

Introduction to Clinical Trials

Lecture 4: Designs for Phase I–III Clinical Trials

Giorgos Bakoyannis, PhD

Associate Professor

Department of Biostatistics and Health Data Science

Indiana University Indianapolis

Notes by Prof. Constantin T. Yiannoutsos, City University of New York

Section 1

Phase I studies

Dose finding designs

In this lecture we will describe designs that attempt to determine the optimal biological dose of a drug (OBD). These are usually undertaken during the Phase I part of the process of drug development.

All dose finding studies are conducted with increasing doses until a predefined clinical outcome is observed.

Goals of Phase I studies

All Phase I studies attempt to investigate toxicity of the drug and establish an optimal biological dose, estimate the pharmacokinetics properties (PK) of the drug and assess its tolerability and feasibility. A number of terms are used in these studies:

- *Dose-ranging studies*

these are designs that specify dose *a priori* along with decision rules for moving from one dose to the next.

- *Dose-finding studies*

These are studies that attempt to detect the optimum dose. They use doses from a continuum rather than a pre-specified set. In contrast to dose ranging studies these studies use a large number of doses.

Everything that we will talk about will be compared or juxtaposed against the classic Phase-I design of oncology studies.

The classic Phase I study of cytotoxic drugs

The basic design found in most oncology studies, but also in many other contexts of cytotoxic medications, proceeds as follows: Once an *a-priori* determine schedule of progressive dose escalation, the following steps are performed:

- Assign three (3) patients to the lowest dose
- If no one experiences the response event of interest (usually a serious toxicity or death) proceed to the next higher dose and expose three more patients to that higher dose
- Continue increasing the dose until one or more patients has the event then proceed as follows:
 - ▶ If only one patient has the event then expose three more patients at the same dose. If two or more out of the six patients have the event of interest then the dose is reduced
- Any dose where two or more toxicities happen is reduced

Different optimality criteria of the OBD

There are various optimality criteria for the OBD.

- *The minimum effective dose (MED)*

For example, in an analgesic, this the optimal dose that completely relieves mild to moderate pain in 90% of recipients

- *Maximum non-toxic dose (MND)*

The optimal dose of an antibiotic may be the dose that causes major side effects in less than 5% of recipients

- *Maximum tolerated dose (MTD)*

In cytotoxic drugs, traditionally the maximum dose that can be tolerated has been thought optimal, such as the dose that yields serious or life-threatening toxicity in no more than 30% of the recipients.

- *Most likely to succeed dose (MLS)*

This is the dose that suppresses 99% of the molecular target activity in at least 90% of patients.

Subjectivity

One of the problems with dose finding studies is the unavoidable subjectivity that creeps in the design. This happens because:

- The dose schedule
- Escalation rules
- Assessment and attribution (to the drug or not) of side effects
- Reactions to unexpected side effects
- The size of the cohort in each dose

The subjectivity results in lack of replication of the results (i.e., two studies of the same drug with different dose and escalation schedules will most likely reach a different result).

An idealized dose finding design

In the ideal world, a dose-finding design would proceed by assigning n_i patients to doses D_i , where $i = 1, 2, 3, \dots$, and then observe the proportion $p_i = r_i/n_i$ of patients that experience the event of interest (usually a toxicity but also efficacy).

A mathematical model is then employed to find the optimal dose. In the mathematical model, invariably, the probability of having the event of interest π_i is associated with a function of the dose D_i as follows:

$$\pi_i = f[\alpha + \beta g(D_i)]$$

then $f^{-1}(\pi_i) = \alpha + \beta g(D_i)$, where f^{-1} is the inverse function of $f(x)$.

Mathematical models in dose-finding studies

The logistic model

For example a logistic-regression model can be applied to the (unknown) population proportion π_i experiencing the event at dose D_i is given by the equation

$$\pi_i = \frac{\exp[\alpha + \beta \log(D_i)]}{1 + \exp[\alpha + \beta \log(D_i)]}$$

so that

$$f^{-1}(\pi_i) = \log \left[\frac{\pi_i}{1 - \pi_i} \right] = \alpha + \beta \log(D_i)$$

Mathematical models in dose-finding studies

The probit model

Another choice is the probit model. In this case the above equation becomes

$$\pi_i = \Phi\{\alpha + \beta \log(D_i)\}$$

where $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2}} e^{-\frac{u^2}{2}} du$ is the standard normal distribution function. The probit model for π_i becomes

$$\Phi^{-1}(\pi_i) = \alpha + \beta \log(D_i)$$

where $\Phi^{-1}(\pi_i)$ is the inverse standard normal distribution.

Representation of the probit and logistic models

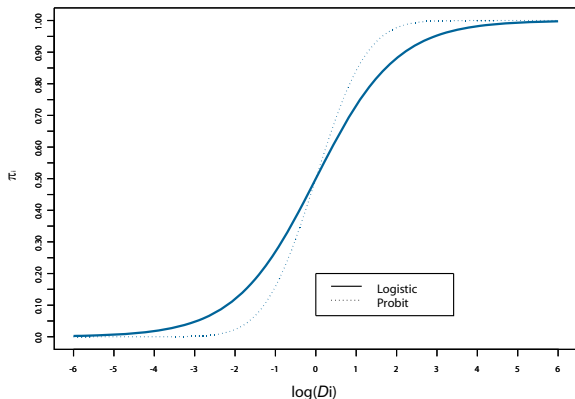


Figure: Pictorial representation of the logistic and probit models of π_i

The two models give similar answers except in the extreme high and low percentiles of π_i .

Why ideal study designs are not normally used

Despite their ability to reliably estimate the optimal dose, ideal designs, such as the one presented above, are not used in clinical trials and are instead used in pre-clinical studies.

There are two reasons that this is the case:

- The ethical limitation of exposing subjects to doses that are expected to cause serious toxicities versus slowly increasing the dose from lower doses that are not expected to result in unacceptable toxicities
- Expose as few patients as possible to the new experimental agent

Operating characteristics (OC) of dose finding designs

We will go over the operating characteristics (OC) of the general dose-escalation design and then apply these to the classic oncology Phase I design. The OC of a study is the probability of stopping before the optimal dose.

To compute the OC of a design, consider the binomial probability of observing k responses out of n subjects as $b(k; n, p)$. Then the cumulative binomial probability

$$P(a < X \leq c) = B(a, c; n, p) = \sum_{i=a+1}^c b(i; n, p)$$

OC of dose finding designs (continued)

If the true probability of the event is p_i , since escalation happens if $r_i \leq u_i$ and de-escalation if $r_i \geq d_i$ while, if $u_i < r_i < d_i$ m_i additional subjects are placed in the same dose, the conditional probability of escalating past the i dose (conditioned on having arrived at this dose) is

$$\underbrace{P_i = B(0, u_i; n_i, p_i)} + \underbrace{\sum_{j=u_i+1}^{d_i-1} b(j; n_i, p_i) B(0, u'_i - j; m_i, p_i)}$$

Responses less than u_i OR more than u_i responses AND less than u'_i responses in expanded cohort

The unconditional stopping probability (considering all conceivable scenarios) and thus the OC of the design is

$$1 - Q_i = 1 - \prod_{k=1}^i \left[B(0, u_k; n_k, p_k) + \sum_{j=u_k+1}^{d_k-1} b(j; n_k, p_k) B(0, u'_k - j; m_k, p_k) \right]$$

OC of the classic Phase I design

In the case of the classic (three-up, three down) design, the previous probability is modified by changing the $u_i = 0$, all $n_i = m_i = 3$ as follows:

$$\underbrace{P_i = b(0; 3, p_i)}_{\text{No responses}} \quad + \quad \underbrace{b(1; 3, p_i)}_{\text{one response in first three patients}} \quad \times \quad \underbrace{b(0, 3, p_i)}_{\text{no responses in next three patients}}$$

The unconditional stopping probability (considering all conceivable scenarios) and thus the OC of the classic design is

$$1 - Q_i = 1 - \prod_{k=1}^i [b(0; 3, p_k) + b(1; 3, p_k)b(0; 3, p_k)]$$

Example of a three-up/three-down design

What is the probability of continuing past the second dose d_2 , assuming that the (unknown) proportion of patients experiencing the event is $p_1 = 0.2$ and $p_2 = 0.3$?

