BIOSTATISTICS METHODS

FOR TRANSLATIONAL & CLINICAL RESEARCH

EARLY PHASE CLINICAL TRIALS

Part A: Concepts & Designs

Drug development is the process of finding and producing therapeutically useful pharmaceuticals and turning them into effective and safe medicines. It is a complex process starting with screening chemicals to identify a lead compound, going through lots of works in toxicology, pharmacodynamics, and pharmacokinetics, and phases of clinical trials.

A successfully completed development and testing program results in lots of information about appropriate doses and dosing intervals, and about likely effects and side effects of the treatment. It is a process carried out by "sponsors" (mostly pharmaceutical companies but also including major medical centers, e.g. "Center for Drug Design" at the University of Minnesota) and is ultimately judged by "regulators" (e.g. FDA of the United States).

There is no aspect of drug development and testing without participation and contributions from biostatisticians. Statisticians and biostatisticians are also becoming more active in the shaping of the pharmaceutical projects.

There are statisticians even on "the other side of the table"; for many years FDA has employed statisticians and biostatisticians to assist in its review process. At medical centers, biostatisticians participate in protocol designs as well as protocol reviews.

Even though there is no aspect of drug development and testing without participation and contributions from biostatisticians; many are unfamiliar with statistical issues in the early phase of the process – **Phase I and Phase II Clinical Trials**. In Phase I Clinical Trials, lower doses are tried first and cautiously increased until a Maximum Tolerated Dose (MTD) maybe established; the focus is safety. A Phase II Clinical Trial, administering MTD, is the first step to explore and confirm an agent's efficacy.

PHASES OF TRIALS

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

Phase I and II clinical trials present special difficulties because they involve use of agents whose spectrum of toxicity and likelihood of benefits are poorly understood/defined. There were "pre-clinical" studies – e.g. In Vivo & In Vitro experiments and bioassays – but the subjects were animals (In Vivo) or human tissues (In Vitro). And inferences across species are never easy, nor precise.

CANCER PHASE I TRIALS

- Different from other phase I clinical trials, phase I clinical trials in cancer have several main features.
- The efficacy of chemotherapy is associated with a non-negligible risk of severe toxic effect, often fatal, so that ethically, such drugs can be investigated only in cancer patients; and only a small number of patients are available for the trial- any trial in any phase.
- These patients are at very high risk of death in the short term under all standard therapies. At low doses little or no efficacy is expected.
- A slow intra-patient dose escalation is not practically possible.

Lack of information about the relationship between dose & probability of toxicity causes a fundamental dilemma; it's a conflict of scientific versus ethical intent.

We put some patients at risk for their own benefits (and benefits of others).

And it's an unavoidable conflict because these patients have failed under all standard therapies.

We need to reconcile the risks of toxicity to patients with the potential benefit to these same patients and make an efficient design to use no more patients than necessary.

Thus, Phase I cancer trial may be viewed as a problem in optimization: maximizing the dose-toxicity evaluation, while minimizing the number of patients treated

GOAL OF PHASE I CANCER TRIALS

- Patients from standard treatment failure
- New Drug: No efficacy at low doses, will have toxicity at high doses- maybe severe, fatal
- Dose range: little known (first in human)
- Goal: Maximum Tolerated Dose (MTD), reasonable efficacy & tolerable toxicity.

STANDARD DESIGN

For Phase I Clinical Trials

Starting Dose:

The starting dose selection of a phase I trial depends heavily on pharmacology and toxicology from pre-clinical studies. Although the translation from animal to human is not always a perfect correlation; generally, toxicity as a function of body weight is assumed roughly constant across species.

From the toxicity and body weight relationship, it is also often implicitly considered that mouse LD10 is about the same as human MTD. However, the use of a second species has also shown to be necessary because in approximately 90 reviewed drugs, mouse data alone was insufficient to safely predict the human MTD (Arbuck, 1996).

For most investigators, the first dose is often chosen at about one-tenth of the mouse LD10 (at which 10% of mice die). However, some are more cautious & proposed to use the smaller of:
(i) one tenth of the mouse LD10, and
(ii) one third of the beagle dog LD10.

