

# Biostatistical Aspects of Rational Clinical Trial Designs

Elizabeth Garrett-Mayer, PhD  
Associate Professor of Biostatistics  
Hollings Cancer Center  
Medical University of South Carolina

Pancreatic Cancer Course  
The Banbury Center, Cold Spring Harbor, NY  
June 21, 2011

# Types of Research Studies in Cancer

- Basic Science
- Translational
- Clinical
  - Exploratory/Pilot/Correlative
  - Phase I
  - Phase II
  - Phase III
  - Other: e.g. prevention, survivorship
- Epidemiological

# Phases of Drug Development

- Phase I
  - Dose finding
  - Usually designed to find the highest safe dose.
  - 12-30 patients
- Phase II
  - Preliminary efficacy and safety
  - Generally not 'head to head' comparison
  - 20-80 patients
- Phase III
  - Definitive comparative trial against the standard of care
  - Usually hundreds or thousands of patients

# Clinical Trials: the beginning

- Write a clinical trial protocol
- Usually 70-180 pages
- ***Not*** like writing a grant
- Every detail spelled out: no page limit!
- There are standard templates that can/should be used.

# Imagine....

- You are principal investigator (PI) of a clinical trial
- In the middle of the trial, you change careers
- You are now an astronaut and fly to the moon
- Meanwhile, a new patient is enrolled.
- The **new** PI needs to know:
  - How should the patient be assigned to a dose?
  - How should dose modifications occur?
  - What measurements should be taken and when?
  - What are the definition of the primary and 2ndary outcomes?
  - Who and how are the data to be reviewed for safety and efficacy?

# Statistical design and development of clinical trials

- Statistical considerations permeate the design and analytic plan
- Requires interaction with your statistician
  - call early!
  - before you have “fixed” the design
  - bad: “i have almost finished writing the protocol, and then i will send to you to insert a statistical plan”
- Really, *we are here to make your life easier*

# Where do I find this statistician?

- Academic cancer centers have biostatistics cores or biostatistics shared resources
- It is the role of these biostatisticians to help design clinical trials
- Find them!
- Other places:
  - University settings usually have biostatistics departments or divisions
  - Pharma will have biostatisticians on site or have biostatistical consultants available
- If your institution does NOT have biostats support, tell them they MUST HAVE IT!!!

# Design of Clinical Trials: Striking a Balance

- Answer the question (correctly)
  - Control risk of errors in conclusions
- Minimize potential harm and maximize potential benefit
  - Limit number of participants treated at sub-therapeutic doses
  - Limit number of participants treated with ineffective therapy or exposed to toxicity
- Maximize feasibility
  - Make it simple enough to carry out
  - Writing a detailed protocol can help avoid unforeseen feasibility issues



# Statistical Considerations: 5 part process

I. Stating research aims

II. Determining your outcome measures

III. Choosing the experimental design

IV. The analytic plan

V. Sample size justification

# Motivating Example

- *Phase II trial of induction gemcitabine/oxaliplatin/cetuximab (GOC) followed by intensity modulated radiotherapy (IMRT) with capecitabine to improve resectability in borderline and frankly unresectable pancreatic cancer*
- Principal Investigator: Nestor Esnaola
- Single arm study
- Treatment plan:
  - Patients are treated with GOC for six 14-day cycles.
  - If resectable, taken to surgery
  - If not, radiochemotherapy (IMRT + capecitabine (RCT)) and restaged

# I. Stating research aims

- Authors devised a protocol, beginning with research aims
- Aims should be concrete and include measurable outcomes
- Bad examples:
  - To evaluate the effect of GOC on cancer.
  - To see if GOC + RCT improves cancer outcomes
  - To determine the safety profile of GOC + RCT
- ***What is wrong with these aims?***
  - ***what does “effect” mean? what kind of cancer, in what patients?***
  - ***“Improves” compared to what? what is the outcome of interest?***
  - ***what does a “safety profile” mean?***
- ***Think about how you are going to determine if this treatment approach works or not***

# I. Stating research aims

- Better examples:
  - **To evaluate the 6 month progression-free survival of GOC + RCT in patients with locally advanced, unresectable or borderline resectable, non-metastatic adenocarcinoma of the pancreas when treated with neoadjuvant GOC with or without RCT followed by definitive surgery.**
  - **To determine the tolerance of this regimen, defined as the proportion of patients who follow the treatment plan.**
- ***Keywords for primary outcome:***
  - *determine, estimate, evaluate, describe*
  - *efficacy, safety*

# Devising your aims

- Generally, there is ONE primary aim and your study is designed to address the primary aim
- Common (generic) aims per phase:
  - Phase I: primary aim is finding the “recommended” dose
  - Phase II: primary aim is determining if there is sufficient efficacy
  - Phase III: primary aim is to determine which of two (or more) treatment combinations yields the longest overall survival
- Secondary aims:
  - important, but do not drive the design
    - pharmacokinetics
    - pharmacodynamic (e.g., methylation)
    - response
    - safety
    - change in gene expression

# Aims and Hypotheses

- Aims are often accompanied by hypotheses.
- Stating the hypothesis to be tested can be a useful guide for the analytic plan:
- “The 6 month PFS will be at least 70%”