The scenarios under $p = 0.2$ and $p = 0.3$ are

k	$b(k, 3, 0.2)$	$B(k, 3, 0.2)$	$b(k, 3, 0.3)$	$B(k, 3, 0.3)$
0	0.512	0.512	0.343	0.343
1	0.384	0.896	0.441	0.784
2	0.096	0.992	0.189	0.973
3	0.008	1.000	0.027	1.000

Example of a three-up/three-down design (cont'd)

We proceed considering the fact that $b(0; 3, 0.2) = 0.512$ and $b(1, 3, 0.2) = 0.384$ and $b(0; 3, 0.3) = 0.343$ and $b(1, 3, 0.3) = 0.441$ in the two doses respectively. Then the previous calculations become

$$\begin{aligned}1 - Q_i &= 1 - \prod_{k=1}^2 [b(0; 3, p_k) + b(1; 3, p_k)b(0; 3, p_k)] \\ &= 1 - [(0.512 + (0.384)(0.512)) (0.343 + (0.441)(0.343))] \\ &\approx 0.65\end{aligned}$$

There is 35% probability of continuing past the second dose or 65% probability of stopping by the second dose (the OC of the design), with 20% toxicity in the first and 30% in the second dose.

Section 2

Phase II studies

Phase II single-sample designs

We will briefly present below the non-comparative designs for Phase II studies based on single cohorts of subjects.

These designs will be of two main kinds:

- Single-stage designs
- Two-stage or multi-stage designs

A Phase-II example (single-stage design)

Consider the following situation:

In a Phase II non-comparative study (i.e, a small study of one treatment that takes a first “stab” at efficacy assessment), we would like to know whether the true response rate is *at least* as high as 15% (the current standard).

Above that rate, the new therapy would be interesting and worth pursuing further, while, below this rate, we would discontinue development of the experimental therapy.

To perform power and sample size calculations we will have to specify an alternative rate $p_1 > p_0$. We set for this example, $p_1 = 0.40$.

Statistical approach

The null hypothesis to be tested is

$$H_0 : p \leq p_0 = 0.15$$

versus the alternative

$$H_A : p > p_1 = 0.40$$

We would like to maintain $\alpha \leq 0.1$ and the power $1 - \beta \approx 0.80$.

We will thus create a one-sided 90% confidence interval and see whether its lower bound excludes (lies above) $p_0 = 0.15$.

The power is the chance that the confidence interval will lie above p_0 if the true response is p_1 .

Monitoring toxicity is done in an identical manner by reversing the roles of p_0 and p_1 and H_0 and H_A above.

The null and alternative distributions

The null and alternative distributions (assuming $n = 16$ and $p_0 = 0.15$ and $p_A = 0.40$ respectively) are as follows:

k	$p = 0.15$		$p = 0.40$	
	$P(X = k p = 0.15)$	$P(X \leq k p = 0.15)$	$P(X = k p = 0.40)$	$P(X \leq k p = 0.40)$
0	0.07425	0.07425	0.00028	0.00028
1	0.20965	0.28390	0.00301	0.00329
2	0.27748	0.56138	0.01505	0.01834
3	0.22851	0.78989	0.04681	0.06515
4	0.13106	0.92095	0.10142	0.16657
5	0.05551	0.97646	0.16227	0.32884
6	0.01796	0.99441	0.19833	0.52717
7	0.00453	0.99894	0.18889	0.71606
8	0.00090	0.99984	0.14167	0.85773
9	0.00014	0.99998	0.08395	0.94168
10	0.00002	1.00000	0.03918	0.98086
11	0.00000	1.00000	0.01425	0.99510
12	0.00000	1.00000	0.00396	0.99906
13	0.00000	1.00000	0.00081	0.99987
14	0.00000	1.00000	0.00012	0.99999
15	0.00000	1.00000	0.00001	1.00000
16	0.00000	1.00000	0.00000	1.00000

The null and alternative distributions (continued)

The decision of whether to accept or reject the null hypothesis $H_0 : p \leq p_0 = 0.15$ will be based on the number of responses X . For some cutoff value of k , the null hypothesis will be rejected if the number of responses is greater or equal to k . This cutoff needs to fulfill two prerequisites:

- 1 $P(X \geq k | p = p_0) \leq \alpha$, i.e., the probability that we will see more than k responses, if the true response rate is $p_0 = 0.15$ is at most $\alpha = 0.10$. This is the probability of rejecting the null when it is true (false positive rate) or $P(X < k | p = p_0) > 1 - \alpha$.
- 2 $P(X \geq k | p = p_A) \geq 1 - \beta$, i.e., the probability that we will see more than k responses, if the true response rate is $p_A = 0.40$ is at most $1 - \beta = 0.80$. This is the probability of rejecting the null when it is false (power) or $P(X < k | p = p_A) < \beta$.

Performing calculations

From the previous table with $k = 5$,

$$P(X < 5|p = 0.15) = 0.9209 > 1 - \alpha$$

so that

$$P(X \geq 5|p = 0.15) = 1 - 0.9209 = 0.0791 < \alpha$$

and

$$P(X < 5|p = 0.40) = 0.1666 < \beta$$

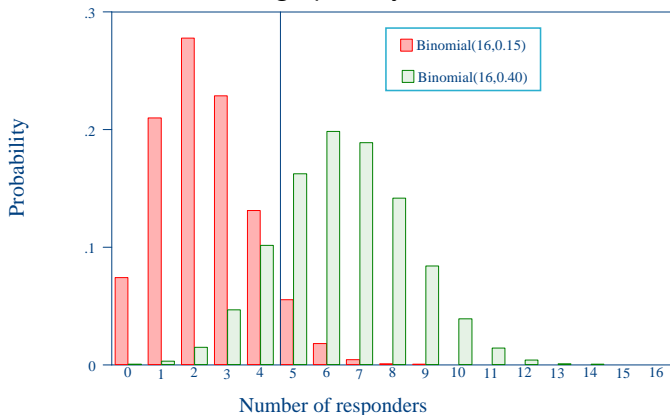
so that

$$P(X \geq 5|p = 0.40) = 1 - 0.1666 = 0.8334 > 1 - \beta$$

This fulfills both the alpha and beta (power) considerations.

