

# AVOIDING BIAS AND RANDOM ERROR IN DATA ANALYSIS

Susan S. Ellenberg, Ph.D.  
Perelman School of Medicine  
University of Pennsylvania

FDA Clinical Investigator Course

Silver Spring, MD

November 14, 2018

# OVERVIEW

- Bias and random error present obstacles to obtaining accurate information from clinical trials
- Bias and error can result at all stages of trial
  - Design
  - Conduct
  - Analysis and interpretation

# POTENTIAL FOR BIAS AND ERROR



- Bias
  - Missing data
    - Dropouts
    - Deliberate exclusions from analysis
  - Handling noncompliance
- Error
  - Multiple comparisons

# MANY CAUSES OF MISSING DATA



- Subject dropped out and refused further follow-up
- Subject stopped drug or otherwise did not comply with protocol and investigators ceased follow-up
- Subject died or experienced a major medical event that prevented continuation in the study
- Subject did not return--lost to follow-up
- Subject missed a visit
- Subject refused a procedure
- Data not recorded

# MANY CAUSES OF MISSING DATA



- Subject dropped out and refused further follow-up
- Subject stopped drug or otherwise did not comply with protocol and investigators ceased follow-up
- Subject died or experienced a major medical event that prevented continuation in the study
- Subject did not return--lost to follow-up
- Subject missed a visit
- Subject refused a procedure
- Data not recorded

