Ethical Issues in the Statistical Analysis of Clinical Research Data

Roger J. Lewis, MD, PhD
Department of Emergency Medicine
Harbor-UCLA Medical Center
Torrance, California

Introduction

- Types of data-centered scientific misconduct:
  - Data fabrication;
  - Data falsification; and
  - Data stealing.
- Not to be covered.
- Will focus on more subtle aspects of ethical statistical practice.

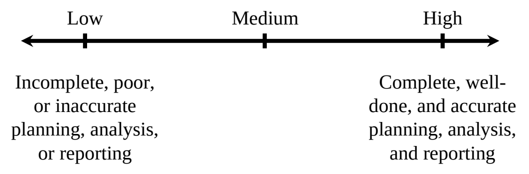
Introduction

- Statistical analysis is a set of quantitative methods for deriving meaningful information from imperfect and incomplete data.
- There is an ethical imperative to perform such analysis in a manner that maximizes the chance the the conclusions drawn are valid and truthful.
- Even without knowing the truth with certainty, we can identify statistical practices that are more or less likely to yield truthful conclusions.

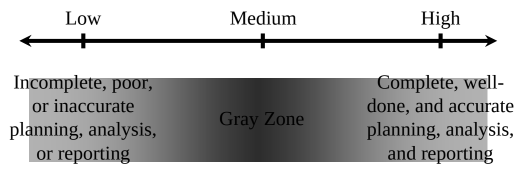
Introduction

- Basic ethical dilemmas:
  - The analysis most likely to yield positive, exciting, and publishable results is often not the analysis most likely to yield the truth.
  - The most complete reporting of analyses may not be the most convincing.
- “Statistical quality” is linked to validity and truthfulness.
- Quality statistical analysis is ethical statistical analysis.

Statistical Quality or Ethical Degree


Low
Medium
High
Incomplete, poor, or inaccurate planning, analysis, or reporting
Complete, well-done, and accurate planning, analysis, and reporting

Statistical Quality or Ethical Degree


Low
Medium
High
Incomplete, poor, or inaccurate planning, analysis, or reporting
Gray Zone
Complete, well-done, and accurate planning, analysis, and reporting

Basic Planning

- Population
  - Inclusion/exclusion
  - Intention-to-treat principle
  - Consecutive sample
  - Sample size
  - Subgroups
- Interventions
- Predictors
- Outcome(s)
  - Primary
  - Secondary
- Blinding

Planned Statistical Analyses

- Primary comparison(s)
- Multiple comparisons
- Missing data/censoring
- Multivariate modeling/stratification
- Options and exploratory analysis, if any
- Interim analyses
- Verification of assumptions

Reporting

- All patients enrolled.
- All analyses performed that might affect interpretation of results.
- Missing/censored information (all denominators).
- Any deviations/additions to original statistical analysis plan.
- Anything you would like to know if you were a reader, reviewer, or editor!

Example 1

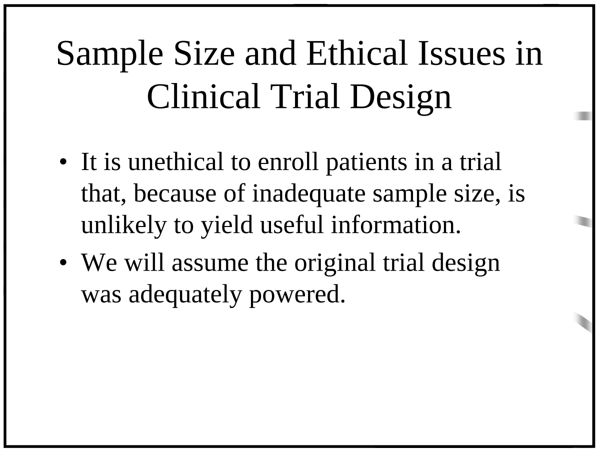
- Prospective interventional clinical trial.
- Planned sample size of 240 patients.
- Abstract deadline approaching and current enrollment is 180 patients.
- Issues:
  - Power;
  - Stopping the trial early;
  - Interim analysis/actual intent; and
  - Reporting.

