta_1}\left(\chi^2_{1-\alpha/2}(1) \leq T_n \leq \chi^2_{\alpha/2}(1)\right)$.

Part III.3.4B Statistical Monitoring: A Short History

Sequential Probability Ratio Test (SPRT) by Wald (1947):

- $X \sim f(x;\theta)$ and $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$
- T_k= the LRT based on sample $X_1, X_2, \ldots, X_k$
- If $T_k \geq b$, accept H_1; if $T_k \leq a$, accept H_0; otherwise, continue to collect X_{K+1}
- Given type I and II error rates, $a \approx \log \frac{\beta}{1-\alpha}$, $b \approx \log \frac{1-\beta}{\alpha}$

Wald and Wolfowitz (1948) proved that SPRT has the theoretical optimal property: it attans the smallest possible expected sample size (average sample number) among all tests with error prob not exceeding α and β.

Part III.3.4B Statistical Monitoring: A Short History

Optimal Stopping Time: the Famous Secretary Problem

However, SPRT is an "open" procedure.

$\Longrightarrow$ a simple modification by Wald: truncated SPRT, to ensure an upper limit on the sample size.

Part III.3.4B Statistical Monitoring: A Short History

- Armitage (1954, 1958, 1975) and Bross (1952, 1958) pioneered the use of sequential methods for comparative clinical trials

  The approaches were fully sequential initially and did not receive widespread acceptance in the medical field: continuouse assessment of study results was often impractical.

- The shift to formal group sequential methods occurred in 1970s.

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 ...

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**
- **Example for Clinical Trial: ACTG359**