# THE BIG WORRY ABOUT MISSING DATA

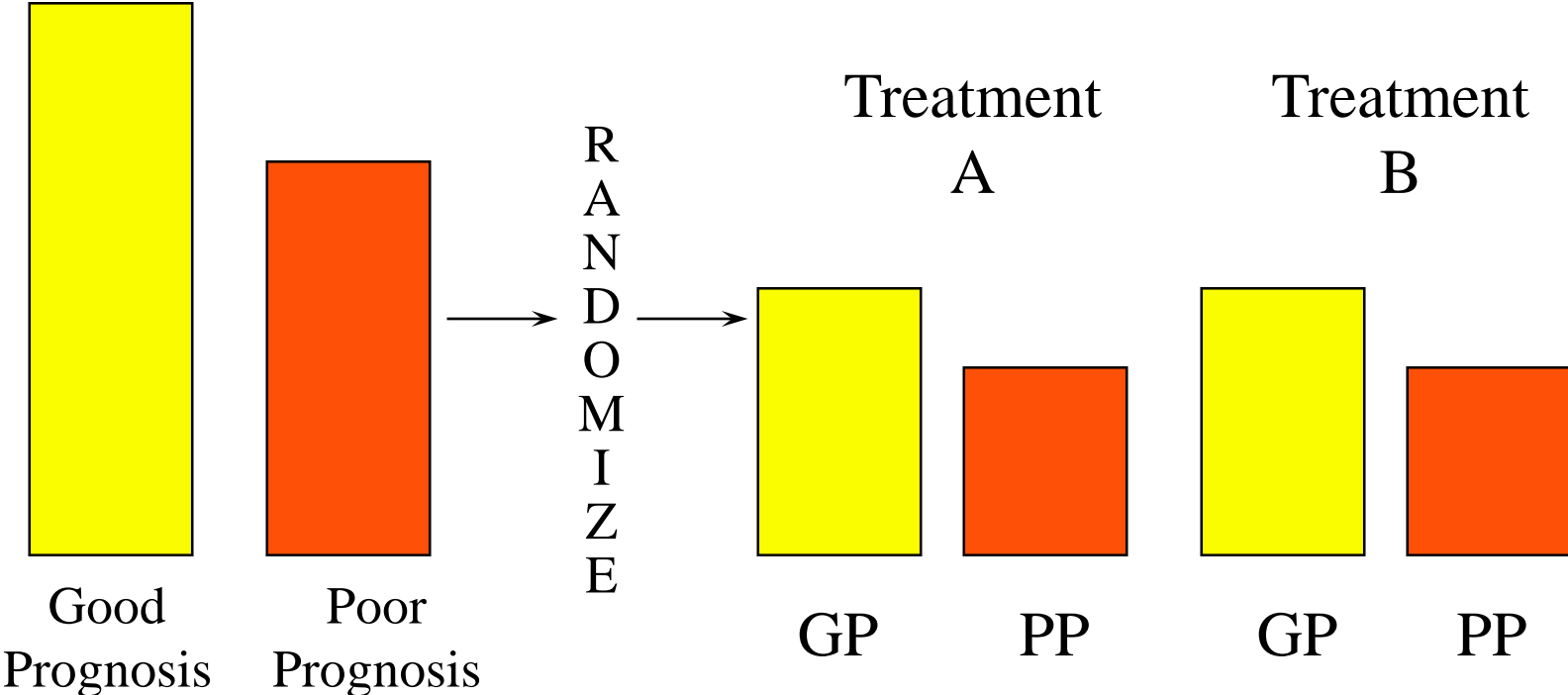


- Missing-ness may be associated with outcome
- We don't know the form of this association
- Nevertheless, if we fail to account for the (true) association, we may **bias** our results

# MISSINGNESS ASSOCIATED WITH OUTCOME: EXAMPLE

- Suppose drug X has certain side effects
- Suppose side effects tend to be worse in subjects who are sicker to begin with
- Then sicker patients assigned to the drug drop out at higher rates than less sick patients
- If we analyze only non-dropouts, results would be biased
  - non-dropouts assigned to drug X would be less sick than the non-dropouts on placebo

# EFFECT OF RANDOMIZATION





# EXAMPLE: CANCER STUDY

- Test of post-surgery chemotherapy
- Subjects randomized to observation only or chemo, following potentially curative surgery
- Protocol specified treatment had to begin no later than 6 weeks post-surgery
  - Assumption: if you don't start the chemo soon enough you won't be able to control growth of micro-metastases that might still be there after surgery
  - Concern that including subjects starting chemo too late would dilute estimate of treatment effect

# EXAMPLE: CANCER STUDY

- Subjects assigned to chemotherapy whose treatment was delayed beyond 6 weeks were dropped from study with no further follow-up—no survival data for these subjects
- No such dropouts on observation arm
  - Whatever their status at 6 weeks, they were included in follow-up and analysis
- Are the subjects whose treatment was delayed different in any way from those who were able to start treatment on time?

# EXAMPLE: CANCER STUDY

- Problem: those with delays might have had poorer prognosis
  - Larger tumors, requiring more complex surgeries
  - Longer time to recover from surgery could reflect co-morbidities or older age
- Concern to avoid dilution of results opened the door to potential bias
- May have reduced false negative error
- Almost surely increased false positive error
- Cannot assess this without follow-up data on those with delayed treatment

# MISSING DATA AND PROGNOSIS FOR STUDY OUTCOME



- For most missing data, very plausible that missingness is related to prognosis
  - Subject feels worse, doesn't feel up to coming in for tests
  - Subject feels much better, no longer interested in study
  - Subject feels study treatment not helping, drops out
  - Subject intolerant to side effects
- Thus, missing data raise concerns about biased results
- Can't be sure of the direction of the bias; can't be sure there IS bias; can't rule it out

# DILEMMA

- Excluding patients with missing values can bias results, increase Type I error (false positives)
- Collecting and analyzing outcome data on non-adherent patients may dilute results, increase Type II error (false negatives), as in cancer example
- General principle: we can compensate for dilution with sample size increases, but can't compensate for potential bias of unknown magnitude or direction

# INTENT-TO-TREAT (ITT) PRINCIPLE



All randomized patients should be included in the (primary) analysis, in their assigned treatment groups