Type II Error

- Concluding that a difference does not exist, when a difference equal to that sought by the clinical trial does exist--a false negative.
- Occurs when a non-significant *p* value is obtained ($p > \alpha$), yet the two groups are different.
- The risk of a type II error, assuming there is a difference, is β.

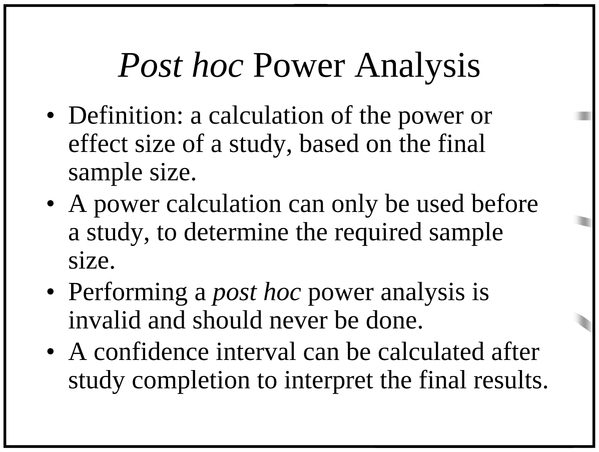
Power

- The chance of obtaining a statistically significant *p* value, if a true difference exists that is equal to that sought by the clinical trial.
- Power = $1 - \beta$.
- Once the study design and analysis method is defined, the power is determined by the sample size, the effect size, and by α.
- Stopping early will result in a lower power.



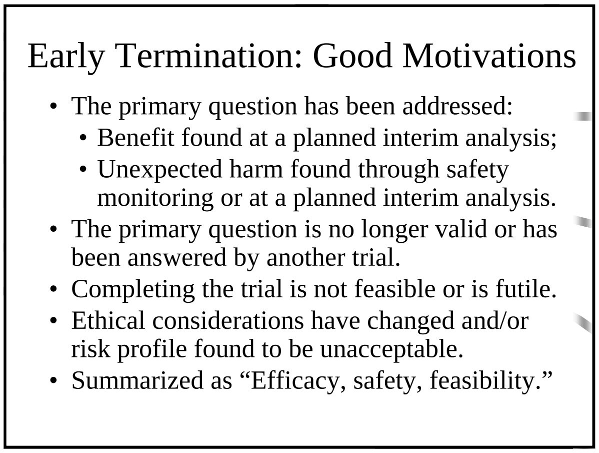
Sample Size and Ethical Issues in Clinical Trial Design

- It is unethical to enroll patients in a trial that, because of inadequate sample size, is unlikely to yield useful information.
- We will assume the original trial design was adequately powered.



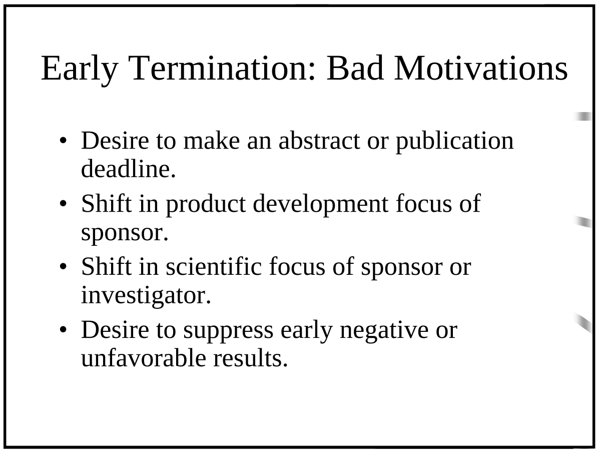
Post hoc Power Analysis

- Definition: a calculation of the power or effect size of a study, based on the final sample size.
- A power calculation can only be used before a study, to determine the required sample size.
- Performing a *post hoc* power analysis is invalid and should never be done.
- A confidence interval can be calculated after study completion to interpret the final results.