Graphical representation

The previous situation is shown graphically below:



Two-stage designs

Consider what would be required in order to be able to insert an interim analysis (a first “stage”) in the study monitoring. The components of a two-stage design are the following:

- Hypotheses to be tested
 - ▶ $H_0 : p \leq p_0$
 - ▶ $H_A : p \geq p_A$
- Type I (α), type II (β) errors and power ($1 - \beta$)
- Sample size (n) and total number of responses (r)
 - ▶ Stage I: Sample size (n_1) and number of responses (r_1)
 - ▶ Stage II: Sample size (n_2) and number of responses (r_2)

Two-stage designs attempt to control the alpha level and power

The Gehan two-stage design

One way to determine the sample size, while controlling the alpha level is to use the following algorithm (Gehan, *J Chron Dis*, 1961):

- **First stage**

- ▶ Enroll n_1 subjects.
- ▶ If at least one subject responds (i.e., if $r_1 \geq 1$) continue to the second stage
- ▶ n_1 is selected so that $P(X = 0 | n_1, p_1) \leq \beta$.

- **Second stage**

Accrue remaining n_2 subjects so that the true response rate p is estimated with the required precision (e.g., based on a maximum allowable half length of the confidence interval)

Implementation of the Gehan two-stage design

First stage

For example, suppose that the null hypothesis is $H_0 : p \leq p_0 = 0.15$ and $H_A : p > p_1 = 0.40$ as before and let $\beta = 0.05$.

After some trial and error, we find that if we accrue $n_1 = 6$ since $P(X = 0 | n = 6, p = 0.40) = 0.0467 < \beta$.

Implementation of the Gehan two-stage design

Second stage

How many additional subjects n_2 will be needed in the second stage in order to estimate the response rate within 20% of the truth?

This translates to selecting a sufficient number of subjects so that the total number $n = n_1 + n_2$ will result in a $(1 - \alpha)\%$ confidence interval with half length $l = z_{\alpha/2} \sqrt{p(1 - p)/n} < 0.20$.

Solving the above inequality with $p = p_A = 0.40$ we end up with

$$n > \frac{z_{1-\alpha/2}^2 p(1 - p)}{0.20^2} = \frac{(1.645)^2 0.40(1 - 0.40)}{0.20^2} = 16.24$$

Thus, $n_2 = 11$ subjects (so that $n = 17$) will be needed to estimate the true response rate $p_A = 0.40$ with precision 20% and $\alpha = 0.10$.

Limitations of the Gehan two-stage design

The key limitations of the Gehan two-stage design are:

- There is no control of type I error in the first stage
- There is no control of power at a desired level

For these reasons the Gehan two-stage design is not very popular nowadays. Instead, Simon's two-stage designs are commonly used today.

The Simon and minimax two-stage designs

The most common type of two stage design unfolds as follows:

- **First stage**

The study is stopped after the first stage for insufficient efficacy if r_1 or less responses out of n_1 total subjects are observed.

The probability of early termination under rate p is

$$PET(p) = P(X \leq r_1 | n, p)$$

- **Second stage**

The study is continued to the second stage if more than r_1 out of n_1 subjects respond during the first stage.

- The study is considered successful (H_0 is rejected) if more than $r = r_1 + r_2$ out of N subjects respond by the end of the second stage.

Components of the usual two-stage designs

Probability to reject the null hypothesis under true response rate p : $R(p)$

$$\begin{aligned} R(p) &= 1 - \left(\underbrace{P(K \leq r_1 | n_1, p)}_{\text{Fail to reject at the first stage}} + \underbrace{P(K \leq r | R_1 > r_1, n_2, p)}_{\text{Fail to reject at the second stage}} \right) \\ &= 1 - \left(B(r_1; n_1, p) + \sum_{k=r_1+1}^{n_1} b(k; n_1, p) B(r - k; n_2, p) \right) \end{aligned}$$

where $b(k; n, p) = P(K = k | n, p)$ and $B(k; n, p) = \sum_{j=1}^k b(j; n, p)$ are the binomial p.d.f. and distribution functions respectively. Then

$$\begin{aligned} \alpha &= R(p_0) \\ 1 - \beta &= R(p_A) \end{aligned}$$

Expected versus maximum sample size

In the two-stage design the maximum sample size is random. The *expected* sample size (also known as *average sample number* or ASN) under rate p is given by the following formula:

$$\begin{aligned}ASN(p) &= n_1 + n_2 \times P(K > r_1 | n_1, p) \\ &= n_1 + n_2 \times (1 - PET(p))\end{aligned}$$

that is, the average sample size equals the number of subjects to be enrolled in the first stage, times the number of subjects enrolled in the second stage probability of continuing to the second stage.

The Simon and *minimax* two-stage designs

Simon (*Cont Clin Trials*, 1985) proposed a two-stage design which minimizes the average sample size $ASN(p_0)$ (i.e., the average sample size under the null hypothesis).

By contrast, the *minimax* design minimizes the maximum sample size n .

Example of Simon's two-stage design

For example, consider the two-stage design with $n_1 = 9$, $r_1 = 1$, $n = 16$ and $r = 4$. Then under the null hypothesis $p = p_0 = 0.15$ we have

n	$P(X = k n = 9, p = 0.15)$	$P(X \leq k n = 9, p = 0.15)$
0	0.23162	0.23162
1	0.36786	0.59948
2	0.25967	0.85915
3	0.10692	0.96607
4	0.02830	0.99437
5	0.00499	0.99937
6	0.00059	0.99995
7	0.00004	1.00000
8	0.00000	1.00000
9	0.00000	1.00000

Simon's two-stage design (continued)

With $r_1 = 1$ and $n_2 = 7$, $r_2 = 3$ (so that $r = 4$) we have

Probability of response $p = 0.15$

Stage I ($n_1 = 9$) responses	Stage II ($n_2 = 7$) responses	Probability	Cum. prob.
0		0.2316	0.2316
1		0.3679	0.5995
2	0	0.0832	0.6827
	1	0.1028	0.7856
	2	0.0544	0.8400
3	0	0.0343	0.8743
	1	0.0423	0.9166
4	0	0.0091	0.9257

Output interpretation: Attained size of the test

The previous output is interpreted as follows:

The probability of not rejecting the null hypothesis $H_0 : p \leq p_0 = 0.15$ when this is true is $1 - \alpha = 0.9257$).

The cumulative probability 0.9257 above is the total probability associated with all scenarios of non-rejection of H_0 . These are:

- **First stage**

The number of responses is $k \leq r_1 = 1$, i.e., $k = 0$, or 1 (this would result in stopping the trial).

- **Second stage**

In order to proceed to the second stage, $k > 1$. In order *not* to reject the null hypothesis, $k \leq r = 4$, i.e., $k = 2, 3, 4$. The probability is given by summing the binomial probabilities of the compatible scenarios.

Thus, the attained size of the test is $\alpha = 1 - 0.9257 = 0.0743$.

Estimation of power

To estimate power we run the same routine with $p = p_A = 0.40$. The results are as follows:

n	$P(X = k n = 9, p = 0.40)$	$P(X \leq k n = 9, p = 0.40)$
0	0.01008	0.01008
1	0.06047	0.07054
2	0.16124	0.23179
3	0.25082	0.48261
4	0.25082	0.73343
5	0.16722	0.90065
6	0.07432	0.97497
7	0.02123	0.99620
8	0.00354	0.99974
9	0.00026	1.00000

With $r_1 = 1$ and $n_2 = 7$, $r_2 = 3$ (so that $r = 4$) we have

Estimation of power (continued)

Probability of response $p = 0.40$

Stage I ($n_1 = 9$) responses	Stage II ($n_2 = 7$) responses	Probability	Cum. prob.
0		0.0101	0.0101
1		0.0605	0.0705
2	0	0.0045	0.0751
	1	0.0211	0.0961
	2	0.0421	0.1383
3	0	0.0070	0.1453
	1	0.0328	0.1780
4	0	0.0070	0.1851

Power of Simon's two-stage design

The previous output is interpreted as follows:

The probability of not rejecting the null hypothesis $H_0 : p \leq p_0 = 0.15$ when this is false (i.e., the Type II of this test) is $\beta = 0.1851$.