## II. Determining your outcome measures

- The outcome measure will depend on the **parameter of interest**
- Examples of possible parameters of interest in phase II:
  - response rate
  - Median progression-free survival
  - 6 month progression-free survival rate
- *Synonyms: outcome, endpoints*
- **Aim  $\neq$  endpoint**
- What is an endpoint or outcome?
  - patient-level measure of “effect” of interest
  - *measured on each patient in the study*
  - *it is QUANTIFIABLE*

# Parameter of interest vs. outcome

Parameter of interest	Outcome
Response rate: proportion of patients with CR or PR	Response (CR or PR)
Median overall survival	Time from enrollment to death (or last follow-up)
6 month overall survival	Time from enrollment to death (or last follow-up)
Mean change in quality of life	Difference in quality of life scores from baseline to follow-up



## II. Determining your outcome measures

- Example:
  - **Parameter of interest is the 6 month progression-free survival rate**
  - endpoint = PFS at 6 months
  - objectively defined: the tumor has not increased by 20% or more comparing the baseline to the 6 month tumor measurement.
  - Each patient is determined to either be or not be “progression-free” at 6 months.
  - BINARY endpoint in this example

# The following are NOT endpoints

- These are estimates of parameters:
  - response rate
  - median survival
  - AE rate
  - safety profile
- These describe the time course of the study in some way (don't let the term 'endpoint' confuse you):
  - length of time of treatment
  - time until patient goes off-study
  - length of study

# Determining clinical outcomes: RECIST criteria

- Definitions of response, stable disease and progression are not quite as 'simple' as they may seem in solid tumors
- RECIST: Response Evaluation Criteria in Solid Tumors.
- Version 1 is from 2000, Version 1.1 published in 2008.
- Key features:
  - Definitions of minimum size of measureable lesions
  - Instructions on how many lesions to follow
  - Use of **unidimensional** measures for overall tumor burden
- See Eisenhauer et al., Eur J of Cancer (2009), 45, 228-247.

# Definitions (briefly\*)

- **Complete Response:** disappearance of all target lesions.
- **Partial Response:** at least a 30% decrease in the sum of the diameters of the target lesions, taking as a reference the baseline sum of diameters
- **Progressive Disease:** At least a 20% increase in the sum of diameters of target lesions. Increase must constitute at least 5mm absolute increase.
- **Stable Disease:** Shrinkage < 30% or increase < 20%.

\* based on target lesions and ignoring lymph node criteria for simplicity

### III. Choosing the experimental design

- Based on the aims and the outcome, a design can be identified.
- Other considerations
  - patient population
  - **accrual limitations**
  - previous experience with the treatment of interest in this or other populations
  - results from earlier phase studies

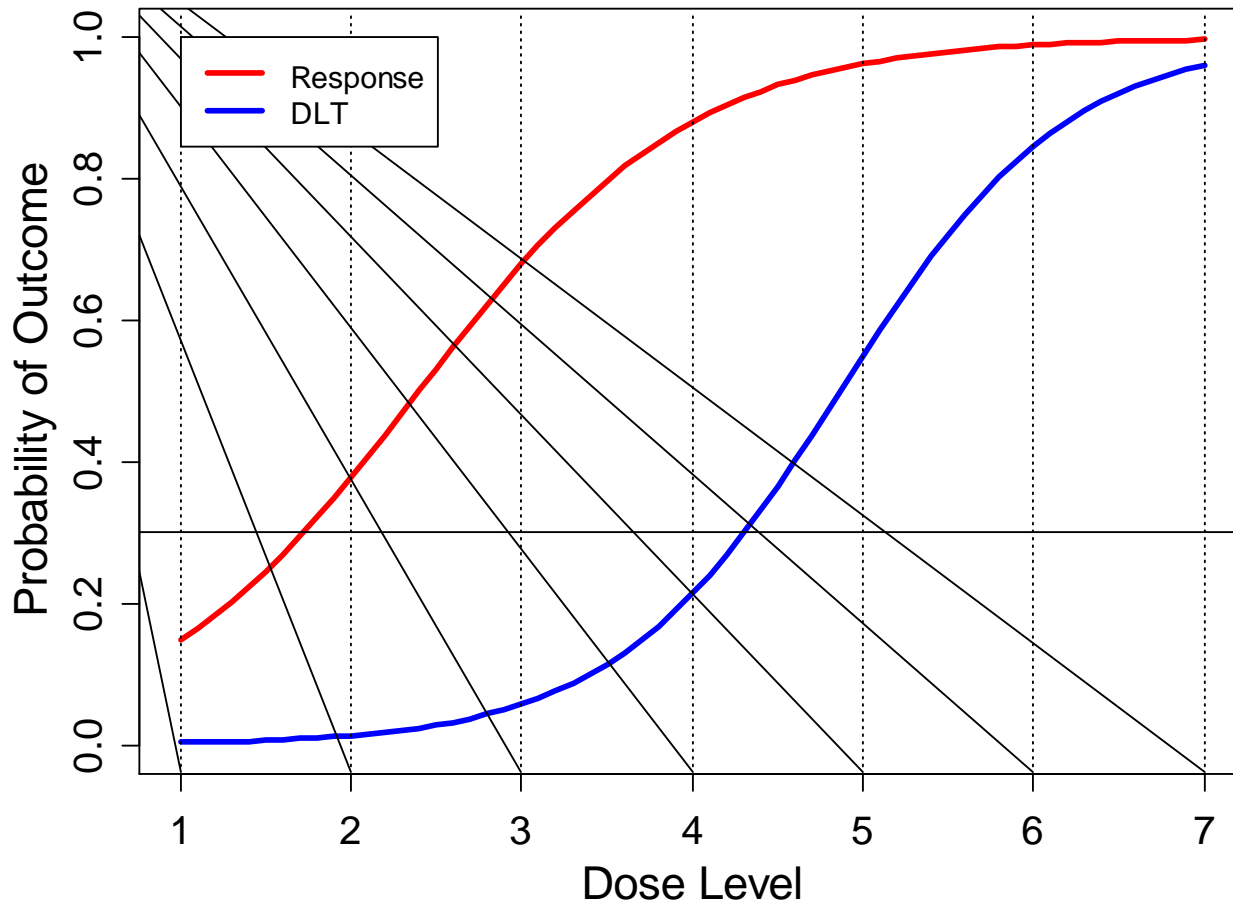
### III. Choosing the experimental design