SPACING BETWEEN DOSES

Generally, the dose levels are selected in order that the percentage increments between successive doses diminish as the dose is increased; for example, (i) equally-spaced on the log scale or (ii) a modified Fibonacci sequence is often employed (increases of 100%, 67%, 50%, 40%, then 33% for subsequent doses if more than 5 are planned); this follows a diminishing pattern, with modest increases

Dose Escalation

STANDARD DESIGN

- Start at lowest dose, enroll groups of 3 pts

(i) Move up if none of first 3 have toxicity;

(ii) If two or three patients have dose-limiting toxicity (DLT), stop.

(iii) If one of 3 has toxicity, enroll 3 more at same dose and move up if none of the second cohort have toxicity. Otherwise, stop

- Moving up, repeat the process at new dose

MAXIMUM TOLERATED DOSE

- If dose escalation not possible, i.e. being stop at a given dose, this dose is considered as "above MTD" – have to go down to find MTD
- When a dose is judged as above the MTD, the next lower dose is (often) declared the MTD in the designs without dose de-escalation (it's not completely universal; some investigators only step down half a dose).

Design Characteristics:

SOME RESULTS

- The cohort is limited to 3 pts each with ≥56% probability: It requires few patients
- For more extreme DLT probabilities (5% & 70%), the cohort is only expanded to 6 patients with probability of less than 20%
- When the rate is 30% or higher, the probability to stop is ≥51%: It's rather safe

- If 2 or 3 pts show DLT ($STOP_3$), we are 90% sure that $r \geq 20\%$
- If no patients show DLT (UP_3), we are 90% sure that $r \leq 55\%$
- If r=10%, 91% chance to escalate; if r=60%, 92% chance to stop

OTHER RESEARCH QUESTIONS

There are other important research questions:

(1) What is the actual expected toxicity rate of the MTD selected by the standard design,
(2) Is the expected toxicity rate of the selected MTD by the standard design robust, and
(3) Can improvement be made to the standard design so that we can get to the next phase quicker but still safe enough? What are alternatives to the Standard Design which could help to accelerate the process?

Fast-Track Design

- This design was created to move through low doses using fewer patients. The design uses cohorts of one or three patients, escalates through the sequence of doses using a one-patient cohort until the first DLT is observed. After that, only three-patient cohorts are used.
- When a DLT is observed in a one-patient evaluation of a dose, the same dose is evaluated a second time with a cohort of three new patients, if no patient in this cohort experiences a DLT, the design moves to the next higher dose with a new cohort of three patients. From this point, it progresses as a standard design.

Potential therapeutic agents for cancer treatment can induce severe safety concern - even at lower dose levels; they can generate severe toxicities than most of pharmaceutical agents for treatment of other diseases. Possible adverse effects may be irreversible, even fatal. As a result, phase I trials are conducted only on cancer patients - not healthy volunteers.

Patients in cancer clinical trials are those with malignant tumors; most cancers are life-threatening and the disease process is usually irreversible. Patients in phase I 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

At this stage, the investigator is facing a classic, fundamental dilemma: it's a conflict of scientific versus ethical intent.

We put some patients at risk for their own benefits and benefits of others; and it's an unavoidable conflict because these patients have failed under all standard therapies.

We need to reconcile the risks of toxicity to patients with the potential benefit to these same patients and make an efficient design to use no more patients than necessary.

Thus, the phase I cancer trial may be viewed as a problem in optimization: maximizing the dose-toxicity evaluation, while minimizing the number of patients treated . And that's a real challenge

A Bayesian Approach:
CONTINUAL REASSESSMENT METHOD

ST DESIGN: COMMON CRITIQUES

- Patients enter early are likely treated sub-optimally; may be we need to move up faster (against the principle of "good medicine"!)
- Only few patients left when MTD reached, not enough to estimate MTD's toxicity rate (against the principle of "good statistics"!)

The standard design is not robust; expected rate of the selected MTD is strongly influenced by the doses used. If the trial is such that there are many dose levels below the MTD then the standard design will choose a dose far too low with greater probability than if there are fewer dose levels below the MTD.