The cumulative probability 0.1851 is given in a manner similar to the calculation of α above by summing the binomial probabilities of the compatible scenarios, but with $p = p_A$ in this case.

The projected power of this study is $1 - \beta = 0.8149$.

Expected sample size

From the output above we can calculate that

- **Under the null hypothesis**

$$\begin{aligned}ASN(p_0) &= n_1 + n_2 \times (1 - B(r_1; n_1, p_0)) \\ &= 9 + 7 \times (1 - 0.59948) = 11.803\end{aligned}$$

- **Under the alternative hypothesis**

$$\begin{aligned}ASN(p_A) &= n_1 + n_2 \times (1 - B(r_1; n_1, p_A)) \\ &= 9 + 7 \times (1 - 0.07054) = 15.506\end{aligned}$$

Recall that the one-stage design with the same parameters required $n = 16$ subjects.

The advantage of this design is that the expected sample size of the two-stage design (under the null hypothesis) is significantly lower than the sample size of the comparable one-stage design.

Implementation of the Simon two-stage design

Implementation of the Simon design above in R produces the following output:

```
> # Optimal and Minimax 2-stage Phase II designs
> # Assume p0 = 0.15, p1 = 0.15,
> # type-I error alpha = 0.10 (one-sided), type-II error beta = 0.20
> library(clinfun)
> ph2simon(pu = 0.15, pa = 0.40, ep1 = 0.10, ep2 = 0.20, nmax = 100)
Simon 2-stage Phase II design
Unacceptable response rate: 0.15
Desirable response rate: 0.4
Error rates: alpha = 0.1 ; beta = 0.2
r1 n1 r n EN(p0) PET(p0)
Optimal 1 7 4 18 10.12 0.7166
Minimax 1 9 4 16 11.80 0.5995
```

Optimal two-stage design

The *optimal* two-stage design testing the null hypothesis $H_0 : p_1 \leq p_0$ versus the alternative $H_1 : p_1 > p_0$ involves $n = 18$ total individuals, of whom $n_1 = 7$ are involved in the first stage. If $r_1 \leq 1$ of them respond, then the study will stop. If $r_1 > 1$ individuals respond, then $n_2 = 11$ additional subjects will be enrolled in the second phase.

If $r_2 \leq 4$ respond, then the study will not reject the null hypothesis (i.e., there won't be enough evidence to reject the null hypothesis that the response probability $p_1 \leq p_0 = 0.15$). Otherwise, the null hypothesis will be rejected and the conclusion will be that the response rate $p_1 > p_0 = 0.15$.

The optimal design is “optimal” in the sense that it exposes, on average, $ASN(p_0) = EN_0 = 10.1$ individuals if H_0 is true (i.e., if $p_1 \leq p_0$).

Minimax two-stage design

The *minimax* two-stage design testing the null hypothesis $H_0 : p_1 \leq p_0$ versus the alternative $H_1 : p_1 > p_0$ involves $n = 16$ total individuals, of whom $n_1 = 9$ are involved in the first stage. If $r_1 \leq 1$ of them respond, then the study will stop. If $r_1 > 1$ individuals respond, then $n_2 = 7$ additional subjects will be enrolled in the second phase.

If $r_2 \leq 4$ respond, then the study will not reject the null hypothesis (i.e., there won't be enough evidence to reject the null hypothesis that the response probability $p_1 \leq p_0 = 0.15$). Otherwise, the null hypothesis will be rejected and the conclusion will be that the response rate $p_1 > p_0 = 0.15$.

The minimax design minimizes the total sample n but may result in larger average sample under the null hypothesis, (here $ASN(p_0) = EN_0 = 11.8$ individuals) if H_0 is true compared to the optimal two-stage design.

A graphical software for the design of two-stage design is available at <http://cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx>

The screenshot shows a web browser window with the URL cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx. The page header features the UNC Lineberger Comprehensive Cancer Center logo and the name of the center. Below the header, the user's name, Anastasia Ivanova, Ph.D., is displayed. A navigation menu includes options for 'Continuous monitoring for toxicity', 'Simon's two-stage design', 'Fleming's two-stage design', 'Simon's like design with relaxed futility stopping', 'Two-stage design for ordinal outcomes', 'The Rapid Enrollment Design (RED) for Phase I trials', and 'Other programs'. The main content area is titled 'Simon's Two-Stage design' and explains that the program generates optimal two-stage designs and admissible designs for Phase II single arm clinical trials. It lists two references: 1. Simon R (1989). *Controlled Clinical Trials* 10: 1-10. [Click here to download Simon's \(1989\) article.](#) 2. Jung SH, Lee TY, Kim KM, George S (2004). *Admissible two-stage designs for phase II cancer clinical trials*, *Statistics in Medicine* 23: 561-569. Below the references, there are input fields for 'Type I error rate, α (one-sided):' (0.1), 'Power:' (0.8), 'Response probability of poor drug, p_1 :' (.15), and 'Response probability of good drug, p_2 :' (.4). A 'Calculate' button is located below these fields. At the bottom of the page, there is a footer with the NCI logo, a statement of support from the National Institutes of Health (RO1 CA120062-01A1), contact information for Anastasia Ivanova, and the UNC Lineberger Comprehensive Cancer Center address and phone number.

Efficient designs for phase II oncol... x Simon's Two-Stage design x +

← → cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx web basd two stage simon

UNC
LINEBERGER

UNC Lineberger Comprehensive Cancer Center

Anastasia Ivanova, Ph.D, University of North Carolina at Chapel Hill

Continuous monitoring for toxicity | Simon's two-stage design | Fleming's two-stage design | Simon's like design with relaxed futility stopping

Two-stage design for ordinal outcomes | The Rapid Enrollment Design (RED) for Phase I trials | Other programs

Simon's Two-Stage design

This program generates Simon's optimal two-stage designs (Simon, 1989) and admissible designs from Jung et al. (2004) for Phase II single arm clinical trials.

1. Simon R (1989). *Controlled Clinical Trials* 10: 1-10. [Click here to download Simon's \(1989\) article.](#)
2. Jung SH, Lee TY, Kim KM, George S (2004). *Admissible two-stage designs for phase II cancer clinical trials*, *Statistics in Medicine* 23: 561-569.

Type I error rate, α (one-sided):

Power:

Response probability of poor drug, p_1 :

Response probability of good drug, p_2 :

The development of this software was supported by funds from the National Institutes of Health [RO1 CA120062-01A1]
For comments, questions and suggestions e-mail to Anastasia Ivanova at avanova@bios.unc.edu


NCI
CCC
A Comprehensive Cancer
Center Program of the
National Cancer Institute

UNC Lineberger Comprehensive Cancer Center
Campus Box 7296, Chapel Hill, NC 27599
Appointments: 1-866-869-1856
Copyright © 2009-2013. All rights reserved.

UNC Cancer Hospital
Chapel Hill

Output

The output of the program looks like this:



UNC Lineberger Comprehensive Cancer Center

Anastasia Ivanova, Ph.D, University of North Carolina at Chapel Hill

Continuous monitoring for toxicity | Simon's two-stage design | Fleming's two-stage design | Simon's like design with relaxed futility stopping

Two-stage design for ordinal outcomes | The Rapid Enrollment Design (RED) for Phase I trials | Other programs

Simon's Two-Stage design

This program generates Simon's optimal two-stage designs (Simon, 1989) and admissible designs from Jung et al. (2004) for Phase II single arm clinical trials.