- There are common approaches within each phase of drug development
- However, there are often many options and seemingly small details that can make big differences.
- Two common ‘philosophies’
  - Frequentist
  - Bayesian
- Buzzword: “adaptive”

# Phase I trial goals

- Classic Phase I trials:
  - Find the highest dose that is deemed safe: the Maximum Tolerated Dose (MTD)
  - DLT = dose limiting toxicity
  - Goal is to find the highest dose that has a DLT rate of x% or less (usually ranges from 20% to 40%)
- Newer Phase I trials:
  - Find the dose that is considered to be safe and have optimal biologic/immunologic effect (OBD).
  - Goal is to optimize “biomarker” response within safety constraints.

# Classic Phase I Assumption: Efficacy and toxicity both increase with dose



DLT =  
dose-  
limiting  
toxicity



# Classic Phase I approach: Algorithmic Designs

- “3+3” or “3 by 3”
- Prespecify a set of doses to consider, usually between 3 and 10 doses.

Treat 3 patients at dose K

1. If 0 patients experience DLT, escalate to dose K+1
2. If 2 or more patients experience DLT, de-escalate to level K-1
3. If 1 patient experiences DLT, treat 3 more patients at dose level K
  - A. If 1 of 6 experiences DLT, escalate to dose level K+1
  - B. If 2 or more of 6 experiences DLT, de-escalate to level K-1

- MTD is considered highest dose at which 1 or 0 out of six patients experiences DLT.
- **Confidence in MTD is usually poor.**

# “Novel” Phase I approaches

- Continual reassessment method (CRM)  
(O’Quigley et al., Biometrics 1990)
  - Many changes and updates in 20 years
  - Tends to be most preferred by statisticians
- Other Bayesian designs (e.g. EWOC) and model-based designs (Cheng et al., JCO, 2004, v 22)
- Other improvements in algorithmic designs
  - Accelerated titration design (Simon et al. 1999, JNCI)
  - Up-down design (Storer, 1989, Biometrics)

# CRM: Bayesian Adaptive Design

- Dose for next patient is determined based on toxicity responses of patients previously treated in the trial
- After each cohort of patients, posterior distribution is updated to give model prediction of optimal dose for a **given level of toxicity** (DLT rate)
- Find dose that is most consistent with desired DLT rate
- Modifications have been both Bayesian and non-Bayesian.

