BIOSTATISTICAL METHODS

FOR TRANSLATIONAL & CLINICAL RESEARCH

Phase II Trials:

EARLY-PHASE CLINICAL TRIALS

CANCER CLINICAL TRIALS

- Phase I: First human trial to focus on safety
- Phase II: Small trial to evaluate efficacy
- Phase III: Large controlled/randomized trial to demonstrate efficacy prior to FDA approval
- Phase IV: Mostly optional, post-regulatory approval, to provide the medicine's more comprehensive safety and efficacy profile

The experiment, called “Phase II Clinical Trial”, in this stage is very simple: The MTD, found in a previous Phase I Clinical Trial, is tested in a small (n=10-25) group of patients to learn about the agent’s efficacy. A few trials are larger, up to 50 patients.

BASIC OBJECTIVES OF PHASE II CLINICAL TRIALS

There are three basic objectives in conducting phase II clinical trials:

(1) Benefit the patients

(2) Screen agent/drug for anti-tumor activity

(3) Extend knowledge of toxicology and pharmacology of drug/agent.;

(Plus: safety is always a major concern; In phase II trials, "efficacy" is the "outcome of interest" whereas "safety" is embedded to serve as "stopping rule").

The first objective is to benefit the patients enrolled in the trial; it must be a primary objective of any therapeutic intervention. It is always the primary "motivation"; however, it's not often stated in research protocols as an "objective". The concerns for patients are/should always be taken very seriously - by everyone - in the design stage; benefiting the patients is the first major objective but not a "research objective" simply because we would not be able to evaluate it. Why?

The second objective is to screen the experimental agent for anti-tumor activity in a given type of cancer; agents which are found to have substantial anti-tumor activity and an appropriate spectrum of toxicity are generally incorporated into combinations to be evaluated for patient benefit in controlled phase III clinical trials. For many investigators, this process of screening for anti-tumor activity is considered as "the" activity of phase II trials – as far as research is concerned.

We should distinguish, clearly from the research point of view, Objective 1 (benefit the patients) from Objective 2 (screen agent used in the trial for anti-tumor activity; this benefit “investigators”).

Primary outcome used in phase II trials is often the "response" which is defined, in the case of solid-tumor cancers, as having a 50% decrease in tumor size, for example, lasting for 4 weeks. The analysis of the resulting binary data is simply based on the "response rate". But response rate is only an appropriate endpoint for evaluating Objective #2 (screen agent used in the trial for anti-tumor activity) – not Objective #1 (benefit the patients) .

Generally, we cannot adequately evaluate the extent to which Objective #1 (benefit the patients) is achieved in one-arm phase II trials; and that is why it is not often stated. First, "response" is only meaningful to and benefit the patient if causing tumor shrinkage means extending survival or, at least, improving quality of life. This may or may not be the case. Logically, we believe so but it has not been convincingly proven. Keep in mind that the sample size is very small and we have only short-term observations (1-2 years)

In addition, when an untreated control group is not available, we generally cannot properly evaluate whether the new agent influences survival so as to benefit the patients. Most phase II trials are one-arm, non-randomized, open-label trials. Effects observed could well be Placebo Effect, i.e. likely psychological.

One could compare, in terms of survival, “responders” versus “non-responders”; however, this not a valid way of demonstrating that there has been an impact of treatment on survival. Such comparisons are biased by the fact that responders must live long enough for a response to be documented. In addition, responders may have more favorable prognostic factors than non-responders, leading to a difference in survival which then be wrongly credited to the treatment.

Response is still “used as” a “surrogate” for the more relevant, more important endpoint of survival even though no-one can prove that they are equivalent (or we can even say that everyone knows that they are not equivalent). Response is still used, and is popular, because: (i) it can be observed on all (or almost all) patients, and (ii) it can typically be determined rather quickly.

It’s an endpoint that everybody understand. For statisticians, it’s the most simple statistic because the “response rate” is just a proportion; the most you can do is forming its confidence intervals. The length of the confidence interval, if specified, would be the basic for sample size determination. In practice, it is hard to afford a large Phase II Clinical Trial.