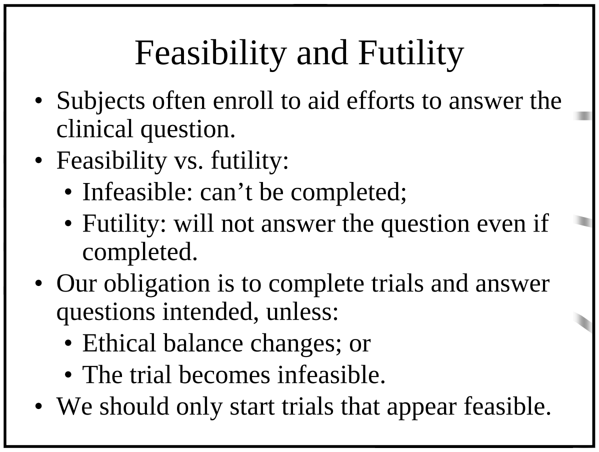
Early Termination: Good Motivations

- The primary question has been addressed:
  - Benefit found at a planned interim analysis;
  - Unexpected harm found through safety monitoring or at a planned interim analysis.
- The primary question is no longer valid or has been answered by another trial.
- Completing the trial is not feasible or is futile.
- Ethical considerations have changed and/or risk profile found to be unacceptable.
- Summarized as “Efficacy, safety, feasibility.”



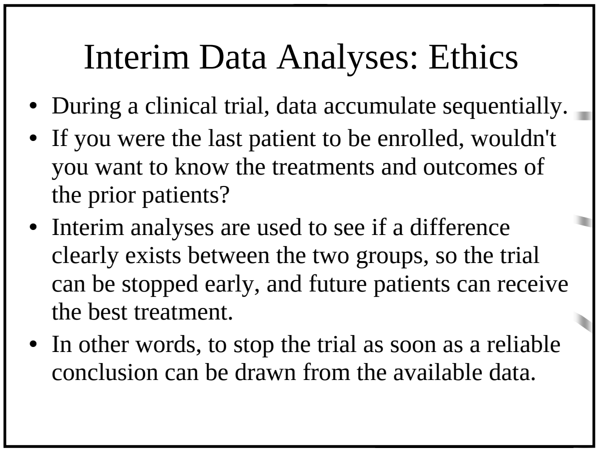
Early Termination: Bad Motivations

- Desire to make an abstract or publication deadline.
- Shift in product development focus of sponsor.
- Shift in scientific focus of sponsor or investigator.
- Desire to suppress early negative or unfavorable results.



Feasibility and Futility

- Subjects often enroll to aid efforts to answer the clinical question.
- Feasibility vs. futility:
  - Infeasible: can’t be completed;
  - Futility: will not answer the question even if completed.
- Our obligation is to complete trials and answer questions intended, unless:
  - Ethical balance changes; or
  - The trial becomes infeasible.
- We should only start trials that appear feasible.



Interim Data Analyses: Ethics

- During a clinical trial, data accumulate sequentially.
- If you were the last patient to be enrolled, wouldn't you want to know the treatments and outcomes of the prior patients?
- Interim analyses are used to see if a difference clearly exists between the two groups, so the trial can be stopped early, and future patients can receive the best treatment.
- In other words, to stop the trial as soon as a reliable conclusion can be drawn from the available data.

Type I Error

- Concluding that a difference exists when it does not.
- A false positive.
- Occurs when a statistically significant *p* value ($p < \alpha$) is obtained when the two groups are not different.
- The risk of a type I error, assuming there is no underlying difference, is α.

Multiple Comparisons

- When two identical groups of patients are compared, there is a chance (α) that a statistically significant *p* value will be obtained (type I error).
- When multiple comparisons are performed, the risk of one or more false-positive *p* values is increased.
- Multiple comparisons include the comparison of two groups at multiple time points.

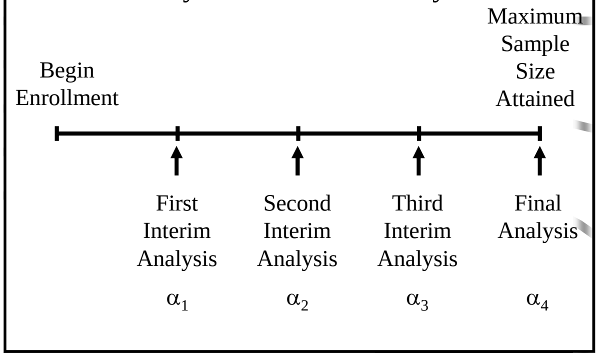
Controlling Type I Error Risk

- Interim data analyses are generally conducted at a few predetermined points throughout the trial to determine if the trial should be stopped because of demonstrated benefit or harm.
- The nominal α (maximum significant *p* value) at each analysis are reduced so that the overall risk of a false positive result, if no treatment effect, is 0.05.