- Simon R (1989). *Controlled Clinical Trials* 10: 1-10. [Click here to download Simon's \(1989\) article.](#)
- Jung SH, Lee TY, Kim KM, George S (2004). *Admissible two-stage designs for phase II cancer clinical trials, Statistics in Medicine* 23: 561-569.

Type I error rate, α (one-sided):

Power:

Response probability of poor drug, p_0 :

Response probability of good drug, p_1 :

n	n_1	r_1	r_2	Type 1 Error	Power	EN_0	Probability of early stopping	Interval for w	Comment
16	9	1	4	0.0743	0.8149	11.8	0.5995	[0.4575, 1]	Minimax
18	7	1	4	0.0880	0.8008	10.1	0.7166	[0,0.4574]	Optimal

Calculated in 3 milliseconds

n is the total number of subjects
 n_1 is the number of subjects accrued during stage 1
 r_1 , if r_1 or fewer responses are observed during stage 1, the trial is stopped early for futility
 r_2 , if r_2 or fewer responses are observed by the end of stage two, then no further investigation of the drug is warranted
 EN_0 is the expected sample size for the trial when response rate is p_0
Interval for w is the set of values w such that the design minimizes $w * n + (1 - w) * EN_0$

Section 3

Phase III studies

Definition and goals

Phase III studies large-scale clinical trials involving one or more experimental therapy against standard therapy or, if standard therapy does not exist, a placebo.

The main objective of Phase III clinical trials is the investigation of the efficacy of the therapy and approval of the experimental therapy by regulatory organizations (e.g., the Food and Drug Administration in the United States).

Phase III Studies

Randomized Trials

Possible designs:

- ① Direct tests (usually when no known treatment for this disease)
 - ▶ Z versus placebo (P)
- ② Active control W (standard treatment for this disease)
 - ▶ Z versus W
- ③ Fairly direct tests of value of adding Z to a standard therapy Y
 - ▶ $Y + Z$ versus $Y + P$
- ④ Amount, Timing of Z
 - ▶ Low dose versus high dose of Z
 - ▶ Z initially versus Z delayed
 - ▶ Z intermittently versus Z continuously

Designs vary as knowledge is gained:

AIDS therapy

Early studies compared single drug, e.g., AZT, to placebo. Once AZT was demonstrated to be effective, studies compared other therapies to AZT, or timing of AZT (based on CD4 count) and value of switching.

More recent studies have established the efficacy of combination therapies (HAART) and have used factorial designs to test two questions (e.g., an antiretroviral agent and a prophylactic agent for opportunistic infection) or a drug combination at the same time.

Phase III clinical trials

Randomization and blinding

The main issue driving the design of any clinical trial is that the treatment groups must ideally be different only in terms of the intervention.

This is accomplished by the following procedures:

- The assignment of the treatments to patients must be performed *randomly*
- A further design “wrinkle” is to “blind” the patients (single-blind studies) and the investigators (double-blind studies) to the true assigned treatment
- Assuming a large sample size, random treatment assignment ensures that known or unknown confounding factors will be equally distributed into the groups under comparison: Thus, any difference in patient outcome can be causally associated with differences in the treatments

Phase III clinical trials

Broad design categories

- **Superiority trials**

Superiority trials compare a new therapy against the established therapy to develop more effective treatments

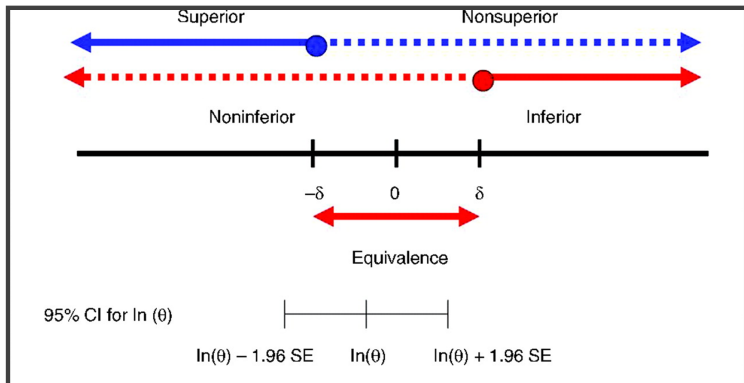
- **Equivalence trials**

Equivalence trials address the question whether a new therapy is as good as the established therapy. Equivalence trials attempt to establish the effectiveness of treatments with favorable toxicity profile compared to the established therapy

- **Non-inferiority trials**

Non-inferiority trials address the question whether a new therapy is not inferior compared to the established therapy. Just as with equivalence trials, non-inferiority trials attempt to establish the effectiveness of treatments with favorable toxicity profile compared to the established therapy.

Characteristics of various designs



Benny Chung-Ying Zee JCO 2006;24:1026-1028

©2006 by American Society of Clinical Oncology

JOURNAL of CLINICAL ONCOLOGY ASCO

Superiority trial example

Adjuvant chemotherapy for HER-2 positive breast cancer

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

OCTOBER 20, 2005

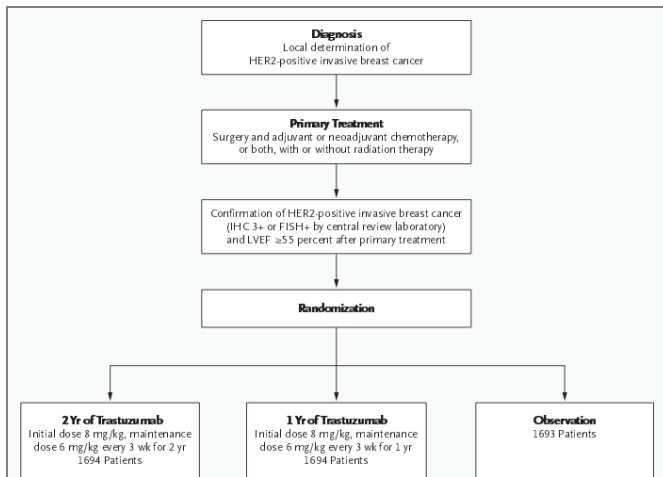
VOL. 353 NO. 16

Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer

Martine J. Piccart-Gebhart, M.D., Ph.D., Marion Procter, M.Sc., Brian Leyland-Jones, M.D., Ph.D., Aron Goldhirsch, M.D., Michael Untch, M.D., Ian Smith, M.D., Luca Gianni, M.D., Jose Baselga, M.D., Richard Bell, M.D., Christian Jackisch, M.D., David Cameron, M.D., Mitch Dowsett, Ph.D., Carlos H. Barrios, M.D., Günther Steger, M.D., Chiun-Shen Huang, M.D., Ph.D., M.P.H., Michael Andersson, M.D., Dr.Med.Sci., Moshe Inbar, M.D., Mikhail Lichinitser, M.D., István Láng, M.D., Ulrike Nitz, M.D., Hiroji Iwata, M.D., Christoph Thomssen, M.D., Caroline Lohrisch, M.D., Thomas M. Suter, M.D., Josef Rüschoff, M.D., Tamás Sütő, M.D., Ph.D., Victoria Greatorex, M.Sc., Carol Ward, M.Sc., Carolyn Straehle, Ph.D., Eleanor McFadden, M.A., M. Stella Dolci, and Richard D. Gelber, Ph.D., for the Herceptin Adjuvant (HERA) Trial Study Team

HERA study

Randomization



HERA study

Statistical design

Statistical analysis Enrollment of 4482 patients was planned to detect a 23 percent relative reduction in the risk of a disease-free–survival event with 80 percent power, with the use of a two-sided significance level of 2.5 percent for each comparison:

- Two years of trastuzumab versus observation
- One year of trastuzumab versus observation.