The third basic objective of a Phase II Trial is to extend our knowledge of toxicology and pharmacology of drug/agent. Ironically, this objective is often listed as "secondary" and, therefore is overlooked by statisticians. Most of the times, details - such data analysis plan - are missing – most Biostatistics students do not often see pharmacology data (which are mostly non-linear regression). (& We know that pharmacology could even be used to guild dose-escalation plan).

Objective #2 of a Phase II Trial is to screen agent used in the trial for anti-tumor activity. There is frequently great variability in the response rates reported from different phase II trials of the same agent. There are a number of factors that contribute to this variability. For example, response criteria and response assessment which are often subjective without universal guidelines. Plus a number of factors related to the conduct of the trials: dosage, protocol compliance, reporting procedure (issue: “evaluable” versus “un-evaluable” patients), etc...

The most important factor leading to variability (of reported response rate) comes from the patient selection process dictated by “inclusion criteria”, “exclusion criteria” - some sections that few statisticians read!

Patients in phase I and phase II trials are mostly terminal cancer patients who have failed all standard therapies and for whom the new anti-tumor agent being tested may be the last hope.

Response rates generally decrease as the extent of prior therapy increases. Patients who have failed several prior regimens are more likely to have tumors composed large numbers of resistant cells, and such patients are also less likely to be able to tolerate full doses of the investigational drug.

Probably the most frequent problem with phase II trials is that some selected patients are so debilitated by disease and prior therapy that an adequate evaluation of anti-tumor activity is impossible. Such patients are more likely to die or withdraw early in the course of treatment; and some investigators consider these patients "inevaluable". The variable proportion of such patients - from study to study - contributes to the variability in reported response rates.

Combining results – based on response rates - in method such as “meta analysis” is, therefore, even judged as invalid or questionable; it’s not a matter of sample size!

To overcome this problem to certain extent, it is recommended that the practice of "intent-to-treat analysis" in phase III trials also be used in the analysis of phase II trials. But some standardization of inclusion and exclusion criteria are very much desirable if ones want to compare and combine results (using meta analysis).

In addition to one-arm trials, there are randomized phase II trials; some with control arms, some without a control arm. But these are not very popular because phase II sample sizes are often small. In addition, large controlled phase III trials involving “real-life treatment regimens” are often involved combinations, not single agents.

RANDOMIZED PHASE II TRIALS WITH A CONTROL

- One type of phase II design involves randomization between an investigational agent and an active standard treatment.
- The purpose, however, is not to determine if the new agent is better or worse than the active control.
- The major objective of the randomization is to help in the interpretation of a poor response rate of the investigational agent.

This type of randomized design is not very popular because : (i) it's only potentially useful where an adequate response rate on the active control is not known or not assured (most of the times, we know more about the controls), and (ii) if we do not know enough about the a control, with the usual phase II small sample sizes, it may be difficult to reliably determine whether the patients are sufficiently responsive to the control treatment. Plus the underlying ethical concern of have some patients served as "controls" because for them the new anti-tumor agent being tested may be the last hope.

RANDOMIZED PHASE II TRIALS WITHOUT CONTROL ARMS

- Two or more treatment arms are possible and the arms are all “experimental”.
- Investigators are puzzled at the rationale for conducting a large randomized phase III trial to compare the two arms either one of which may have no activity (i.e. efficacy) in the disease.
- Phase II trials may provide needed early stopping rules because toxicity profiles are still not known.

MAJOR ADVANTAGES

- Randomization helps to ensure that patients are centrally registered before treatment starts
- Central registration is essential for checking patient's eligibility, terminating accrual when the target sample size is reached, and establishing reliable records.
- There will be some limited form of comparison - in addition to response rate - the "degree" of anti-tumor activity (extent of tumor shrinkage), the durability of responses, etc…

However, this type of randomized design is only used for certain “limited” form of comparison; it does not involve formal statistical tests of significance, nor phase III-type sample size determination. The most prevalent forms (of randomized trials without a control) are what we usually called “designs for selection” – also called “screening trials”.

Design For Selection

THE NEED FOR SELECTION

- The process starts with a dose-finding phase I trial leading to MTD
- Next, a small one-arm phase II trial to study anti-tumor activity – through “response rate”;
- If the results from the phase II trial are promising (safe, effective), the agent becomes a “candidate”.
- Problem: There may be too many candidates for phase III trial (to compare efficacy to a standard treatment or placebo); sometimes differences between candidates are small.