# MISSING DATA AND THE INTENTION -TO-TREAT (ITT) PRINCIPLE

- ITT: Analyze all randomized patients in groups to which randomized
- What to do when data are unavailable?
- Implication of ITT principle for design: collect all required data on all patients, regardless of compliance with treatment

Thus, avoid missing data

# WHAT ABOUT PEOPLE ASSIGNED TO PLACEBO WHO CROSS OVER TO ACTIVE ARM OR START OTHER AVAILABLE THERAPY?

- This will dilute observable difference
- Can be planned for by enlarging sample size
- If large numbers are expected to cross over, can consider crossover as a failure of therapy



# ROUTINELY PROPOSED MODIFICATIONS OF ITT

- All randomized **eligible** patients...
- All randomized **eligible** patients **who received any of their assigned treatment...**
- All randomized patients **for whom the primary outcome is known...**

# MODIFICATION OF ITT (1)

- OK to exclude randomized subjects who turn out to be ineligible?

# MODIFICATION OF ITT (1)

- OK to exclude randomized subjects who turn out to be ineligible?
  - probably won't bias results--unless level of scrutiny depends on treatment and/or outcome
  - greater chance of bias if eligibility assessed after data are in and study is unblinded

# 1980 ANTURANE REINFARCTION TRIAL: Mortality Results



---

	Anturane	Placebo	P-Value
Randomized	74/813 (9.1%)	89/816 (10.9%)	0.20
“Eligible”	64/775 (8.3%)	85/783 (10.9%)	0.07
“Ineligible”	10/38 (26.3%)	4/33 (12.1%)	0.12
P-Values for eligible vs. ineligible	0.0001	0.92	

---

Reference: Temple & Pledger (1980) *NEJM*, p. 1488

(slide courtesy of Dave DeMets)

# MODIFICATION OF ITT (2)

- OK to exclude randomized subjects who never started assigned treatment?

# MODIFICATION OF ITT (2)

- OK to exclude randomized subjects who never started assigned treatment?
  - probably won't bias results if study is double-blind
  - in unblinded studies (e.g., surgery vs drug), refusals of assigned treatment may result in bias if “refusers” excluded

# MODIFICATION OF ITT (3)

- OK to exclude randomized patients who become lost-to-followup so outcome is unknown?

# MODIFICATION OF ITT (3)

- OK to exclude randomized patients who become lost-to-follow-up so outcome is unknown?
  - Possibility of bias, but can't analyze what we don't have
  - Model-based analyses may be informative if assumptions are reasonable
  - Sensitivity analysis important, since we can never verify assumptions



# NONCOMPLIANCE: A PERENNIAL PROBLEM



- There will always be people who don't use medical treatment as prescribed
- There will always be noncompliant subjects in clinical trials
- How do we evaluate data from a trial in which some subjects do not adhere to their assigned treatment regimen?

# CLASSIC EXAMPLE

- Coronary Drug Project
  - Large RCT conducted by NIH in 1970's
  - Compared several treatments to placebo
  - Goal: improve survival in patients at high risk of death from heart disease
- Results disappointing
- Investigators recognized that many subjects did not fully comply with treatment protocol

# CORONARY DRUG PROJECT

Five-year mortality by treatment group



Treatment	N	mortality
clofibrate	1065	18.2
placebo	2695	19.4

Coronary Drug Project Research Group, *JAMA*, 1975

# CORONARY DRUG PROJECT



Five-year mortality by treatment group

Adherence	N	% mortality
< 80%	357	24.6
≥80%	708	15.0

Coronary Drug Project Research Group, *NEJM*, 1980

# CORONARY DRUG PROJECT



Five-year mortality by adherence to clofibrate and placebo

Adherence	<u>Clofibrate</u>		<u>Placebo</u>	
	N	% mortality	N	% mortality
<80%	357	24.6	882	28.2
<u>&gt;</u> 80%	708	15.0	1813	15.1