According to the (unstated) principle of "good medicine", each patient should be treated optimally: each patient should be treated with the "best" treatment that the doctor. According to this principle, each patient in phase I trial should be given a dose equal to the MTD - if the doctor knows what it is. In most cases, doctors may not know what is the the MTD but they all "know" that, according to the standard design, the first few doses are likely "below" the MTD.

In addition:

(1) It does not target a particular toxicity rate associated with MTD.

(2) It does not make use of all available toxicity data; escalation rule depends solely of toxicity outcomes of the current dose.

STATISTICAL FORMULATION

The MTD could be statistically interpreted as some "percentile" of a tolerance distribution or dose-response curve in terms of the presence or absence of the DLT. In other words, the MTD is the dose at which a specified proportion of patients, say, θ, experiencing DLT. Storer (1997) indicated that the value of θ is usually in the range from .1 to .4. In published literature, it has been considered in the range of 30% to 40%. Generally, it depends on the severity of the side effects and if these side effects are treatable.

A STATISTICAL MODEL

Let Y be the binary response such that Y=1 denote the occurrence of a pre-defined DLT and $\{d_i ; i= 1,2,\ldots,n\}$ be a set of fixed dose levels used in a phase I trial. Let

$p(x) = Pr[Y=1|x]$ &

$Logit [p(x)] = \log \{p(x)/[1-p(x)\}$;

The relationship between Y and dose level d could be described by the logistic model:

$logit [p(x)] = \alpha + \beta x$,

where x is usually the log of the dose d.

MAXIMUM TOLERATED DOSE (MTD)

Let $l_\theta = \text{logit}(\theta)$, then the MTD is defined by $x_\theta = \log(\text{MTD}) = (l_\theta - \alpha)/\beta$ (it's likely not be one of the doses used in a phase I trial under standard design).

Model: $\text{logit}[p(x)] = \alpha + \beta x$

Note that if we have some estimates of the toxicity rates associated with the dose levels used in a phase I standard design, we could estimate to toxicity rate of the resulting MTD and compare to the θ of choice. Of course, the exact rates are not known, but clinicians should have some estimates; otherwise, it would be difficult for the investigator to "justify" the selected dose levels.

The process could consist of these steps:
(i) choose the "maximum tolerated level" θ,
(ii) choose a design and calculate its MTD's expected toxicity rate r_0, and
(iii) compare the calculated expected rate r_0 to the selected level θ; then

(iv) Do it again if needed: trial by error

Besides many elements of arbitrariness (choosing the level θ for the problem and estimating the rates r_i's for the planned doses); the basic problem is, according to Storer (1989, 1993), the standard dose escalation design "frequently failed to provide a convergent estimate of MTD" (so, even if we know what we want, i.e. θ, the standard design might not get us there). The alternative is a newer design, "continual reassessment method".

The primary objective of phase I trials is to find the maximum dose, called MTD, with an acceptable and manageable safety profile for use in subsequent phase II trials. But that's the investigator's objective, not the patient's. Patients in phase I trials are mostly terminal cancer patients for which the new anti-tumor agent being tested may be the last hope. Designs, such as "standard design", do not serve them - at least not ethically.

According to “principle of “good medicine”, the patient should be treated with the best treatment the doctor knows. Patients enter early to a Phase I trial with Standard Design are likely treated sub-optimally; they receive a treatment level that the attending physician knows to be inferior. Some of these patients would likely die before any other therapy can be attempted. The newer design, the “continual reassessment method (CRM)” is an attempt to correct that by giving each patient a better chance of a favorable response.

In addition to the attempt to treat each patient more ethically, the CRM also updates the information of "the dose-response relationship" as observations on DLT become available and then to use this information to concentrate the next tep of the trial around the dose that might correspond to the anticipated target toxicity level. It does so using a Bayesian framework.

The CRM is very attractive and has fostered a heated debate or debates which last for more than a decade. There are many variations of the CRM, we'll briefly describe here a scheme based on a specific prior; the principle and the process are the same if another model is selected.

Step 1: Choose the "maximum tolerated level" θ, the toxicity rate at the recommended dose level or MTD's (say, θ=.33 or whatever); this is a basic difference with standard design (SD).

Step 2: Choose a fixed number of patients to be enrolled; usually n = 19-24; this is another difference with SD (where the number of patients needed is variable).