SPECIFIC AIM

- At this stage, the aim is not to make a definite conclusion about the "superiority" of one treatment (or one mode of administration) as compared to the other.
- If "correct ordering" is the goal, a properly-powered a phase II trial would be required; but we can't afford a phase III trial before a phase III trial!
- The goal is to ensure that if one treatment is clearly inferior, it is less likely carried forward to the phase III trial (versus standard/placebo).

(1) There are no Standard/Placebo; both treatments (or modes of treatment) A & B are experimental.

(2) Decision (i.e. selection) has to be made; does not fit framework of “statistical test of significance” where “not statistically significance” is a possibility. We cannot afford it!

OTHER CHARACTERISTICS

- Because the goal is not “superiority”, Type I errors are less relevant; emphasis is on the “probability of correct selection” – called “Designs for Selection” or “Screening Trials”
- Trial is randomized.
- In addition to efficacy, other criteria may be also be considered (toxicity, cost, ease of administration, or quality of life); investigators want that flexibility.

A and B were identified, likely separately from phase II trials. So, this screening trial could be referred to as “phase II and a half”; but just “phase II” for simplicity.

CRITERION

- If the “observed outcome” (e.g. response rate, but could some sample mean) of one arm is greater than “d units” than the other, the arm with “better observed outcome” (larger proportion or larger sample mean) will be selected for use in the next phase III trial.
- If the difference is smaller than d units (“d” may or may not be 0), selection may be based on other factors
- For example:

$$\text{If } "p_A - p_B > d\%"; \text{ treatment A is selected}$$

Or

$$\text{If } "\bar{x}_A - \bar{x}_B > d"; \text{ treatment A is selected}$$

CORRECT OUTCOME

- Suppose that the outcome variable is response rate and Treatment A is assumed to be better:

$$\pi_A - \pi_B = \delta$$

- The "probability of correct outcome" is:

$$P_{corr} = \Pr[p_A - p_B > d \mid \pi_A, \pi_B]$$

CORRECT SELECTION

- If the "observed outcome is ambiguous", i.e. difference is less than "d", treatment A could still be chosen (by factors other than efficacy), with – say - probability ρ;
- The probability of correct selection is:

$$P_{corr} = \Pr[p_A - p_B > d \mid \pi_A, \pi_B]$$

$$P_{Amb} = \Pr[p_B - d \leq p_A \leq p_B + d]$$

$$\lambda = P_{CorrSel} = P_{Corr} + \rho P_{Amb}$$

For simplicity, we could set “d=0”; in that case the decision rule requires that at the end of the trial, whichever arm is ahead by any margin be carried forward to the phase III trial. However, this may be less desirable because the rule does not allow the inclusion of factors other than efficacy be included in the decision process.

If the “observed outcome is ambiguous”, i.e. difference is less than “d”, treatment A could still be chosen (by factors other than efficacy), with – say - probability ρ; Conservatively (and likely), ρ = 0 but we could have ρ = .5

WHAT DO STATISTICIANS DO?

- The size of "d" is a clinical decision; at the end of the trial, compute ($p_A - p_B$) and compare to d.
- Statistician is responsible for "the design", to find sample size n (per arm) to ensure that "the probability of correct selection" exceeding certain threshold; say $\lambda \geq .90$ (similar to power).
- Population parameters (such as π_A and π_B, or π_A and δ) are in "Alternative Hypothesis"; ideas from separate phase II trials.
- We cover the case of response rate but method is applicable to continuous outcome variables.