Coronary Drug Project Research Group, NEJM, 1980

# HOW TO EXPLAIN?

- Taking more placebo can't possibly be helpful
- Must be explainable on basis of imbalance in prognostic factors between those who did and did not comply
- Adjustment for 20 strongest prognostic factors reduced level of significance from  $10^{-16}$  to  $10^{-9}$
- Conclusion: important but unmeasured prognostic factors are associated with compliance

# ANALYSIS WITH MISSING DATA

- Analysis of data when some are missing requires assumptions
- The assumptions are not always obvious
- When a substantial proportion of data is missing, different analyses may lead to different conclusions
  - reliability of findings will be questionable
- When few data are missing, approach to analysis probably won't matter

# COMMON APPROACHES

- Ignore those with missing data; analyze only those who completed study
- For those who drop out, analyze as though their last observation was their final observation
- For those who drop out, predict what their final observation would be on the basis of outcomes for others with similar characteristics



# ASSUMPTIONS FOR COMMON ANALYTICAL APPROACHES

- Analyze only subjects with complete data
  - Assumption: those who dropped out would have shown the same effects as those who remained in study
- Last observation carried forward
  - Assumption: those who dropped out would not have improved or worsened
- Modeling approaches
  - Assumption: available data will permit unbiased estimation of missing outcome data

# ASSUMPTIONS ARE UNVERIFIABLE (AND PROBABLY WRONG)



- Completers analysis
  - Those who drop out are almost surely different from those who remain in study
- Last observation carried forward
  - Dropout may be due to perception of getting worse, or better—interim outcome may not indicate true effect of treatment
- Modeling approaches
  - Can only predict using data that are measured; unmeasured variables may be more important in predicting outcome (CDP example)

# SENSITIVITY ANALYSIS

- Analyze data under variety of different assumptions—see how much the inference changes
- Such analyses are essential to understanding the potential impact of the assumptions required by the selected analysis
- If all analyses lead to same conclusion, will be more comfortable that results are not biased in important ways
- Important to pre-specify sensitivity analyses and consider what outcomes might either confirm or cast doubt on results of primary analysis

# “WORST CASE SCENARIO”

- Simplest type of sensitivity analysis
- Assume all missing outcomes for investigational drug arm were treatment failures, all for control group were successes
- If drug still appears significantly better than control, even under this extreme assumption, home free
- Note: if more than a tiny fraction of data are missing, this approach is unlikely to provide persuasive evidence of benefit

# MANY OTHER APPROACHES

- Different ways to model possible outcomes
- Different assumptions can be made
  - Missing at random (can predict missing data values based on available data)
  - Nonignorable missing (must create model for missing mechanism)
- Simple analyses (eg, completers only) can also be considered sensitivity analyses
- If different analyses all suggest same result, can be more comfortable with conclusions

# CLARIFYING THE “ESTIMAND”

- An “estimand” is what we want to estimate
- What we want to estimate is not always what we can estimate
  - Treatment effect in all randomized subjects?
  - Treatment effect in all randomized subjects who adhered to assigned regimen?
  - Treatment effect in all randomized subjects with follow-up data?
- New draft guidance (Estimands and Sensitivity Analysis in Clinical Trials: Addendum to ICH E9, Statistical Principles for Clinical Trials) discusses trial design and analytical approaches specific to the estimand chosen

# THE PROBLEM OF MULTIPLICITY



- Multiplicity refers to the multiple judgments and inferences we make from data
  - hypothesis tests
  - confidence intervals
  - graphical analysis
- Multiplicity leads to concern about inflation of Type I error, or false positives

# EXAMPLE

- The chance of drawing the ace of clubs by randomly selecting a card from a complete deck is  $1/52$
- The chance of drawing the ace of clubs at least once by randomly selecting a card from a complete deck 100 times is....?