Step 3: The CRM uses binary response (DLT or not); Let Y be the binary response such that Y=1 denote the occurrence of a pre-defined DLT. Let

$p(x) = Pr[Y=1|x]$ and

$\text{logit}\,[p(x)] = \log\{p(x)/[1-p(x)]\}$

The next step is to choose a statistical model representing the relationship between Y and dose level; for example, it could be described by the logistic (or probit) model :

$\text{logit}\,[p(x)] = \alpha + \beta x$

where x is the log of the dose d; or x is dose d.

Step 4: Use the baseline response and toxicity,adverse-effect rate (dose = 0) to calculate and fix the "intercept" α.

Step 5: Under the Bayesian framework, choose a prior distribution for the "slope" β; for example, "unit exponential" - one with probability density function $g(\beta) = \exp(-\beta)$

Step 6: From the model: logit [p(x); β] = α + βx, with β placed at "the prior mean" and set p(x) equal to the target rate θ, solve for dose x. This is dose for the first patient, a dose determined to reflect the current belief of the investigator or doctor as the dose level that produces the probability of DLT closest to the target rate θ - the "maximum dose with an acceptable and manageable safety" . This step fits the "principle of good medicine"! -the patient is treated at the MTD.

Step 7: After the first patient's toxicity/adverse-effect result becomes available, the "posterior distribution" of β is calculated and the posterior mean of β is substituted in logit $[p(x); \beta] = \alpha + \beta x$.

The next patient is treated at the dose level x whose probability p(x) is the target rate θ (with calculated posterior mean of β). This step is repeated in subsequent patients every time toxicity/adverse-effect result becomes available and the posterior distribution of β is re-calculated.

Finally the MTD is estimated as the dose level for the hypothetical (n+1)th patient; n has been pre-determined, usually 19-24.

The strength of the CRM are its three properties: (1) it has a well-defined goal of estimating a percentile of the dose-toxicity relationship, (2) it should converge to this percentile with increasing sample sizes, and (3) the accrual is pre-determined. The standard design does not have these characteristics.

Phase II clinical Trials

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 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 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.

To think about this using a mathematical terminology: "Response" is necessary but might not be sufficient to conclude that a drug is benefitting the patients.

Secondly, 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.

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 already se 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.

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.

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 objective of this type of randomized trials is to help in the interpretation of a poor response rate of the investigational agent. Is it really poor?

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…

Designs 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.

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:**

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

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

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

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.

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.

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.

TWO-STAGE PHASE II TRIALS

This part covers a very special form of phase II clinical trials: two-stage design.

A small group of patients are enrolled in the first stage; and the enrollment of another group of patients in stage 2 is "conditional" on the outcome of the first group.

The activation of the second stage depends on an adequate number of responses observed from the first stage.

Rationale:

Why two stages? We do not want to enroll a large group of patients (in conventional one-stage designs) when not sure if the treatment is effective.

Aim of Two-stage Design:

Do not activate the second stage if the first group/stage shows that the treatment is not effective

There are more than one method – some are recent, but the most popular method is "two-stage Simon's Design".

This design uses a "computer search" to meet certain optimal requirement; it does require some special program; research organizations and health centers have this software.

Phase I trials provide information about the MTD; it is important because most cancer treatments must be delivered at maximum dose for maximum effect.

Patients may die from toxicity or side effects and, if not treated "enough", they might die from the disease too. Phase I trials provide little or no information about efficacy; patients are diverse with regard to their cancer diagnosis and are treated at different doses - only 3 or 6 at a dose – even one at a dose by fast-track design.

A phase II trial of a cancer treatment is an uncontrolled trial (most trials of phase II are one-arm, open-label) to obtain an estimate of the "degree of anti-tumor effect". The proportion of patients who "tumors shrink by at least 50% which lasts for at least 4 weeks" is often the primary endpoint. The aim is to see if the agent has sufficient activity against a specific type of tumor to warrant its further development (to combine with other drugs in a phase III trial comparing survival results with a standard treatment).

It is desirable to find out about the anti-tumor capacity of new agents and to determine if a treatment is sufficiently promising to warrant a major controlled evaluation. However, recall that 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.