# CRM Designs

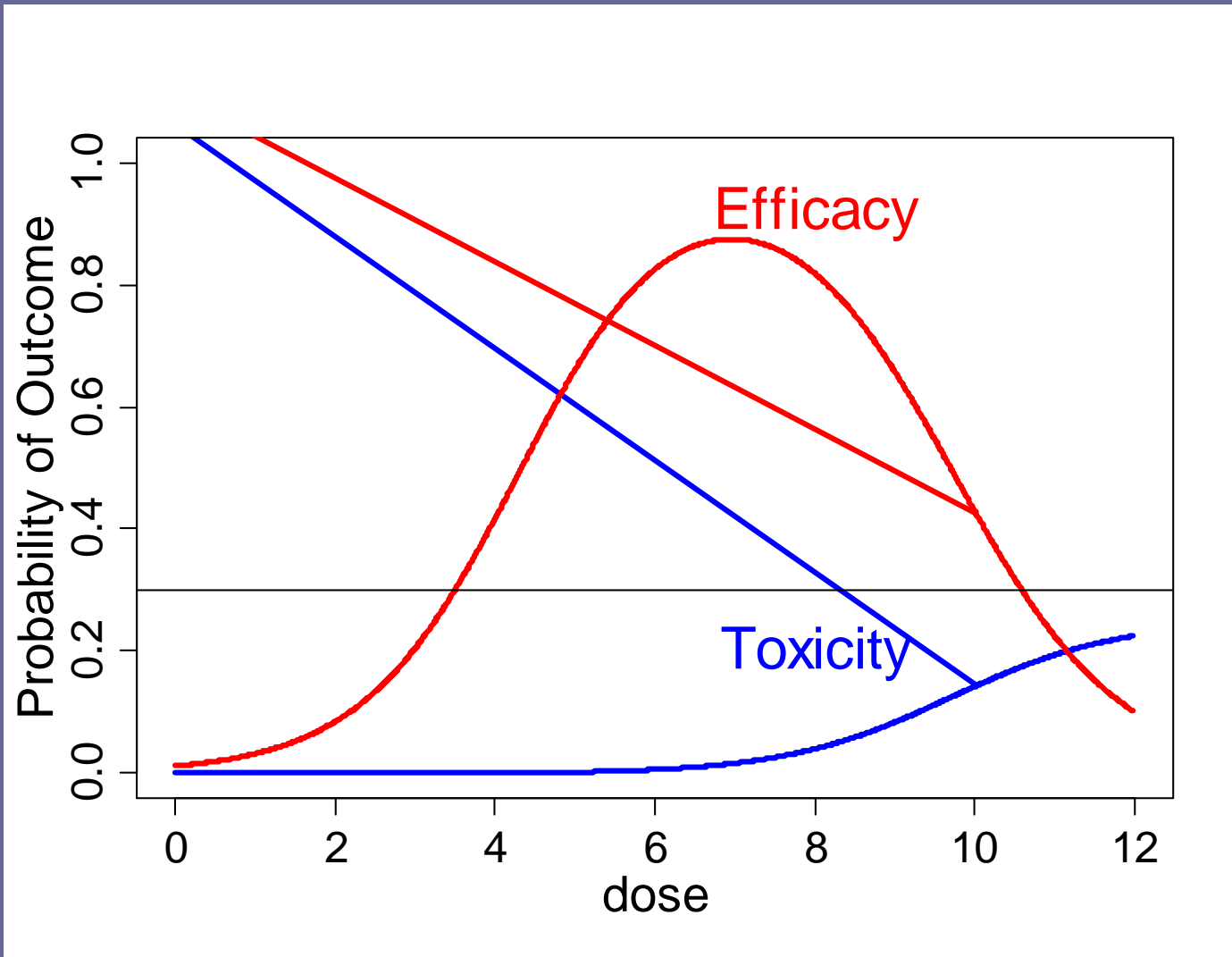
- Underlying mathematical model
- Doses can be continuous or discrete
- Compared to the '3+3' the CRM is
  - safer: fewer patients treated at toxic doses
  - more accurate: selected MTD is closer to the true MTD
  - more efficient: more patients are treated at doses near the MTD.
- Disadvantages:
  - requires intensive involvement of statistician because future doses depend on model prediction
  - need more lead time: statisticians need time (weeks?) to select the appropriate CRM design for a given trial
    - simulations
    - need to ensure that it will “behave” in a smart way
- “TiTE-CRM” was developed to allow incorporation of long-term toxicity evaluation.

## *New paradigm: Targeted Therapy*

How do targeted therapies change the early phase drug development paradigm?

- Not all targeted therapies have toxicity
  - Toxicity may not occur at all
  - Toxicity may not increase with dose
- Targeted therapies may not reach the target of interest
- Implications for study design: Previous assumptions may not hold
  - Does efficacy increase with dose?
  - Endpoint may no longer be appropriate
  - Should we be looking for the MTD?
  - What good is phase I if the agent does not hit the target?

# Possible Dose-Toxicity & Dose-Efficacy Relationships for Targeted Agent



# Phase II

- Provide preliminary information on whether a treatment is efficacious
- Provide preliminary data about the relationship between dose and efficacy.
- **Often controlled but**
  - **They are small: generally cannot find large differences in treatment effects**
  - **Their endpoints are “short-term”**
    - **Phase II endpoint: response**
    - **Phase III endpoint: overall survival**
- Often unblinded

# GOC + RCT trial

- “This is a single arm phase II trial to evaluate the 6 month PFS rate in patients with borderline and frankly unresectable pancreatic cancer.”
- The goals: determine if the 6 month PFS rate is significantly better than 50%.
- Study design:
  - A single arm, single stage study was designed
  - The null hypothesis is that the 6 month PFS rate is 50%
  - The alternative hypothesis is that the 6 month PFS rate is 70%
- Alternatives:
  - Randomized phase II design
  - Early stopping for futility (e.g. Simon two-stage).



# Why randomized phase II?

- Classic phase II studies
  - Single arm study where results are compared to historical control rate (or other parameter).
  - Problem: this is not always satisfying
    - Requires patient populations to be comparable
    - Might not have information to derive control rate (e.g. disease progression is of interest and not response)

# Why randomized phase II?

- Two most common randomized phase II studies
  - Phase II selection design (prioritization)
  - Phase II designs with reference control arm (control)
- Other phase II designs:
  - phase II/III studies
  - randomized discontinuation designs
  - Biomarker driven designs (also phase III)

# Common design of randomized phase II study

- Two parallel one arm studies (classic case)
- **Do not directly compare arms to each other**
- Compare each to “null rate”
- Example:
  - Randomized phase II: Two parallel arms in study
    - Test each treatment to see if it is better than null rate
  - Comparative (phase III) study:
    - Test to see if one treatment is better than the other treatment
    - Sample size can be 2 to >10 times greater, depending on the outcome

# Classic Randomized Phase II designs

- Phase II selection designs (Simon, 1985)
  - “pick the winner“
  - Appropriate to use when:
    - Selecting among NEW agents
  - Each arm is compared to a null rate
    - Must satisfy efficacy criteria of Simon design
    - Move the “winner” to phase III
    - Only have to pick winner if more than one arm shows efficacy
  - Can be used when the goal is prioritizing which (if any) experimental regimen should move to phase III when no *a priori* information to favor one.

