PHASE II CLINICAL TRIALS

• Phase II trials are undertaken, in addition to patient care, to assess the anti-tumor “therapeutic efficacy” of a specific treatment regimen (investigational drug).
• They are first human trials looking at “efficacy”
• “Dose” used in a phase II trial is the MTD found in previous phase I trials.
These trials are primarily designed to lead to a recommendation whether or not the treatment regimen deserves further investigation (in future Phase III trials – which are often in “combination”). The agent/drug being tested is recommended if the response is “high” and not recommended if the response is “low”.
The definition of “high/low” is always subjective but, generally, a response of 10% or less is considered “poor/low” and a response 20% or higher is considered “promising/high”.

A response of 50% or more is considered very high, a great success - a rare finding.
EARLY TERMINATION

• Typically, a Phase II trial is designed to have one sample, single stage in which n evaluable patients accrue, are treated, and are then observed for possible response; there are also some two-arm randomized phase II trials.

• Recommendation is made at the end of the trial.

• However, for ethical reasons, the conduct of the trial sometime should allow for early termination if early results are extreme.
If early estimate of response is “low”, say 10% or less, the trial should be stopped so (next) patients could get better treatment - agents being tested are these patients’ last hope; if early estimate of response is “very high”, say 50%-60%, the trial should be stopped so as to proceed to Phase III trial faster (and more patients could benefit from this good treatment)
The termination of a Phase II trial involving a poor agent, due to its low response rate, can be accomplished by a proper study design – for example, the popular two-phase Simon’s design.

If early estimate of response is “very high”, trial should also be stopped so as to proceed to Phase III trial faster (and more patients could benefit from this good or effective treatment). But this cannot be achieved by a design.
Therefore the current conventional approach is to achieve is to reach early termination for poor efficacy by study design and early termination for excellent efficacy by data analysis. One can also handle both types of early termination in a same trial by data analysis but it's more complicated and a bit confusing.
In order to reach a decision “early”, we would need to analyze data more than once; at least once before the (planned) end of the study.
Each analysis leads to a decision by a statistical test of significance. For binary outcomes, such as “response”, the test is “Chi-square”; and one can apply the statistical test once or more than one times. Of course, we are all aware of the “multiple-decision problem”.
THE CONCERNS IN MULTIPLE DECISIONS

The central objective is to essentially preserve the “size”, the “statistical power” (involved in the decision to recommend or not to recommend further investigation), and – as much as possible – the simplicity of single-stage procedure.
The most obvious/serious concern is the “size”. Why?

To perform many tests increases the probability that one or more of the comparisons will result in a Type I error.

**Aim:** We should be prepared to “preserve” this “size” - to keep the probability of type I errors at or below a preset value, say 5%
You can think of simple adjustment like “Bonferroni”, by dividing .05 by the number of tests. But that is not optimal and:

(1) In a two-arm trial, **the most important test is still the last one** (when we have more data), and

(2) if you do often enough, you would not be able to prove anything.
There are more one-arm, open-label phase II trials but it is easier to handle multiple decisions in two-arm trials; let start with those.

RANDOMIZED TWO-ARM TRIALS

(O’Brien and Fleming, 1979)
USUAL SETTING FOR TWO-ARM PHASE II TRIALS

- Randomized clinical trials for comparing two treatments; phases II and III are treated similarly.
- Response is dichotomous and immediate.
- Single-phase, with sample sizes fixed in advance.
- At the end of the trial, compare “success rates” - i.e. proportions- using a formal test of significance based on the usual Pearson Chi-square test.
Ethical concern: Possibility of early termination of the study should “early” results indicate a marked superiority of one treatment over the other. Now we want to build in provision for multiple decisions.
Common application: results indicate a marked superiority of a new treatment over the placebo; trial should be stopped so future patients could get this better treatment
**Aim:** A multiple testing procedure which provides the investigators with opportunity to conduct periodic reviews of the data as they accumulate and, thereby, offers the chance for early termination should one treatment prove superior to the other early on while continuing to use essentially the single-phase decision rule should early termination not occur.
O’BRIEN & FLEMING PROCEDURE