The first aim is to benefit the patients

The problem is that, if the agent has no or low anti-tumor activity, patients in the phase II trial might die from the disease. Therefore, we often wish to minimize the number of patients treated with an ineffective drug.

Early acceptance of an highly effective drug is permitted but very rare in phase II trials; however, it is ethically imperative to exercise early termination when the drug has no or low anti-tumor activity.

GEHAN'S TWO-STAGE DESIGN

- The first and most commonly used design for many years was developed by Gehan (1961); this design has been popular- much more so in the 70's.
- The same Gehan who invented the "generalized Wilcoxon test" (two years later).
- It has two stages; the primary aim Gehan's design is to estimate the response rate - two-stage feature is an option for "screening" of agents worthy of further development.

GEHAN'S DESIGN

- The first stage enrolls 14 patients; if no responses are observed, trial is terminated;
- If at least one response is observed among the first 14 patients of stage 1, the second stage of accrual is activated in order to obtain an estimate of the response probability having a pre-specified standard error (SE).
- Patients from both stages are used in the estimation of the response rate.

RATIONALE FOR EARLY TERMINATION

The probability of observing no responses among 14 patients is less than .05 if the response probability is greater than 20%

$$(.20)^0(.80)^{14} = .044$$

$$\pi^0(1-\pi)^{14} \leq .044 \text{ if } \pi \geq .20$$

Implicitly, response rates over 20% are considered promising for further studies.

If no responses are observed in the first stage of 14 patients, trial is terminated. It is stopped because we can conclude that π<.2 or 20%, not worthy of further investigation.

SECOND STAGE

The number of patients n_2 accrued in the second stage depends on the number of responses observed in the first stage (because patients from both stages are used in the estimation of the response rate) and the pre-determined standard error;

$$SE(p) = \sqrt{\frac{\pi(1-\pi)}{14+n_2}}$$

CRITIQUES

- **The size of the first stage is "fixed"; it may not be optimal for the underlying aim of early termination "if the drug has no or low anti-tumor activity".**
- **It serves investigators & drug companies more - not the patients enrolled in the trial.**
- **It does not help to achieve the aim of early termination when the drug has no or low anti-tumor activity, a very important ethical concern**

For example, even a poor drug with a true response probability of 5%, there is a 51% chance of obtaining at least one response in the first 14 patients and, therefore, activating the second stage accrual.

$$\Pr(\text{at least 1 response}) = 1 - (.05)^{0}(.95)^{14} = .51$$

TWO-STAGE SIMON'S DESIGN

In phase II trials, the ethical imperative for early termination occurs when the drug has low anti-tumor activity; the "Two-stage Simon's Design" is currently a popular tool to achieve that.

The trial is conducted in two stages with the option to stop the trial after the first or after the second stage (and not recommending the agent for further development).

The basic approach is to **"minimize" <u>expected sample size when the true response is low</u>** - say, less than some pre-determined uninterested level.

PROBLEM'S STATISTICAL SETUP

- **Endpoint: (binary) Tumor Response: yes/no**
- **Null Hypothesis: H_0: $\pi = \pi_0$; π is the true response (say, proportion of patients whose tumors shrink by at least 50%) and π_0 is a pre-determined uninterested/undesirable level.**
- **Alternative Hypothesis: H_A: $\pi = \pi_A$; π_A is some desirable level that warrant further development.**
- **Type I and type II errors: α and β**
- **Basis for decision: minimize the number of patients treated in the trial if H_0 is true.**

THE DESIGN

- **Two stages (the design could have more than two stages, but less practical and not often used)**
- **Enroll n_1 patients in stage 1; the trial is stopped if r_1 or fewer responses are observed, goes on to the second stage otherwise.**
- **Enroll n_2 patients in stage 2; the trial is not recommended for further development if a total of r (of course: $r > r_1$) or fewer responses are observed in both stages.**

PET: PROB OF EARLY TERMINATION

$$\mathbf{PET}(\pi) = \mathbf{B}(\mathbf{r}_1; \mathbf{n}_1, \pi)$$

where B(.) denotes the cumulative Binomial probability, and π is the true response.