Controlling Type I Error Risk

- Interim analyses must be planned in advance, including the amount of type I error risk to be taken at each analysis.

Group Sequential Trial with Three Interim Analysis and a Final Analysis



Begin Enrollment

Maximum Sample Size Attained

First Interim Analysis

Second Interim Analysis

Third Interim Analysis

Final Analysis

α_1

α_2

α_3

α_4

Nominal α Levels

- α values (the maximum significant *p* value) for each interim analysis are adjusted downward, so that the true type I error rate for the entire study is 0.05.

Max No.		Pocock	O'Brien-Fleming	
Groups	Analysis	α_i	α_i	
2	Interim 1	.0294	.0052	
	Final	.0294	.0480	
3	Interim 1	.0221	.0005	
	Interim 2	.0221	.0141	} > 0.05
	Final	.0221	.0451	
4	Interim 1	.0182	5E-5	
	Interim 2	.0182	.0042	
	Interim 3	.0182	.0194	
	Final	.0182	.0430	

Question

- What would the *p* value obtained after 180 patients be compared to?
  - 0.05
  - 0.0052
  - Whatever the investigator needs to use to claim a statistically significant result

Example 2

- Prospective, randomized clinical trial with a generally negative result.
- Some enrolled patients are found to never have taken their prescribed study medication.
- After completion, some patients are also found to have met prospectively-defined exclusion criteria.
- Issues:
  - Purpose of clinical research;
  - Inclusion/exclusion and ITT principle; and
  - Reporting.

Purpose of a Clinical Trial

- To generate data that will influence physician selection of effective therapy and improve patient outcome.

Determinants of Efficacy

- The effectiveness or efficacy of a therapy is determined by:
  - One's ability to administer the therapy to the patient or to get the patient to take the medication (i.e., "compliance");
  - Inherent or "chemical efficacy;"
  - Patient characteristics captured in inclusion and exclusion criteria; and
  - Other patient characteristics that you may not be able to anticipate, measure, or control

Compliance, Prognosis, and Bias

- Compliant and noncompliant patients often differ in many characteristics, including prognosis.
- Even in a randomized, double-blind study compliance is rarely equal in the different treatment groups.
- This can potentially introduce bias, in that the non-compliant, poor-prognosis (or good-prognosis) subgroup will tend to leave one treatment more than the other.

Intention-to-Treat Analysis: Definition

- Patients are considered to be members of the treatment group to which they are originally assigned, regardless of whether or not they receive that therapy.
- In other words, patients are members of treatment groups according to the treatment the were intended to receive.

Analysis by Treatment Received: Definition

- Patients are considered to be members of treatment groups based on what treatment they actually received.
- Thus a patient assigned to an active drug treatment, who freely admits to never taking any tablets, would be considered a member of the control group.

Don’t do this!

Intention-to-Treat Analysis: Motivation

- The effectiveness of a therapy in clinical practice is determined both by “biologic” activity and by patient compliance.
- To estimate the effectiveness of a treatment in clinical practice, one must allow for differences in compliance.
- This is the purpose of the intention-to-treat principle.

Protocol Violations

- What about the patients that met exclusion criteria and shouldn’t have been enrolled?
  - In general, the primary analysis plan should anticipate such patients and define whether they are included.
  - A secondary analysis should be conducted the “other” way (included or excluded) to ensure no qualitative difference.
  - When in doubt, include all patients in the primary analysis.

Example 3

- Prospective clinical trial with generally negative result.
- “Exploratory” analyses show statistically significant difference in a clinically-important subgroup.
- Issues:
  - Avoiding this scenario
  - Data torturing/data dredging
  - Reporting
  - Interpretation

Subgroups

- Subgroups that are likely to be clinically important, or respond to treatment differently, should be defined prospectively in the protocol.
- Such subgroups should almost always be defined in terms of characteristics that are available at presentation.