- Investigators plan to “test” N times, including the final comparison at the end of the trial.
- Data are viewed periodically with \( m_1 \) subjects receiving treatment 1 and \( m_2 \) treatment 2 between successive tests; total \( N(m_1 + m_2) \) subjects.
- Want to maintain an overall size \( \alpha \), say \( \alpha = .05 \).
- **Rule:** After the \( n \)th test, \( 1 \leq n \leq N \), the study is terminated and \( H_0 \) is rejected if \( (n/N)X^2 \geq P(N, \alpha) \) where \( X^2 \) is the usual Pearson Chi-square statistic.
RESULTS

Using the theory of “Brownian motion”, O’Brien and Fleming obtained the values for $P(N,\alpha)$ but, more importantly, they concluded that they are approximately the $(1-\alpha)$th percentile of the Chi-square distribution with 1 degree of freedom - almost independent of $N$. 
EXAMPLE #1

• Take N=2 (one “interim” analysis) and \( \alpha = .05 \), \( P(2,.05)= 3.928 \) (“exact”)

• **Rule**: reject the Null Hypothesis after the interim analysis if: \( X^2 \geq (2)(3.928) = 7.86 \) which is equivalent to having the p-value less than .005 and reject the Null Hypothesis after the final analysis if: \( X^2 \geq (1)(3.928) \) which is equivalent to having the p-value less than .045.

\[
H_0 \text{ is rejected if } (n/N)X^2 \geq P(N,\alpha)
\]
A Simple Application of Procedure:

1. Use approximate value (the 95th percentile of the Chi-square distribution with 1 degree of freedom, i.e. 3.84) instead of $P(N,\alpha)$ (i.e. 3.928)

2. Calculate “cut-point for p-value” for the interim analyses (In application of this rule, one would assign .5% to the interim analysis) and subtract them out of the planned size (say 5%) to obtain “cut-point for p-value” for the final analysis.

3. In the example, we can use usual Chi-square tests at .5% and 4.5% respectively.
EXAMPLE #2

- Take $N=3$ (two “interim” analyses) and $\alpha = .05,$

**Rule:** Reject the Null Hypothesis after the first interim analysis if: $X^2 \geq (3)(3.84) = 11.52$ which is equivalent to having the p-value less than .001; after the second if $X^2 \geq (3/2)(3.84) = 5.76$ which is equivalent to having the p-value less than .014

- In application of this rule, one would assign .1% and 1.4% to the interim analyses and 3.5% to the final analysis for an overall 5% size.

$H_0$ is rejected if $(n/N)X2 \geq P(N,\alpha)$
Overall Observation: If you want to look very often (N ≥ 3), early inspections (say, first or second in a series of 4 or 5) are very unlikely to lead to the rejection. For larger N, the threshold for p-value will be reduced substantially. For example, if N=5, We reject the Null Hypothesis after the first interim analysis if:

\[ X^2 \geq (5)(3.84) = 19.20 \]

which is equivalent to having the p-value less than .00002 - almost impossible with only 1/5 of the planned sample size.
The problem and the rule were formulated assigning constant \((m_1 + m_2)\) subjects accrued between successive periodic tests. However, O’Brien and Fleming’s simulations showed that their conclusions/results virtually remain valid in the more general settings. In other words, the only major factors affecting the rule is the number of tests \(N\) and the overall size \(\alpha\). The number of test \(N\) affects the rule but not \(P(N,\alpha)\); in practice most use \(N=2\) or \(3\).
Multiple decision is more popular for two-arm randomized trials but it is possible to make multiple decisions in one-arm, open label trials too. The idea sounds kind of “new” but the methods are not that new; we just form it as an extension of “one-sample test”.

ONE-ARM TRIALS

(Schultz et al., 1973; Fleming, 1982)
USUAL SETTING FOR ONE-ARM PHASE II TRIALS

• Assume that a Phase I trial has been completed from which an MTD has been established.