That's the probability to have r_1 responses or fewer in first stage.

EXPECTED SAMPLE SIZE

$$EN(\pi) = n_1 + [1 - PET(\pi)] * n_2$$

Both PET(π), the probability of early termination, and EN(π), expected sample size, are function of the response rate π.

DECISION NOT TO RECOMMEND

- The drug may not be recommended either after 1 or 2 stages; the probability is PNC(π).
- We will terminate the trial at the end of the first stage and not recommending the drug if r_1 or fewer responses are observed, or
- We will not recommend the drug at the end of the second stage if r ($r = r_1 + r_2$) or fewer responses are observed; some of the responses after the first (r_1) may come from stage 1, some from stage 2.

PNC: PROB OF NOT RECOMMENDING

The drug is not recommended if the trial is terminated early (i.e. fewer than r_1 responses are observed in the first stage) OR fewer than $r = r_1 + r_2$ are observed; some of the responses (say, x) may come from stage 1 and some from stage 2 (say, r-x).

$$\mathrm{PNC}(\pi) = \mathrm{B}(r_1; n_1, \pi) + \sum_{x=r_1+1}^{\min[n_1, r]} \mathrm{b}(x; n_1, \pi)\mathrm{B}(r - x; n_2, \pi)$$

TWO TYPES OF ERRORS

$$(\text{Type I errors}) : \alpha = 1 - PNC(\pi_0)$$

$$(\text{Type II errors}) : \beta = PNC(\pi_A)$$

SIMON'S APPROACH

The design approach considered by Simon is to specify the parameters π_0, π_A, α, and β; then determine the two-stage design that satisfies the errors probabilities α and β and minimizes the expected sample size EN when the response probability is π_0; i.e. minimizing EN(π_0).

“AE” MONITORING IN PHASE II TRIALS

In phase II trials, "efficacy" is the "outcome of interest" whereas "safety" is embedded to serve as "stopping rule"

In planning a clinical trial of a new treatment, we should always be aware that severe, even fatal, side effects are a real possibility. If the accrual or treatment occurs over an extended period of time, we must anticipate the need for a decision to stop the trial – at any time - if there is an excess of these unwanted events.

FOCUS ON PHASE II TRIALS

- In phase I trials, toxicity may be considered the "Outcome Variable" and dose escalation plan serves as the stopping rule.
- In phases II trials, we start to focus on efficacy which requires conventional analysis at the end. "Response" becomes the Outcome Variable; however, toxicity (or other adverse effects) may still turn out to be a problem during the trial.
- The monitoring of side-effect events is a separate activity that may require special consideration

Two-stage designs stops trials for "Efficacy Reason"; here we want rules to stop trials for "Safety Reason" – both, not treated enough or excessive adverse effects, put patients at risk.

Two-stage Designs are optional (decision by investigators) but stopping rules for safety reason are "required" by regulatory affairs agencies/entities.

SEQUENTIAL PROCESS

- The most common method for monitoring toxicity or adverse effects is to design a formal sequential "stopping rule" based on the limit of acceptable side-effect rates; the sequential nature of the rule allows investigators to stop the trial as early as the evidence that the event's rate becomes excessive.
- In multi-site trials, a "data safety and monitoring board" (DSMB) is required; in local phase II trials, it's the statistician's responsibility to form the rule and the Clinical Trial Office's staff is responsible for its implementation.

The most common method for monitoring toxicity or adverse effects is to design a formal sequential "stopping rule"; and a sequential stopping rule could be formed in two different ways:

(i) a Bayesian approach to evaluating the proportion of patients with side effects, or

(ii) a Hypothesis testing approach - using the sequential probability ratio test (SPRT) - to see if the normal, acceptable side-effect's rate has been exceeded.

HYPOTHESIS TESTING

- Let start with the hypothesis testing approach because it's more "conventional" (with statisticians)
- Let π be the proportion of patients with adverse side effects; the problem becomes testing for the null hypothesis H_0: $\pi = \pi_0$ against alternative H_A: $\pi = \pi_A$; where π_0 is the normal baseline side-effect's rate (say, 5%) and is the "maximum tolerated rate" (say, 20%) - anything over that are considered excessive.
- Baseline rate is determined/estimated from historical data but the setting of a "ceiling" rate is subjective- by investigator.