# EXAMPLE

- The chance of drawing the ace of clubs by randomly selecting a card from a complete deck is  $1/52$
- The chance of drawing the ace of clubs at least once by randomly selecting a card from a complete deck 100 times is....?
- And suppose we pick a card at random and it happens to be the ace of clubs—what probability statement can we make?

# MULTIPLICITY IN CLINICAL TRIALS



- There are many types of multiplicity to deal with
  - Multiple endpoints\*
  - Multiple subsets
  - Multiple analytical approaches
  - Repeated testing over time

\*FDA Draft Guidance: Multiple Endpoints in Clinical Trials (January 2017)

# MOST LIKELY TO MISLEAD: DATA-DRIVEN TESTING



- Perform experiment
- Review data
- Identify comparisons that look “interesting”
- Perform significance tests for these results

# EXAMPLE: OPPORTUNITIES FOR MULTIPLICITY IN AN ONCOLOGY TRIAL



- Experiment : regimens A, B and C are compared to standard of care
  - Intent: cure/control cancer
  - Eligibility: non-metastatic disease

# OPPORTUNITIES FOR MULTIPLE TESTING

- Multiple experimental treatment arms: A, B, C
- Subsets: gender, age, tumor size, marker levels...
- Site groupings: country, type of clinic...
- Covariates accounted for in analysis
- Repeated testing over time
- Multiple endpoints
  - different outcome: mortality, progression, response
  - different ways of addressing the same outcome: different statistical tests

# RESULTS IN SUBSETS

- Perhaps the most vexing type of multiple comparisons problem
- Very natural to explore data to see whether treatment works better in some types of patients than others
- Two types of problems
  - Subset in which the treatment appears beneficial, even though no overall effect
  - Subset in which the treatment appears ineffective, even though overall effect is positive

# REAL EXAMPLE

- International Study of Infarct Survival (ISIS)-2 (Lancet, 1988)
  - Over 17,000 subjects randomized to evaluate aspirin and streptokinase (A and S) post-MI in a factorial design
    - Subjects randomized to A + S placebo, S + A placebo, A + S (no placebos), both placebos
  - Both treatments showed highly significant survival benefit in full study population
  - Subset of subjects born under Gemini or Libra showed slight adverse effect of aspirin

# REAL EXAMPLE

- New drug for sepsis
  - Overall results negative
  - Significant drug effect in patients with APACHE score in certain region
  - Plausible basis for difference
  - Company planned new study to confirm
  - FDA requested new study not be limited to favorable subgroup
  - Results of second study: no effect in either subgroup



# DIFFICULT SUBSET ISSUES

- In multicenter trials there can be concern that results are driven by a single center
- Not always implausible
  - Different patient mix
  - Better adherence to assigned treatment
  - Greater skill at administering treatment

# A PARTICULAR CONCERN IN MULTIREGIONAL TRIALS



- Very difficult to standardize approaches in multiregional trials
- May not be desirable to standardize too much
  - Want data within each region to be generalizable within that region
- Not implausible that treatment effect will differ by region
- How to interpret when treatment effects do seem to differ?

# REGION AS SUBSET

- A not uncommon but difficult problem
  - Assumption in a multiregional trial is that if treatment is effective, it will be effective in all regions
  - We understand that looking at results in subsets could yield positive results by chance
  - Still, it is natural to be uncomfortable about using a treatment that was effective overall but with zero effect where you live
  - Particularly problematic for regulatory authorities— should treatment be approved in a given region if there was no apparent benefit in that region?