Data Dredging or Torturing

- Definition: making a large number of overt or covert statistical comparisons, without the guidance of previously defined, enumerated, and specific hypotheses.
- "Seeing what looks promising."
- "Exploratory data analysis."
- Common before an abstract deadline?
- *Post hoc* subgroups usually equals data dredging.

Multiple Comparisons: Risk of ≥ 1 False Positive

Number of Comparisons	Probability of at Least One Type I Error
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
10	0.40
20	0.64
30	0.79

Assumes α= 0.05, uncorrelated comparisons

Multiple Comparisons: Bonferroni Correction

- A method for reducing the overall risk of a type I error when making multiple comparisons.
- The overall (study-wise) type I error risk desired (e.g., 0.05) is divided by the number of tests, and this new value is used as the α for each individual test.
- Controls the type I error risk, but reduces the power (increased type II error risk).

Subgroup Analysis: Motivation

- Patient populations are heterogeneous, composed of subgroups.
- This is especially true for populations of emergency department patients.
- A treatment effect detected in the entire population may or may not exist for a particular subgroup.
- Data from subgroups are often clinically important and analyzed separately.

Subgroup Analysis: Problems

- Analysis of multiple subgroups requires the use of multiple comparisons, increasing the overall risk of a type I error.
- Since each subgroup is smaller than the whole study population, the power of subgroup comparisons is smaller, increasing the risk of type II error.
- These problems occur even if the subgroups were defined prior to data collection.

Subgroup Interpretation

- Interpretation of results from subgroups should always be conservative and tentative, especially if the effect was unexpected or differs in direction between subgroups.

Example 4

- Observational study of outcome in patients with hypotension after penetrating trauma treated with new resuscitation algorithm.
- Many prehospital and ED clinical characteristics and other treatments collected as potential predictors of outcome.
- Univariate analysis shows strong association between resuscitation algorithm (and many other predictors) and outcome.

Example 4 (Continued)

- Planned multivariate analysis shows marginal association between algorithm and outcome.
- *Post hoc* multivariate analysis, with additional predictors, shows no association.
- Issues:
  - What is the primary result?
  - What has to be reported?
  - What should be reported?
  - Interpretation

Adjusting for Covariates

- The primary result is the result of the analysis defined in the protocol as the primary endpoint analysis.
- No clear rule on what analyses, other than the planned ones, must be reported.
- All probably should be reported.
- Ask yourself "what would you want to know?"
- Interpretation is a matter of judgment.

Example 4 (Continued)

- Five cases of ARDS requiring mechanical ventilation were observed in patients treated using the algorithm, while none were seen in the control patients.
- ARDS was not a planned outcome to be measured or reported.
- Issues:
  - What should be reported?
  - Interpretation

Goals of Safety Monitoring

- Detection of intervention-associated AEs against background rates in the population.
- Identification of unanticipated intervention-associated AEs.
- Identification of subgroups at increased risk of AEs.
- Verification that expected AEs are not occurring more often than expected.

Adverse Events

- Adverse event (AE): "Any untoward medical occurrence in a … subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment." (ICH Guideline E2A)
- An adverse drug reaction (ADR) is an AE that occurs after the patient is given a drug (ICH E2A and 21 CFR§312.32).
- An adverse device effect is an AE in a device trial (21 CFR§812.3).

Serious Adverse Event (SAE)

- An AE that results in any of the following:
  - Death;
  - A life-threatening adverse drug reaction;
  - Inpatient hospitalization or prolongation of existing hospitalization;
  - Persistent or significant disability/incapacity;
  - Congenital anomaly/birth defect.

  (ICH E2A and 21 CFR§312.32)
- Other “important medical events” may also be SAEs, based on medical judgment.

Reporting

- SAEs generally require expedited reporting to the IRB, sponsor, and/or monitor.
- AEs are generally tabulated and reported to the IRB or other bodies at the time of interim analysis and/or at the end of the study.
- Why isn’t this information routinely put in manuscripts?