HERA study

Sample size

A total of 951 disease-free–survival events were required for the final analysis.

One interim efficacy analysis was planned after 475 events, with a specified significance level of $P \leq 0.001$ required, with the use of a sequential plan according to the O'Brien–Fleming boundary¹ as implemented by Lan and DeMets.

The independent data monitoring committee reviewed data on patient enrollment, deaths, compliance, and safety every six months and conducted the interim cardiac safety and efficacy reviews as preplanned.

¹More on this later.

HERA study

Pre-planned statistical analyses

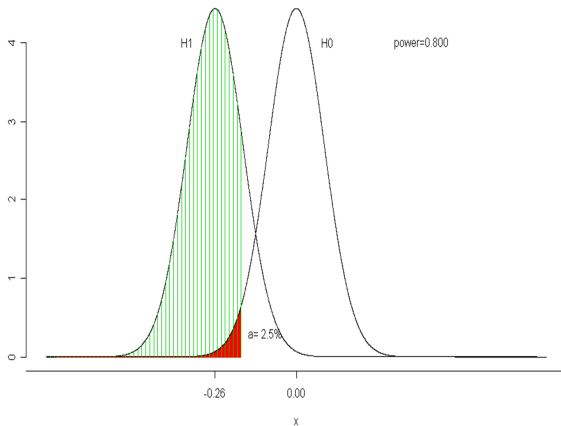
The efficacy analyses were conducted according to the intention-to-treat principle.

Chi-square tests for categorical data and log-rank tests for time-to event end points provided two-sided P values.

Kaplan–Meier curves and Cox proportional hazards regression analysis was to be used to estimate hazard ratios and 95 percent confidence intervals.

HERA study

Statistical considerations



$$\Delta = \ln(\text{HR}) = \ln(0.77) = -\ln(1-0.23) = -0.26$$

Equivalence trials

What can we learn

The null hypothesis in an equivalence trial is that the new therapy (N) is not equivalent with the standard therapy (S).

Equivalence is shown when this null hypothesis is rejected in which case we conclude that N is *equivalent* with S within a pre-specified window.

Note: In equivalence studies one must select a treatment that has been demonstrated to be superior to placebo, since there is no internal control within the trial itself (i.e., there is no placebo arm)

Characteristics of equivalence trials

In general we test whether the experimental therapy is as effective as the standard therapy.

Equivalence studies are also called studies of proof of the null hypothesis or, sometimes, non-inferiority studies (although non-inferiority studies are somewhat different).

If the experimental treatment is “equivalent” but has lower cost or less toxic side-effects, the improvement is self-evident.

Equivalence trial example

The GATE study

Research

Original Investigation | CLINICAL TRIAL

Equivalence of Generic Glatiramer Acetate in Multiple Sclerosis A Randomized Clinical Trial

Jeffrey Cohen, MD; Anna Belova, MD; Krzysztof Selmaj, MD; Christian Wolf, MD; Maria Pia Sormani, PhD;
Janine Oberyé, MSc; Evelyn van den Tweel, PhD; Roel Mulder; Norbert Koper, PhD; Gerrit Voortman, MSc;
Frederik Barkhof, MD; for the Glatiramer Acetate Clinical Trial to Assess Equivalence
With Copaxone (GATE) Study Group

The GATE trial

Premise

The idea behind the GATE study was, simply put, that “the patents for the older treatments for relapsing-remitting multiple sclerosis are expiring, creating the opportunity to develop generic alternatives” (Cohen et al., JAMA Neurology, 2015).

Thus, the objective of the study was “To evaluate in the Glatiramer Acetate Clinical Trial to Assess Equivalence With Copaxone (GATE) study whether **generic glatiramer acetate (hereafter generic drug) is equivalent to the originator brand glatiramer acetate (hereafter brand drug) product**, as measured by imaging and clinical end points, safety, and tolerability” .

The GATE study

Design

Participants were randomized 4.3:4.3:1 to receive generic glatiramer acetate (20 mg), brand glatiramer acetate (20 mg), or placebo by daily subcutaneous injection for 9 months.

The **primary endpoint** was the total number of gadolinium-enhancing lesions during months 7, 8, and 9.

Secondary endpoints included other magnetic resonance imaging parameters, annualized relapse rate, and Expanded Disability Status Scale score.

Safety and tolerability were assessed by monitoring adverse events, injection site reactions, and laboratory test results.

The GATE study

Statistical considerations



Statistical Analysis

Based on the European/Canadian Glatiramer Acetate trial,⁴ we estimated that the mean number of gadolinium-enhancing lesions during months 7 through 9 would be 1.75 times higher with placebo treatment compared with brand glatiramer acetate treatment. The upper limit of the equivalence margin was set at 1.375, representing 50% of the treatment effect vs placebo observed in the aforementioned trial. The lower limit of the equivalence margin was set at 0.727 to create a symmetrical margin in the log scale. To conclude equivalence between generic glatiramer acetate and brand glatiramer acetate, efficacy in the combined active treatment groups needed to be superior to placebo (confirming study sensitivity), and the 2-sided 95% CI for the estimated ratio of generic drug to brand drug needed to be fully enclosed in the prespecified equivalence margin. Given the sample size as calculated and the estimated width of the 95% CI for the ratio of generic drug to brand drug, the maximal allowable difference between the point estimates to show equivalence would be approximately 10%. With a dropout rate of 12%, we estimated that 336 evaluable participants in each of the generic drug and brand drug groups and 78 evaluable participants in the placebo group would provide 98% power to demonstrate study sensitivity, 92% power to show equivalence of generic drug and brand drug, and 90% power to show study sensitivity and equivalence.

Non-inferiority trial example

Gefitinib vs. Docetaxel for non-small-cell lung cancer

VOLUME 26 • NUMBER 26 • SEPTEMBER 10 2008

JOURNAL OF CLINICAL ONCOLOGY

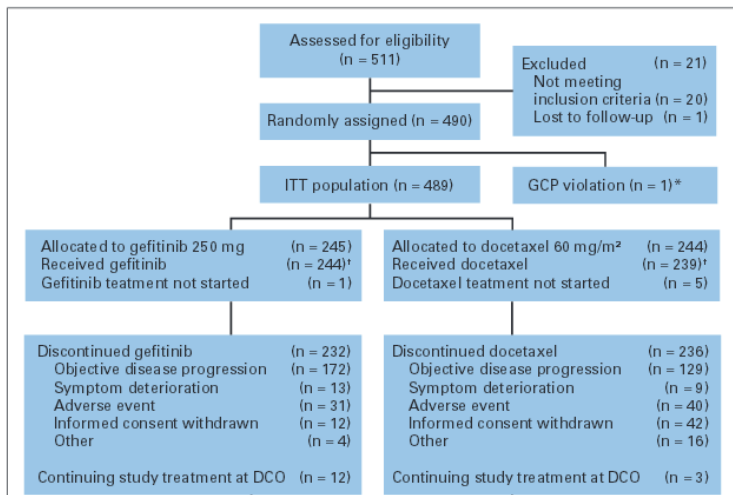
ORIGINAL REPORT

Phase III Study, V-15-32, of Gefitinib Versus Docetaxel in Previously Treated Japanese Patients With Non–Small- Cell Lung Cancer

Riichiroh Maruyama, Yutaka Nishiwaki, Tomohide Tamura, Nobuyuki Yamamoto, Masahiro Tsuboi, Kazuhiko Nakagawa, Tetsu Shinkai, Shunichi Negoro, Fumio Imamura, Kenji Eguchi, Koji Takeda, Akira Inoue, Keisuke Tomii, Masao Harada, Noriyuki Masuda, Haiyi Jiang, Yohji Itoh, Yukito Ichinose, Nagahiro Saijo, and Masahiro Fukuoka

Non-inferiority trial example

Randomization



Non-inferiority example

Study design

Multi-center, randomized, open-label (i.e., nonblinded), clinical study of gefitinib versus docetaxel in Japanese patients who had pretreated, locally advanced/metastatic (stages IIIB to IV) or recurrent NSCLC.