• The investigator is first asked to specify the largest response rate, $\pi_0$, which if true the treatment does not warrant further investigation.

• Secondly, the investigator is asked to judge the smallest response rate, $\pi_A$, which if true would imply that the treatment has adequate therapeutic efficacy to warrant further investigation.
It seems it would be easier to understand if we set $\pi_0 = \pi_A$; however, the “gap” would also allow for consideration of other factors: cost, ease of applications, safety, etc… It is also similar to the case of monitoring for toxicity or adverse effects (where $\pi_0$ is the baseline rate but the study is only stopped if the toxicity rate exceeds $\pi_A$).
• After acquiring needed components, we state the (one-sided) hypotheses to be tested as:

\( H_0: \pi = \pi_0 \) versus \( H_A: \pi = \pi_A \).

• In addition, we should specify the size (\( \alpha \)) and the power (1-\( \beta \)) from to determine the sample size \( n \):

\[
\alpha = \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) \\
\beta = \Pr(\text{reject } H_A \mid H_A \text{ is true})
\]
SINGLE-STAGE DECISION

• The number of responses “x” is distributed as Binomial Bin(n,π); but if nπ ≥ 10, say, the distribution of Y is approximately normal.

\[ Y(\pi) = \frac{x - n\pi}{\sqrt{n\pi(1 - \pi)}} \]

• Using the normal approximation, the Null Hypothesis H₀ is rejected when

\[ Y(\pi_0) = \frac{x - n\pi_0}{\sqrt{n\pi_0(1 - \pi_0)}} > z_{1-\alpha} \]

\[ x > x_r = n\pi_0 + z_{1-\alpha} \sqrt{n\pi_0(1 - \pi_0)} \]
SINGLE-STAGE DECISION

\[ Y(\pi_0) = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} > z_{1-\alpha} \]

\[ x > x_r = n\pi_0 + z_{1-\alpha} \sqrt{n\pi_0(1-\pi_0)} \]

The rejection rule (rejecting \( H_0 \) when \( y(\pi_0) > z_{1-\alpha} \)) preserves the size \( \alpha \) but may lead to a power somewhat less than (1- \( \beta \)); we need some “continuity correction”
In order to main the power (may result from the normal approximation), the single-stage procedure should more appropriately reject $H_0$ whenever

$$x \geq x_r = \left[ n\pi_0 + z_{1-\alpha} \sqrt{n\pi_0 (1-\pi_0)} \right]^* + 1$$

Where $x^*$ denotes the nearest integer to $x$ (round up and add “1”); **Note: No need to specify the power**
SAMPLE SIZE

If statistical power is specified, we can determine sample size:

\[
n = \left\{ \frac{z_{1-\beta} \sqrt{\pi_A (1-\pi_A)} + z_{1-\alpha} \sqrt{\pi_0 (1-\pi_0)}}{(\pi_A - \pi_0)} \right\}^2
\]
Note that it is relatively simple to specify the size $\alpha$ and $\pi_0$, but more difficult to specify $\pi_A$; For example, one can take as $\pi_0$ the response rate of the “standard treatment”. Then, with a “practical estimate” of $n$, we can see “what it can do”.
FLEMING’S RESULT

Instead of independent specification of power 
(1-β) and Alternative Response Rate \( \pi_A \); Fleming (Biometrics, 1982) prove that, for a 
given \( n \) the usual single-stage procedure 
where one rejects \( \pi_0 \) if \( y(\pi_0) > z_{1-\alpha} \) has power 
(1- \( \alpha \)) against the alternative rate:

\[
\pi_A^+ = \frac{\sqrt{n\pi_0} + z_{1-\alpha}\sqrt{1-\pi_0}}{n + z_{1-\alpha}^2}^2
\]
EXAMPLE A