The difference $(p_A - p_B)$ is distribute d as Normal with Mean μ and Variance $= \sigma$,

$$\mu = (\pi_A - \pi_B) = \delta$$

$$\sigma^2 = \frac{1}{n}[\pi_A(1-\pi_A) + \pi_B(1-\pi_B)]$$

$$P_{corr} = \Pr[p_A - p_B > d \mid \pi_A, \pi_B]$$

$$= 1 - \Phi[\frac{d - \delta}{\sigma}]$$

$$P_{Amb} = \Pr[-d \le p_A - p_B \le d]$$

$$= \Phi[\frac{d - \delta}{\sigma}] - \Phi[\frac{-d - \delta}{\sigma}]$$

$$\lambda = P_{CorrSel} = P_{Corr} + \rho P_{Amb}$$

Example: Let take $\pi_A = .35$ and $\pi_B = .25$ (or $\delta = .10$) and d = .05, n = 50 ($\sigma = .09$)

$$P_{corr} = 1 - \Phi[\frac{d-\delta}{\sigma}]$$

$$= 1 - \Phi(-.55) = .71$$

$$P_{Amb} = \Phi[\frac{d-\delta}{\sigma}] - \Phi[\frac{-d-\delta}{\sigma}]$$

$$= \Phi[-.55] - \Phi[-1.65] = .24$$

$$\lambda = P_{CorrSel} = P_{Corr} + \rho P_{Amb}$$

$$= \begin{cases} .71 \; for \; \rho = 0 \\ .83 \; for \; \rho = .5 \end{cases}$$

Changing n will change σ and, therefore, the probability of correct selection λ; that's key to sample size determination

If it was a “test of significance’, at the conclusion of the trial, one would compute the sample proportions and reject the Null Hypothesis if:

$$\frac{|p_A - p_B|}{\sigma} \geq z_{\alpha/2}$$

The statistical power of this test would be, which is very different from the probability of correct selection:

$$1-\beta = 1-\Phi[z_{\alpha/2} - \frac{\delta}{\sigma}] + \Phi[-z_{\alpha/2} - \frac{\delta}{\sigma}]$$

Some Results from Sargent and Goldberg (2001)				
n	δ	P_{corr}	P_{Amb}	$P_{corr}+(.5)*P_{Amb}$
50	0.1	0.71	0.24	0.83
50	0.15	0.87	0.12	0.93
75	0.1	0.82	0.15	0.89
75	0.15	0.95	0.05	0.97
100	0.1	0.88	0.1	0.93
100	0.15	0.98	0.02	0.99

Odd rows: $\pi_A = .35$, $\pi_B = .45$ ($\delta = .10$), $d = .05$

Even rows: $\pi_A = .35$, $\pi_B = .50$ ($\delta = .15$), $d = .05$

(Corresponding powers are much lower!)

Example:

$\pi_A = .35$, $\pi_B = .45$ ($\delta = .10$), $d = .05$ & $n = 100$

$P_{corr} = .88$; $P_{corr} + (.5)P_{Amb} = .93$

$$\sigma = \sqrt{\frac{.35(1-.35)+.45(1-.45)}{100}}$$

$$= .07$$

$$Power = 1 - \Phi[z_{\alpha/2} - \frac{\delta}{\sigma}] + \Phi[-z_{\alpha/2} - \frac{\delta}{\sigma}]$$

$$= 1 - \Phi(1.96 - \frac{.10}{.07}) + \Phi(-1.96 - \frac{.10}{.07})$$

$$= 1 - \Phi(.53) + \Phi(-3.39)$$

$$= 1 - .7019 + .0003$$

$$= .2984$$

If a selection (of one treatment over the other) is made when two treatments are equally effective, it's type I "error". But in our context, it's fine because whatever treatment is selected patients are equally well-served.

A design for selection has no concern for Type I errors, so required sample size is much smaller – The probability of correct selection is like the counterpart of statistical power.

REFERENCES

- Simon et a. Randomized phase II clinical trials; *Cancer Treatment Reports 69:* 1375-1381, 1985
- Thall et al. Two-stage selection and testing designs for comparative clinical trials. *Biometrika 75*: 303-310, 1988
- Liu et al. False positive rates of randomized phase II designs. *Control Clinical Trials 20*: 343-352, 1999.
- Sargent and Goldberg. A flexible design for multiple armed screening trials. *Statistics in Medicine 20*: 1051-1060, 2001

Suggested Exercise:

Consider the configuration given in the last example: $\pi_A = .35$ and $\pi_B = .25$ and let take $d = .05$ and $\rho = 0$.

(1) Verify that if n=50, then $\sigma = .09$

(2) Find the sample size (trial by error) if we want $\lambda = .90$