# Classic Randomized Phase II designs

- Randomized Phase II designs with reference arm
  - Includes reference arm to ensure that historical rate is “on target”
  - Reference arm is not directly compared to experimental arm(s) (due to small N)
  - Can see if failure (or success) is due to incomparability of patient populations

# Prevalence of Randomized Phase II Designs

- Lots of randomized studies are calling themselves randomized phase II studies these days:
- If outcome of interest is surrogate
  - Correlative (biomarker)
  - Clinical (response)
- If sample size is relatively small but direct comparison is made
  - BEWARE: Underpowered Phase III!
- If study is comparative, but is not definitive for whatever reason.

# Important considerations in phase II

- Randomized phase II studies should not be considered “short cut” to comparative study “
- “I want to do a comparative study, but there is no way I can enroll 200 patients”
- Current research climate: many candidates! Critical to screen these because we cannot take so many forward to Phase III.
- **Important : these studies need to protect the ability to perform definitive phase III trials**

# Phase III

- If agent (or combination) succeeds in phase II, the next logical step is phase III.
- Usually designed by large companies or cooperative groups (e.g. ECOG, CALG-B).
- Comparative trial
  - Two or more arms
  - Standard outcome is overall survival (with rare exception in cancer)
  - Goal: show a significant improvement in survival
- Generally very large and expensive
- Must be strong evidence in phase II to conclude that Phase III study will succeed



# Phase III

- Very large undertaking
  - Multicenter
    - Infrastructure
    - IRB and scientific approvals at each site
  - Talks with FDA and other regulators
  - Establishment of DSMB specifically for trials
- The statistical design is relatively simple compared to the practical issues of running a Phase III trial.
- Practical issues will often drive the design

## IV. Analytic Plan

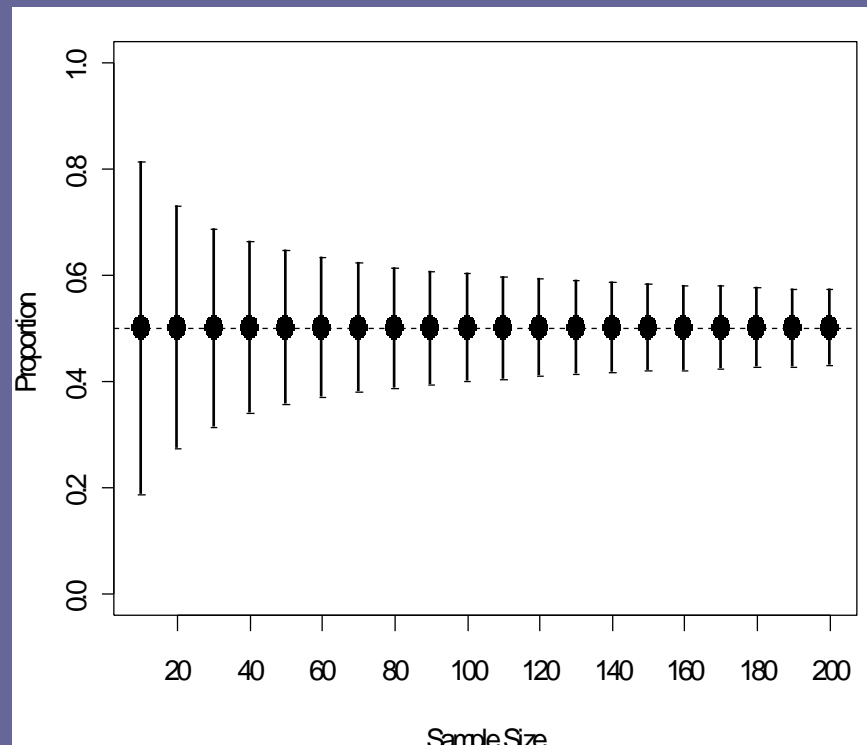
- Do you want to compare?
- Do you want to estimate?
- Do you want to test a hypothesis?
- These questions, in regards to your stated aims, will determine your analytic plan
- Recall primary aim: **To determine the 6 month progression-free survival rate.**
- Recall primary endpoint: **6 month progression-free survival.**

## IV. Analytic Plan

- The analytic plan for the primary outcome usually involves two things:
  - estimating a parameter of interest
  - testing that the parameter is different than in another setting (e.g., different treatment)
- **Estimation:** a point estimate and some measure of precision
- Example: “The 6 month PFS rate will be estimated with its confidence interval.”
  - this provides us with an estimate of the proportion of patients who are progression-free at 6 months.
  - it also provides us with a measure of precision about the estimate

# The 95% confidence interval

- an interval that contains the true value of the parameter of interest 95% of the time.
- “we are 95% confident that the true 6 mo. PFS rate lies in this interval”
- Example: below shows examples where **observed** rate is 0.50. **95% confidence interval width depends on the sample size**
- Depending on the sample size, we have greater or less precision in our estimate



## IV. The analytic plan

- **Hypothesis testing:** Determining if the treatment is worthy of further study.
- Recall our hypotheses:
  - The 6 month PFS rate of patients in this study will be at least 70%.
- What is a sufficiently LOW 6 month PFS rate that we are not interested in further pursuit?
- Based on the study team's experience, a 6 month PFS rate of **50%** is too low to warrant further study of this treatment approach.