Let take \( \pi_0 = .20 \), \( \alpha = .05 \), and \( n = 20 \), then the rejection rule “\( y(\pi_0) > z_{1-\alpha} \)” has power 95% against the alternative rate:

\[
\pi_A^+ = \frac{\sqrt{n\pi_0 + z_{1-\alpha}\sqrt{1-\pi_0}}^2}{n + z_{1-\alpha}^2}
\]

\[
= \frac{\sqrt{(20)(.2) + (1.65)\sqrt{.8}}}^2}{20 + (1.65)^2}
\]

\[
= .53
\]
EXAMPLES B

• **Example #1**: $\pi_0 = .20$, $\pi_A = .40$, and $n = 25$; Null Hypothesis is rejected if there are 9 responses or more.

$$x_r = [n\pi_0 + z_{1-\alpha} \sqrt{n\pi_0(1-\pi_0)}]^* + 1$$

$$= [(25)(.2) + (1.65)\sqrt{(25)(.2)(1-.2)}]^* + 1 = 9$$

• **Example #2**: $\pi_0 = .25$, $\pi_A = .50$, and $n = 25$; Null Hypothesis is rejected if there are 11 responses or more.

Note: Use “1.65” instead of “1.96”
Multiple Testing Design

Suppose one decides to perform $N$ tests (usually $N = 2$ or $3$) and to allow $n_i$ patients accrue between the $(i-1)$ th and $i$ th tests, so that $n = n_1 + n_2 + \ldots + n_k$.

Let $x_1, x_2, \ldots, x_k$ represent the number of responses among the $n_1, n_2, \ldots, n_k$ evaluable patients (so that $x = x_1 + x_2 + \ldots + x_k$). In addition, denote the set of (cumulative) rejection points (of $H_0$) by $\{x_{r1}, x_{r2}, \ldots, x_{rk}\}$.
CUMULATIVE SET OF REJECTION POINTS

Note that:

\[ x_{ri} \text{ are determined after } i \text{ tests, } (n_1 + n_2 + \ldots + n_i): \]

\[ x_{ri} = \left[ \left( \sum_{j=1}^{i} n_j \right) \pi_0 + z_{1-\alpha} \sqrt{\left( \sum_{j=1}^{i} n_j \right) \pi_0 (1 - \pi_0)} \right]^* + 1 \]
RULE BY SCHULTZ ET AL.

After test \#n, $1 \leq n \leq N$; Stop and reject $H_0$ if:

$$\sum_{i=1}^{j} x_i \geq x_{rj}$$
EXAMPLE C

• Let take: $\pi_0 = .20$, $\pi_A = .40$, and $n = 30$
• And decide to use $k=3$ (2 interim analyses)
• After the first 10 patients: $x_r = 4$; reject $H_0$ & stop if 4 responses or more
• After the first 20 patients: $x_r = 7$; reject $H_0$ & stop if total 7 responses or more
• After all 30 patients: $x_r = 10$; reject $H_0$ & stop if total 10 responses or more
A PROBLEM & SOLUTION

(1) Fleming observed that the true significance level of a multiple testing procedure can be considerably higher than the nominal level.

(2) To preserve the nominal significance level in multiple testing, Fleming proposed to “inflate” the variance used in determining the rejection points.
FLEMING’S RULE

\[ x_{ri} = \left[ \left( \sum_{j=1}^{i} n_j \right) \pi_0 + z_{1-\alpha} \sqrt{\left( \sum_{j=1}^{i} n_j \right) \pi_0 (1 - \pi_0)} \right]^* + 1 \]
EXAMPLE E

• Let take: $\pi_0 = .20$, $\pi_A = .40$, and $n = 30$
• And decide to use $k=3$ (2 interim analyses)
• After the first 10 patients: $x_r = 4$ by Schultz’s and $x_r = 6$ by Fleming’s
• After the first 20 patients: $x_r = 7$ by Schultz’s and $x_r = 8$ by Fleming’s
• After all 30 patients: $x_r = 10$ unchanged
REFERENCES

• Armitage et al (1969) JRSS series A 132: 235-244
• Herson (1979). Biometrics 35: 775-783
• Lee et al (1979). Cancer Treatment Reports 63: 1721-1726
• Fleming (1982). Biometrics 38: 143-151