Patients were randomly assigned by using stratification factors of sex, performance status (PS; 0 to 1 versus 2), histology (adenocarcinoma versus others), and study site.

The **primary endpoint** was overall survival, and the study aimed to show *non-inferiority* of gefitinib versus docetaxel.

Secondary endpoints were progression-free survival (PFS), time to treatment failure, objective response rate (ORR), disease control rate (DCR), quality of life (QoL), disease-related symptoms, safety, and tolerability.

Non-inferiority trial example

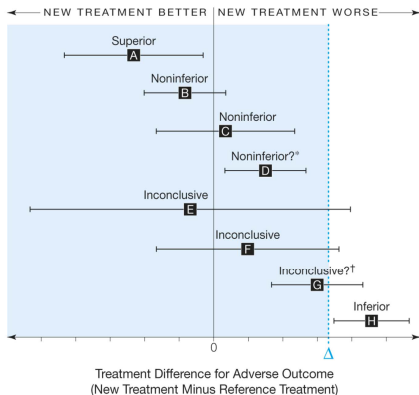
Pre-planned analyses

The primary overall survival analysis was conducted in the intent-to-treat (ITT) population by estimating the HR and two-sided 95% CI for gefitinib versus docetaxel, derived from a Cox regression model without covariates (significance level adjusted because of interim analysis).

- *Non-inferiority* was to be concluded if the upper CI limit was $\delta \leq 1.25$.
- *Superiority* was concluded if the upper CI limit was less than 1.

A total of $D = 296$ death events were required for 90% power to demonstrate non-inferiority, with the assumption that gefitinib had better overall survival than docetaxel (median survival, 14 vs. 12 months), and the study plan was to recruit 484 patients.

More on non-inferiority studies



 The JAMA Network

From: **Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement**

JAMA. 2006;295(10):1152-1160. doi:10.1001/jama.295.10.1152

Error bars are 2-sided 95% confidence intervals. Tinted area is the zone of inferiority.

Comments

A: If the CI lies wholly to the left of zero, the new treatment is superior.

B and *C*: If the CI lies to the left of Δ and includes zero, the new treatment is noninferior but not shown to be superior.

D: If the CI lies wholly to the left of Δ and wholly to the right of zero, the new treatment is non-inferior in the sense already defined, but it is also inferior in the sense that a null treatment difference is excluded. This puzzling case is rare, since it requires a very large sample size. It can also result from having too wide a non-inferiority margin².

²This CI indicates noninferiority in the sense that it does not include Δ , but the new treatment is (statistically) significantly worse than the standard. Such a result is unlikely because it would require a very large sample size.

More comments

E and *F*: If the CI includes Δ and zero, the difference is nonsignificant but the result regarding noninferiority is inconclusive.

G: If the CI includes Δ and is wholly to the right of zero, the difference is statistically significant but the result is inconclusive regarding possible inferiority of magnitude Δ or worse³.

H, If the CI is wholly above Δ , the new treatment is inferior.

³This CI is inconclusive in that it is still plausible that the true treatment difference is less than Δ , but the new treatment is (statistically) significantly worse than the standard.

Factorial designs

Factorial designs are used to test the effect of more than one treatment and uses a design that permits the assessment of interaction between them.

A typical 2×2 factorial design comparing the effect of treatment A and treatment B (assuming that A&B can be given in combination) is as follows:

Treatment A	Treatment B		Total
	No	Yes	
No	n	n	$2n$
Yes	n	n	$2n$
Total	$2n$	$2n$	$4n$

Effect estimation in factorial designs

The treatment effects in a factorial design in the previous slide are as follows:

A typical 2×2 factorial design comparing the effect of treatment A and treatment B (assuming that A&B can be given in combination) is as follows:

Treatment A	Treatment B	
	No	Yes
No	\bar{Y}_o	\bar{Y}_B
Yes	\bar{Y}_A	\bar{Y}_{AB}

Interaction effects

Interaction effect between factors A and B is the modification of the effect of factor A by factor B . This, in the context of two drugs is either

- *Synergism*

A positive (synergistic or potentiated) interaction between the two drugs (i.e., a larger additive effect than would be expected by adding the individual effects of the two drugs)

- *Antagonism*

A negative (antagonistic) interaction between two drugs (i.e., a smaller additive effect than would be expected by adding the individual effects of the two drugs)

Efficiency of factorial designs

In the absence of interaction between factor A and B , the estimates of the effect of these two factors can be averaged from the estimates of their treatment effect versus placebo. This is

$$\hat{\beta}_A = \frac{(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)}{2}$$

for factor A and,

$$\hat{\beta}_B = \frac{(\bar{Y}_B - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_A)}{2}$$

for factor B .

Efficiency of factorial designs (continued)

The efficiency of the factorial design becomes obvious if one considers that, if the variance of the patient response is σ^2 and is the same in all treatment groups, then the variance of $\hat{\beta}_A$ (similarly for $\hat{\beta}_B$) is

$$\begin{aligned}\text{Var}(\hat{\beta}_A) &= \text{Var}\left[\frac{(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)}{2}\right] \\ &= \frac{1}{4} \frac{4\sigma^2}{n} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Now, considering the variance of the treatment effect

$$\text{Var}(\hat{\beta}'_A) = \text{Var}(\bar{Y}_A - \bar{Y}_0) = \frac{2\sigma^2}{n}$$

Thus, a factorial design is more efficient compared to a one-treatment-at-a-time design (i.e., conducting two separate trials for evaluating treatments A and B).

Testing of interactions

Factorial designs are the only designs where interactions between factor A and B can be measured. The definition of an interaction is that the effect of A is different in the presence versus absence of B .

This can be estimated by comparing with zero

$$\hat{\beta}_{AB} = (\bar{Y}_A - \bar{Y}_0) - (\bar{Y}_{AB} - \bar{Y}_B)$$

We note that the variance of β_{AB} is

$$\text{Var}(\hat{\beta}_{AB}) = \frac{4\sigma^2}{n}$$

which is four times larger than the variance of the individual effects when there is no interaction.

Testing for interactions (continued)

To get the same precision for estimating the interaction effect we need four times the sample size.

Alternatively, the interaction effect must be twice as large as the main effects in order for it to be detected with the same power by the sample size necessary to detect the effect size of the main effects. This is rarely, if ever, the case.

This shows that estimation of interaction effects with a desired power level requires a larger sample size.