## IV. Analytic Plan

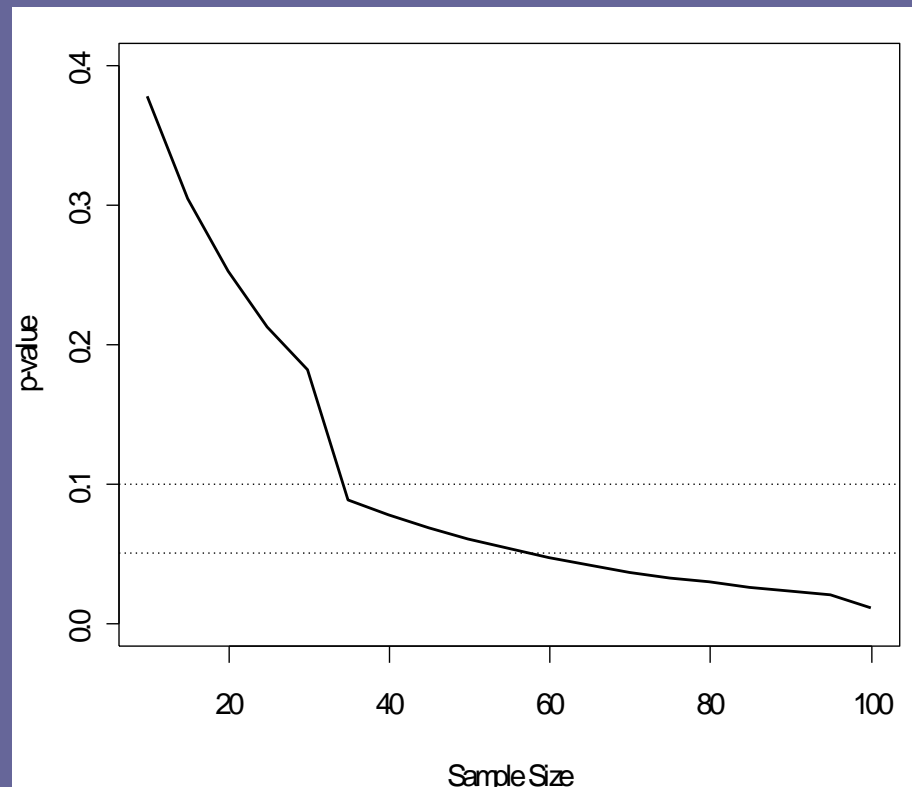
- We perform a hypothesis test:
  - $H_0: p = 0.50$  (null)
  - $H_a: p = 0.70$  (alternative)
- This test is performed using an exact binomial procedure.
- The result is a p-value that provides “evidence” to either reject or fail to reject the null hypothesis
- If this were a randomized phase II study:
  - the test is performed in each arm
  - the arms are not directly compared to one another (that is a different test)

# Recall the p-value

- p-value: the probability of observing a result as or more extreme than we saw in our study if the null hypothesis is true.
- **Small p-value:** evidence that the null is not true (“significant result”)
- **Large p-value:** not sufficient evidence to reject the null (“not significant”)
- Threshold for significance? we usually think of 0.05, but in phase II, often use 0.10.

# P-value depends on the sample size

- For the same observed rate, a larger sample size will lead to a smaller p-value
- Example: With an **observed** PFS rate of 0.61 (24/39), the p-value gets smaller as the sample size increases
- **Important point:** a large p-value does not always mean that “the null is true”. It may mean that the sample size was not large enough to reject the null









## IV. Analytic Plan

- Depends on the design and the goals
- Example is a Phase II trial
  - single arm approach to analysis
  - compare to historical 6 month PFS rate (e.g., 0.50)
- Phase I studies
  - often the analysis plan is descriptive
  - rare to see hypothesis testing (for primary aim)
- Phase III studies
  - head to head comparison of groups
  - Hazard ratio compares event rates per group
  - Time to event methods are required:
    - Log rank test
    - Cox regression

# V. Sample size justification

- Two basic approaches
  - power (most common)
  - precision
- Recall:
  - Limit number of participants treated at sub-therapeutic doses
  - Limit number of participants treated with ineffective therapy or exposed to toxicity
- But, also we need to enroll enough patients to achieve our aims
- Balancing act:
  - Too few patients: you cannot answer the question
  - Too many patients: you have wasted resources and potentially exposed patients to an ineffective treatment unnecessarily
- **Most commonly motivate sample size by a hypothesis testing approach**

# Refresher of alpha, beta and power

	$H_0$ is True	$H_0$ is NOT True
Accept $H_0$		 <span style="color: red;">Type II error</span>
Reject $H_0$	 <span style="color: red;">Type I error</span>	

$\alpha$  = probability of Type I error (level of significance)

$\beta$  = probability of Type II error

$1 - \beta$  = Power

# V. Sample size justification

- Usual motivation: hypothesis testing
- Power = the probability of rejecting the null if it is false
- If a study is “underpowered”, it is too small to detect a clinically meaningful difference
- Example:  $H_0: p=0.50$  vs.  $H_a: p=0.70$ 
  - this is the assumed “clinically meaningful” difference
  - Investigators chose power of 0.80 (beta = 0.20)
  - alpha (one-sided) was chosen to be 0.05
- Other design issues
  - Interim analyses require larger sample sizes (more later)
  - Stopping rules for ‘futility’ can be hard to implement when you need to wait to determine outcome per patient (e.g., 6 month PFS takes 6 months from enrollment per patient to determine).

# “Plug and Chug”

- With power of 80% and one-sided alpha of 0.05, and  $H_0$  and  $H_a$ , a one-stage design was selected.

