Statistical Considerations in Protocol Development: From Hypothesis to Analysis

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

Methods in Clinical Cancer Research
Vail, Colorado
August 2009
The process for statistical design and development

- 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
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
  - Simple enough to carry out
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

- Randomized Phase II study evaluating two administration schedules of flavopiridol given in timed sequential combination with cytosine arabinoside (ara-C) and mitoxantrone for adults with newly diagnosed, previously untreated, poor-risk acute myelogenous leukemias (AML)
- Principal Investigator: Judy Karp
- Two different administration schedules:
  - bolus
  - “hybrid bolus-infusion”
1. 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 flavopiridol on cancer.
  - To see if flavopiridol improves cancer outcomes
  - To determine the safety profile of flavopiridol

**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 works or not**
1. Stating research aims

**Better examples:**

- To evaluate the **efficacy of flavopiridol** administered by two different schedules followed by ara-C and mitoxantrone in adults with newly diagnosed AML with poor-risk features
- To evaluate the **toxicities** of flavopiridol administered by two different schedules followed by ara-C and mitoxantrone in adults with newly diagnosed AML with poor-risk features

**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.

- Usually:
  - Phase I: primary aim is finding the “recommended” dose
  - Phase II: primary aim is determining if there is sufficient efficacy

- Secondary aims:
  - important, but do not drive the design
  - Examples in Phase I:
    - pharmacokinetics
    - pharmacodynamic (e.g., methylation)
    - response
  - Examples in Phase II:
    - overall survival
    - 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:
- Examples:
  - The complete remission rate of patients in the \textit{bolus infusion} arm will be at least 55%.
  - The complete remission rate of patients in the \textit{hybrid-bolus infusion} arm will be at least 55%.
  - The median disease-free survival time across both arms will be at least 14 months.
Q1: Complete the following with the best answer:
The primary objective of this study is to________ the ______ of paclitaxel as second-line therapy in endometrial adenocarcinoma.

1. study.....effect
2. observe.....consequences
3. evaluate.....toxicity
4. determine.....endpoints
5. show.....financial benefits
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
  - complete remission rate
  - 6 month progression-free survival.

**Synonyms:** outcome, endpoints

**Aim ≠ endpoint**

- What is an endpoint or outcome?
  - patient-level measure of “effect” of interest
  - measured on each patient in the study
  - it is QUANTIFIABLE
II. Determining your outcome measures

Example:

- **Parameter of interest is the complete remission rate**
- **endpoint = complete remission (CR)**
- **objectively defined:** *Complete remission (CR):* Bone marrow showing less than 5% myeloblasts with normal maturation of all cell lines, an ANC of at least 1000/μL and a platelet count of 100,000 μL, absence of blast in peripheral blood, absence of identifiable leukemic cells in the bone marrow, clearance of disease-associated cytogenetic abnormalities, and clearance of any previously existing extramedullary disease. A CR must be confirmed 4 to 6 weeks after the initial documentation. If possible, at least one bone marrow biopsy should be performed to confirm the CR.

- Each patient is determined to either have had or not had a CR
- **BINARY endpoint in this example**
II. Determining your outcome measures

- Example:
  - Parameter of interest is change in quality of life
  - SF-36 is a quality of life scale
  - Each patient has SF-36 measured at baseline and 6 months after treatment initiation.
  - The difference in scores is the endpoint.
  - CONTINUOUS outcome in this case

- Example:
  - parameter of interest is median overall survival
  - the time from treatment initiation until death (or last known date alive) is recorded on each patient
  - Patients who are still alive at the time of data analysis are considered “censored”
  - time until death is the endpoint.
  - TIME TO EVENT outcome in this case
The following are also 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
Q2: An investigator wrote a trial with the following primary aim:

To goal of this study is to define the presence or absence of malignancy in mediastinal lymph nodes of patients with known or suspected lung cancer, who undergo endobronchial ultrasound with fine needle aspiration for staging purposes.

This aim:
1. is not really an aim. it describes the outcome which is presence or absence of malignancy in a node.
2. is perfect. nice work.
3. is too long
4. is for a surgery trial which explains why it seems so complicated
5. 1 and 4
Q2: An investigator wrote a trial with the following primary aim:

To goal of this study is to define the presence or absence of malignancy in mediastinal lymph nodes of patients with known or suspected lung cancer, who undergo endobronchial ultrasound with fine needle aspiration for staging purposes.

Revised aim:

To goal of this study is to determine if there are factors that can predict the presence of malignancy in mediastinal lymph nodes of patients with known or suspected lung cancer, who undergo endobronchial ultrasound with fine needle aspiration for staging purposes.
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

- **Phase I:**
  - how many dose levels and why?
  - combination or single agent?
  - one or multiple disease types?
  - is expansion at MTD feasible?