Example: Omega-3 for the prevention of heart failure

In a (fictional⁴) clinical trial of Omega-3 for the prevention of heart failure, the following 2×2 factorial design was adopted:

	Marvistatin 5 mg	Maristatin 80 mg	Total
Omega-3	100 participants	100 participants	200
Placebo	100 participants	100 participants	200
Total	200	200	400

The primary endpoint was to see whether Omega-3 had a protective effect for patients receiving statins and, secondarily, whether the effect, if any, had an interaction with statin dose.

⁴<https://prsinfo.clinicaltrials.gov/trainTrainer/Factorial-Design-Fiction-Manuscript.pdf>

Analysis of the Omega-3 study

The results corresponding to the primary endpoint were as follows:

	Marvistatin 5 mg	Maristatin 80 mg	Total
Omega-3	27/100 events	25/100 events	52/200
Placebo	26/100 events	24/100 events	50/200
Total	53/200 events	49/200 events	102/400

It is obvious that the events corresponding to Omega-3 can be combined across the two groups of subjects receiving Omega-3 and these can be compared against the combined groups receiving placebo.

If there is no interaction (which, in turn, will allow us to combine the two groups in each case), this design effectively doubles the sample size versus a study which would test the two interventions independently (via conducting two separate trials).

Crossover designs

In contrast to the designs where participants are treated with a single or combination treatment *concurrently*, in crossover designs, treatments are administered *sequentially*. The main advantage of this study is that treatment effects can be compared within the same subjects (thus eliminating within-subject or biological variability).

Crossover designs are different in objective and scope from trials that give treatments sequentially (e.g., $A \rightarrow B \rightarrow C$ versus $A \rightarrow B$) but assess the incremental effect of a treatment (here treatment C) or from factorial designs where patients are administered a combination of treatments *simultaneously*.

Randomization in crossover studies

Crossover designs do not employ randomization in the same way that comparative studies do. This is because, all patients receive all treatments.

There are randomized crossover studies where the *order* of treatment administration is randomized among subjects (e.g., $A \rightarrow B$ versus $B \rightarrow A$). The lack of randomization of treatment allocation is not a problem *per se* since the homogeneity of the treatment groups is ensured by having the same subjects receive all treatments.

Nevertheless, randomization of period of treatment administration is not sufficient to ensure that treatment groups are similar because it is not known in advance whether period of administration is significantly associated with the treatment effect (which means that, only after analysis of the study data, will we know whether the treatment groups defined by different sequence of treatment administration will be comparable).

Efficiency of crossover studies

To see why crossover designs are efficient, we consider that each subject is its own control. Thus, a potentially large component of treatment effect variability is removed from the estimation. To see this, consider the variance of the difference in treatment effects A and B , $\hat{\Delta}_{AB}$, noted here as \bar{Y}_A and \bar{Y}_B respectively, will be:

$$\begin{aligned}\text{Var}(\hat{\Delta}_{AB}) &= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - 2\text{cov}(\bar{Y}_A, \bar{Y}_B) \\ &= 2\frac{\sigma^2}{n}(1 - \rho_{AB})\end{aligned}$$

We note that, in comparative studies, $\rho_{AB} = 0$ because the groups receiving the treatments A and B are independent. So $\text{var}(\hat{\Delta}_{AB}) = 2\sigma^2/n$. However, if $\rho_{AB} > 0$, as it is usually expected when treatment is administered to the same patient, the variance of the effect difference will be smaller in a crossover than a comparative study.

Difficulties with crossover studies

There are several difficulties with crossover designs

- *Carryover effects.*

This can happen because

- ▶ Physiological persistence of the effects of the first treatment. This can be addressed with longer “washout” periods.
- ▶ The first treatment may change or cure the condition artificially affecting the effect of the second drug

- *Dropout effects*

Dropout rate between the first and second treatment may be exacerbated because of the increased length of the study and the effects of the first treatment. Dropouts have strong effect on the analysis more so than in a comparative study.

Difficulties with crossover studies (continued)

- *Complexities with the analysis of crossover studies.*

In particular, these are

- ▶ Use a *staged plan* where carryover effects are studied in the first stage and, if none are found, the main effects are estimated in the second stage
- ▶ Use baseline measurements in each period to test for carryover effect

Prerequisites for application of crossover designs

There are a number of factors that must be in place before undertaking a crossover design

- Ensure that there is a large positive correlation between successive evaluations on the same patient
- Disease intensity must be constant throughout the study
- Small or no carryover effect (or if there is carryover effect, a sufficient washout period is applied)
- Expected (negative) view of the crossover trial by regulators (FDA considers RCT as the *de facto* standard for proof of treatment effect)

Efficiency of the crossover design

To comment on the efficiency of the crossover design, we consider the estimation of the treatment effects. In the absence of a treatment \times period interaction, the treatment effect difference of treatments A and B is

$$\hat{\beta}_1 = \frac{1}{2} (\bar{Y}_{B2} - \bar{Y}_{A2} + \bar{Y}_{B1} - \bar{Y}_{A1})$$

and the period effect is

$$\hat{\beta}_2 = \frac{1}{2} (\bar{Y}_{B2} - \bar{Y}_{B1} + \bar{Y}_{A2} - \bar{Y}_{A1})$$

where \bar{Y}_{Ai} and \bar{Y}_{Bi} , $i = 1, 2$ are the treatment effects of treatments A and B during period 1 and 2 respectively. The variance of $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{n}(1 - \rho)$$

Efficiency of the crossover design (continued)

On the other hand, the treatment \times period interaction effect, is

$$\hat{\beta}_3 = (\bar{Y}_{A2} - \bar{Y}_{A1}) - (\bar{Y}_{B2} - \bar{Y}_{B1})$$

If $\hat{\beta}_3 \neq 0$ then the treatment effect is estimated by

$$\hat{\beta}_1 = (\bar{Y}_{B1} - \bar{Y}_{A1})$$

The variance of $\hat{\beta}_3$ is

$$\begin{aligned}\text{Var}(\hat{\beta}_3) &= \text{var}(\bar{Y}_{A2} - \bar{Y}_{A1}) - \\ &= 4\frac{\sigma^2}{n}(1 + \rho)\end{aligned}$$

which is more than four times larger than the variance of $\hat{\beta}_1$ or $\hat{\beta}_2$ for $\rho \geq 0$. Thus, crossover studies are not as efficient in testing the carryover effect (treatment \times period interaction).

Example: Cross-over trial for angina pectoris

In a cross-over trial, the effects of pronethanol on angina pectoris were investigated⁵. Twelve patients received placebo for two weeks and pronethalol for two weeks, in random order. The results of the study are as follows:

Patient	Placebo	Pronethalol	Difference
1	71	29	42
2	323	348	-25
3	8	1	7
4	14	7	7
5	23	16	7
6	34	25	9
7	79	65	14
8	60	41	19
9	2	0	2
10	3	0	3
11	17	15	2
12	7	2	5

⁵<https://www-users.york.ac.uk/~mb55/msc/trials/cross.htm>

Analysis of the pronethanol trial

The summary statistics of the previous table are as follows:

Placebo		Pronethanol		Difference	
Mean (\bar{Y})	Std. dev (s)	Mean (\bar{Y})	Std. dev (s)	Mean (\bar{Y})	Std. dev (s)
53.42	89.07	45.75	97.19	7.67	15.11

It is obvious that analyzing this study as a two-independent-sample clinical trial is associated with dramatically higher variability than the (correct) analysis of a matched cross-over design.