STATISTICAL MODEL

- We can assume that the number of adverse events "e" follows the usual Binomial Distribution B(n, π), where n is the total number of patients.
- This leads to the log likelihood function: $L(\pi;e) = \text{constant} + e \ln\pi + (n-e) \ln(1- \pi)$

SEQUENTIAL PROBABILITY RATIO TEST

- **When "e" adverse events are observed out of n "evaluable" patients, the test for null hypothesis H_0: $\pi = \pi_0$ against alternative H_A: $\pi = \pi$ can be based on "the log likelihood ratio statistic" LR_n: $LR_n = e(\ln \pi_A - \ln \pi_0) + (n-e)[\ln (1-\pi_A) - \ln (1-\pi_0)]$**
- **In conventional sequential testing, the statistic is calculated as each patient's evaluation becomes available and plotted against n; the trial is stopped if the plot goes <u>outside predefined boundaries</u> which depends on pre-set type I and type II errors.**

SEQUENTIAL STOPPING RULE

- In testing for null the hypothesis H_0: $\pi = \pi_0$ against the alternative H_A: $\pi = \pi_A$, the decision is:

(i) to stop the trial and reject H_0 if $LR_n \geq \ln(1-\beta)-\ln\alpha$

(ii) to stop the trial and accept H_0 if $LR_n \leq \ln\beta-\ln(1-\alpha)$

(iii) continue the study otherwise

- In (i) there are too many events and in (ii) there are too few events - enough to make a decision.

SIDE EFFECTS MONITORING

- **We do not stop the trial because there are too few events; we only stop the trial early for an excess of side effects, that is when:**
  $e(\ln\pi_A - \ln\pi_0)+(n-e)[\ln(1-\pi_A)- \ln(1-\pi_0)]\geq \ln(1-\beta)-\ln\alpha$
- **The lower boundary is ignored; trial continues**
- **Solving equation for "e" yields for upper boundary**
- **We can also solve the same equation for n .**

RESULT

Stop the trial as soon as n, as a function of e, satisfies the following equation:

$$n(e) = \frac{\ln(1-\beta) - \ln\alpha + e[\ln(1-\pi_A) - \ln(1-\pi_0) - \ln\pi_A + \ln\pi_0]}{\ln(1-\pi_A) - \ln(1-\pi_0)}$$

n(e) is the number of evaluable patients for having e of them with adverse effects.

Rule: To stop the trial when we have “e” adverse effects before reaching a total of “n(e)” patients.

INTRODUCTION TO MULTIPLE DECISIONS

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 <u>extreme</u>.

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 (test is significant but null hypothesis is true); For example, suppose the null hypothesis is true and we perform 100 tests---each has a 0.05 probability of resulting in a Type I error; then 5 of these 100 tests would be wrongly statistically significant simply as the results of Type I errors (false positives).

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.

EXAMPLE: 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. If results indicate a marked superiority of a new treatment over the placebo; trial should be stopped so future patients could get this better treatment

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 α, say $\alpha = .05$
- Rule: After the nth 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.
- O'Brien & Fleming showed that $P(N,\alpha)$ is approximately the $(1-\alpha)$th percentile of the Chi-square distribution with 1 degree of freedom - almost independent of N.

A Simple Application of Procedure for N=2:

(1) Use the value of the 95th percentile of the Chi-square with 1 degree of freedom, i.e. 3.84)

(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) For N=2, we can use usual Chi-square tests at .5% and 4.5% respectively.

EXERCISES

#1. When Phase I Cancer Trial following the Standard Design reaches a dose level with a toxicity rate 0f 40%, what is the probability that it would pass to the next higher dose?

#2. Consider a Phase I Cancer Trial with three doses (and toxicity rates): 15%, 35%, and 55%. Using the Standard Design, what is the probability that a subject would be treated at the last dose?

#3. Suppose we are conducting a small phase II trial with N=25 patients. Form a sequential stopping rule with these two parameters: $\pi_0 = .05$ and $\pi_A = .20$ by applying the SPRT (Sequential Probability Ratio Test).