A single stage design was chosen. To achieve a power of 80% with a one-sided alpha assuming a null 6 month PFS of 50% and an alternative rate of 70%, 39 patients need to be enrolled. If 25 or more of the patients are progression-free at their 6-month visit, then we will reject the null hypothesis at the 5% level and conclude that the treatment approach is worthy of further study.

# “Plug and Chug” with interim look

- With power of 80% and one-sided alpha of 0.05, and  $H_0$  and  $H_a$ , a **Simon two-stage design** could have been adopted.

The Simon's two stage design used is defined as follows. Our null hypothesis is that the 6 month PFS rate is 50% and our alternative hypothesis is that it is 70%. At the first stage we will enroll 15 patients. We will close accrual to an arm if  $\leq 8$  patients are PF at 6 months. If 9 or more are progression-free at 6 months, then the study will remain open for an additional 28 patients. The treatment approach will be considered promising if at least 27 patients are progression-free in 43 patients. This study has power of 80% and a one-sided alpha of 5%.

- The sample size per arm will be 15 patients or 43 patients (depending on early stopping)

## V. Sample size justification

- Hypothesis testing is not always the way to go
- Sometimes estimation is sufficient (but not always! it is not an 'escape route')
- In that case, sample size can be justified by precision
- Example: with 39 patients, we will be able to estimate the 6 month PFS rate with a 90% confidence interval with half-width no greater than 0.16.
- Difficult part: is 0.16 half-width sufficiently precise? how to rationalize that?

# Sample size is generally chosen based on

1. budget
2. expected accrual
3. the clinical effect size of interest
4. type I and type II errors
5. 3 and 4
6. all of the above



# Feedback loop

- The process is actually not completely linear as stated
- Examples:
  - Design issues may cause you to change your outcome or restate your aim
  - Accrual limitations may cause you to change the design
- “Dynamic process”

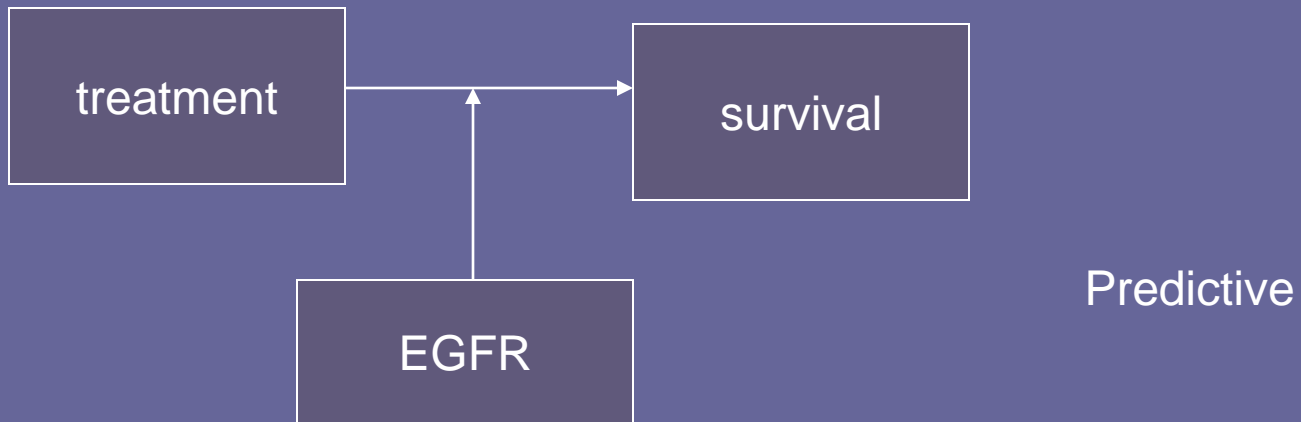
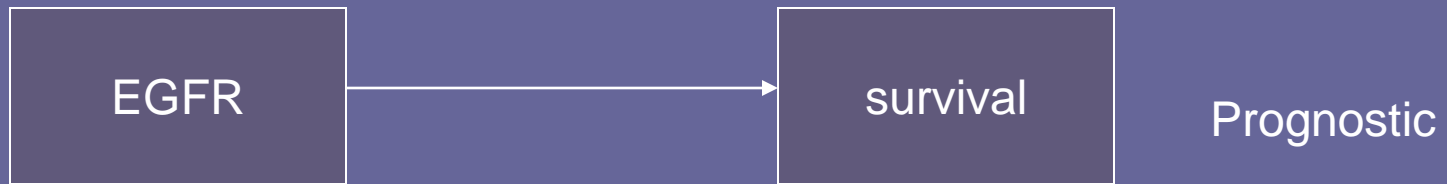
# Additional aims (correlatives, etc.)

- VERY important aims!
- Not discussed here due to space/time.
- Same principles apply for stating aims, determining outcomes, writing analytic plan
- Usually power/sample size is less of a concern for secondary aims
- “correlative” does not mean you can be vague!
  - these need to be well-conceived
  - often on biopsy tissue, pre post design
  - will you really learn anything?