- **Phase II**
  - what is historical control rate?
  - is a “reference arm” needed because the historical control rate is not well-defined (randomized phase II?)?
  - is there more than one schedule being considered? (randomized phase II?)?
  - how well is safety profile defined?
  - safety versus efficacy or both?
AML flavopiridol trial

- “This is a randomized Phase II study to evaluate two different schedules of flavopiridol administration in combination with ara-C and Mitoxantrone for response and toxicities. The primary outcome is complete remission.”

- The goals:
  - identify if either schedule is sufficiently efficacious
  - if both efficacious, to choose the better schedule.

- Study design:
  - Simon’s two-stage designs will be used in each arm which will allow an arm to stop early if there is strong early evidence of futility (i.e., lack of efficacy). If both arms proceed through to the second stage and reject the null hypothesis, the schedule with the higher response rate will be selected for further study.
  - A “pick the winner” approach will be used which has a 90% probability of selecting the best schedule if the true difference in CR rates is at least 15%.
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 evaluate the efficacy of two schedules of flavopiridol administration**

Recall primary endpoint: **complete remission**
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 CR rate in each arm will be estimated with its confidence interval.”
  - this provides us with an estimate of the CR rate
  - it also provides us with a measure of precision about the estimate
Recall 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 CR lies in this interval”
- Example: below shows examples where observed CR rate is 0.40. 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 complete remission rate of patients in the bolus infusion arm will be at least 55%
  - The complete remission rate of patients in the hybrid-bolus infusion arm will be at least 55%

- What is a sufficiently LOW CR rate that we are not interested in further pursuit?

- Based on Dr. Karp’s experience, a CR rate of 30% is too low to warrant further study.
IV. Analytic Plan

- In each arm, we perform a hypothesis test:
  - Ho: \( p = 0.30 \) (null)
  - Ha: \( p = 0.55 \) (alternative)

- This test is performed using an exact binomial procedure or a chi-square test.

- The result is a p-value that provides “evidence” to either reject or fail to reject the null hypothesis.

- In our a randomized phase II example:
  - 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 CR rate, a larger sample size will lead to a smaller p-value.
- Example: With an observed CR rate of 0.40, 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 CR rate (e.g., 0.30)
- 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 two groups
  - more common to see overall survival as the outcome of interest. (time to event methods are required)
Q3: It is important to have an analytic plan written for each aim of your protocol because

1. it shows that you have considered how you are going to address each aim of the study
2. otherwise the protocol will not pass scientific peer-review at many institutions
3. it provides job security for the statisticians at your institution
4. 1 and 2

It is easy to write aims. it is more challenging to show that you can actually achieve your aims!
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

<table>
<thead>
<tr>
<th>Ho is True</th>
<th>Ho is NOT True</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="happy.png" alt="Smiley" /></td>
<td><img src="sad.png" alt="Sad" /></td>
<td></td>
</tr>
<tr>
<td>Accept Ho</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="sad.png" alt="Sad" /></td>
<td><img src="happy.png" alt="Smiley" /></td>
<td>Type I error</td>
</tr>
<tr>
<td>Reject Ho</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \alpha = \text{probability of Type I error (level of significance)} \]
\[ \beta = \text{probability of Type II error} \]
\[ 1 - \beta = \text{Power} \]
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: Ho: $p=0.30$ vs. Ha: $p=0.50$
  - this is the assumed “clinically meaningful” difference
  - we chose power of 0.90 ($\beta = 0.10$)
  - alpha was chosen to be 0.10

- Other design issues
  - we wanted to allow early stopping in each arm
  - chose Simon two-stage approach
“Plug and Chug”

- With alpha = beta = 10% and Ho and Ha, a Simon two-stage design is selected.

The Simon’s two stage design we will use is defined as follows. Our null hypothesis is that the response rate is 30% and our alternative hypothesis is that the response rate is 55%. At the first stage we will enroll 15 patients in each arm. We will close accrual to an arm if < 4 responses are seen in that arm in the first stage. If 5 or more CRs are observed, then the arm (s) will remain open for an additional 20 patients per arm. An arm will be considered promising if the CR rate is >42% (i.e., at least 15 responses in 35 patients). This study has power of 90% and a one-sided alpha of 10%.

- The sample size per arm will be 15 patients or 35 patients (depending on early stopping)
- Total study size will be
  - N = 30 if both arms stop early
  - N = 50 if one arm stops early and one continues
  - N = 70 if both arms continue to the 2nd stage
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 45 patients, we will be able to estimate the CR rate with a 95% confidence interval with half-width no greater than 0.15.
- Difficult part: is 0.15 half-width sufficiently precise? how to rationalize that?
Q4: 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.
- Stay tuned for biomarkers talks
- 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?