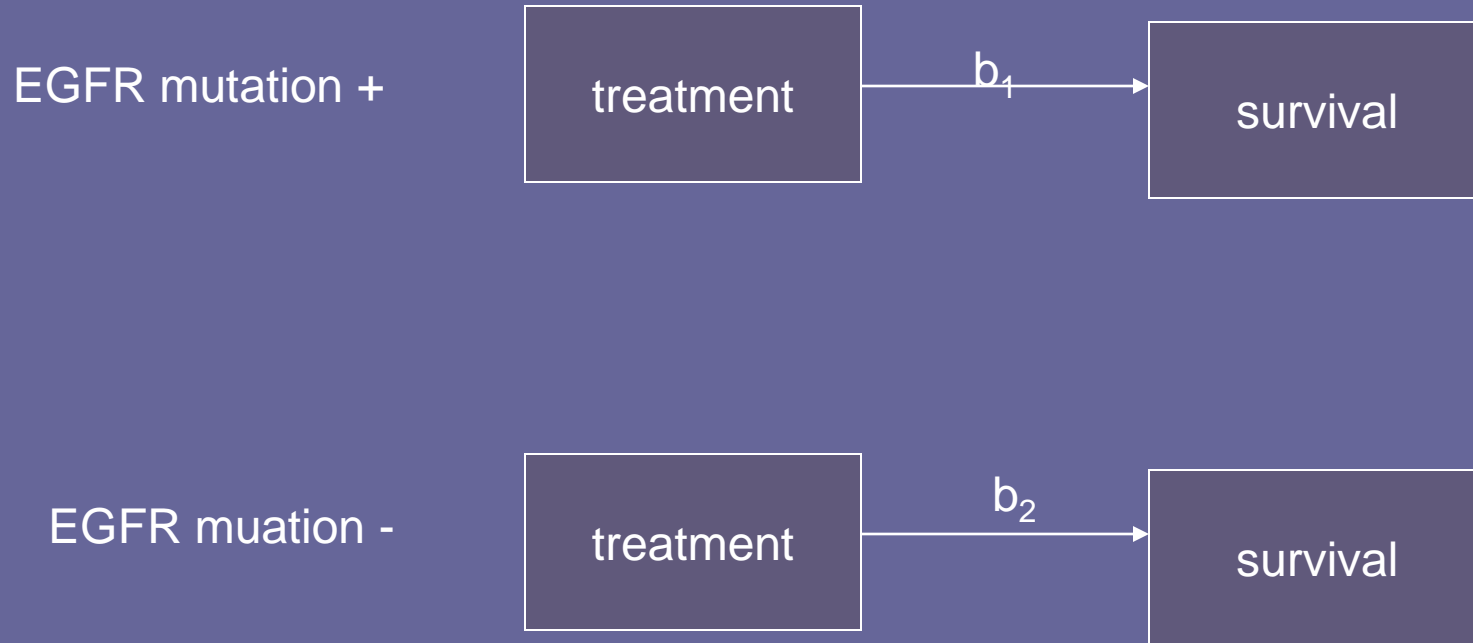
# Prognostic and Predictive Biomarkers

- Prognostic markers: patient or tumor factors that, independent of treatment, predict survival outcome
  - Age
  - Stage
- Predictive markers: factors that may influence and predict the outcome of **treatment** in terms of either response or survival benefit
  - EGFR mutation
  - HER2 status in breast cancer

# Prognostic vs. Predictive



# Another conception of predictive



# Early Stopping Rules and Interim Analyses

- As a general rule, consider incorporating early stopping rules
- Why? Ethics and resources
- Lots of reasons for stopping
- Example: phase II designs
  - Early stopping for safety (one or more arms)
  - Early stopping for futility
  - Early stopping for harm

# Implications of early/interim looks

- Early looks can/will affect
  - Type I error
  - Type II error
- Consequences? They need to be “built” into study design and power calculations
- Misconception: The DSMB will stop the study early if needed for safety or harm so there is no need to account for early looks.
- Having ‘independent’ review does not mean that the interim looks should not be built in to design.

# Approaches for interim analyses

- Different statistical philosophies
- Interim analysis is an area where the philosophies lead to possibly very different approaches
- **Frequentist**: typical p-value approach leads to 'inflated' errors with multiple looks. Can 'spend' type I and II errors during the course of the study.
- **Bayesian**: little or no effect of multiple looks on error rates. Can look essentially 'sequentially'
- Given 'bottleneck' + niche populations, options for stopping trials early is becoming more necessary



# Data Safety Monitoring Board/Committee

- DSMB or DSMC
- Standard in phase III trials.
- Independent body of experts, usually clinical researchers + 1 or more statisticians + an ethicist.
- They periodically review ongoing trial results
- Have access to unblinded treatment assignments if necessary.
- Often assumed they will do more than they should (e.g. redesign the study in midstream)

# Take-home points

- Talk to your statistician early and often when you want to design a study
- Write clear aims and define clear endpoints
- Let the statistician help you with the design, analysis plan and power calculation: that is our job.
- How to learn more?
  - Visit your institution's cancer protocol review committee
  - Try a workshop (check out AACR workshop schedule)

# Some good text books on trials

- General Trials:
  - Clinical Trials: A Methodologic Perspective (Piantadosi)
  - Clinical Trials (Meinert)
- Specific to Cancer:
  - Classic:
    - Clinical Trials in Oncology (Green, Crowley, Benedetti and Smith)
  - Recently published
    - Principles of Anti-Cancer Drug Development (Hidalgo, Eckhardt, Garrett-Mayer, Clendenin)
    - Oncology Clinical Trials: Successful Design, Conduct, and Analysis (Kelly, Halabi, Schilsky)