# EXAMPLE: HEART FAILURE TRIAL



- MERIT-HF trial (*Am Heart J*, 2001)
- Total enrollment: 3991
  - US: 1071
  - All others: 2920
- Total deaths observed: 362
  - US: 100
  - All others: 262
- Overall relative risk: 0.67 ( $p < 0.0001$ )
  - US: **1.05** 95% confidence interval (0.73, 1.53)
  - All others: **0.56** 95% confidence interval (0.44, 0.71)

# MULTIREGIONAL TRIALS

- FDA recently issued a guidance document (developed by the International Conference on Harmonization) on planning multiregional clinical trials: ICH E17  
<https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM519603.pdf>
- Emphasizes importance of considering how to evaluate consistency of treatment effects across regions
  - Descriptive and graphical approaches
  - Statistical models that include region as a covariate
  - Testing for treatment by region interaction

# FOUR BASIC APPROACHES TO MULTIPLE COMPARISONS PROBLEMS



1. Ignore the problem; report all interesting results
2. Perform tests of all comparisons of interest at the nominal (typically 0.05) level and warn reader that no accounting has been taken for multiple testing
3. Limit yourself to only one test
4. Adjust the p-values/confidence interval widths in some statistically valid way

# IGNORE THE PROBLEM

- Probably the most common approach
- Less common in the higher-powered journals, or journals where statistical review is standard practice
- Rarely accepted in regulatory submissions
- Even when not completely ignored, often not fully addressed

# DO ONLY ONE TEST

- Single (pre-specified) primary hypothesis
- Single (pre-specified) analysis
- No consideration of data in subsets
  
- Not really practicable
- (Common practice: pre-specify the primary hypothesis and analysis, consider all other analyses “exploratory”)



# NO ACCOUNTING FOR MULTIPLE TESTING, BUT MAKE THIS CLEAR



- Message is that readers should “mentally adjust”
- Justification: allows readers to apply their own preferred multiple testing approach
- Appealing because you show that you recognize the problem, but you don’t have to decide how to deal with it
- May expect too much from statistically unsophisticated audience

# USE SOME TYPE OF ADJUSTMENT PROCEDURE



- Divide desired  $\alpha$  by the number of comparisons (Bonferroni); often used when two primary endpoints are designated
- Bonferroni-type stepwise procedures
  - Somewhat less conservative; might be used to assess a series of secondary endpoints
- Control “false discovery” rate
  - Even less conservative
  - Allows for a small proportion of false positives, rather than ensuring only small probability of each test producing a false positive conclusion

# ONE MORE ISSUE: BEFORE- AFTER COMPARISONS

©2002 by King Features Syndicate, Inc. World rights reserved.

RING!  
RING!!  
RING!!



RUFF!  
RUFF!!  
RUFF!!



I'VE FOUND IF YOU  
BARK LONG ENOUGH  
THAT THING SHUTS  
UP



MORT  
WALKER  
• COM

# BEFORE-AFTER COMPARISONS

- In order to assess effect of new treatment, must have a comparison group
- Changes from baseline could be due to factors other than intervention
  - Natural variation in disease course
  - Patient expectations/psychological effects
  - Regression to the mean
- Cannot assume investigational treatment is cause of observed changes without a control group, with rare exceptions

# SIDEBAR: REGRESSION TO THE MEAN

- “Regression to the mean” is a phenomenon resulting from using threshold levels of variables to determine study eligibility
  - Must have blood pressure  $> X$
  - Must have CD4 count  $< Y$
- In such cases, a second measure will “regress” toward the threshold value
  - Some qualifying based on the first level will be on a “random high” that day; next measure likely to closer to their true level
  - Especially true for those whose true level is close to threshold value

# REAL EXAMPLE

- Ongoing study of testosterone supplementation in men over age of 65 with low T levels and functional complaints
- Entry criterion: 2 screening T levels required; average needed to be  $< 275$  ng/ml, neither could be  $> 300$  ng/ml
- Even so: about 10% of men have baseline T levels  $> 300$



# CONCLUDING REMARKS

- There are many pitfalls in the analysis and interpretation of clinical trial data
- Awareness of these pitfalls will limit errors in drawing conclusions
- For some issues, no consensus on optimal approach
- Statistical rules are best integrated with clinical judgment, and basic common sense

# REFERENCES



- FDA guidance documents
  - E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials. <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM582738.pdf>
  - Multiple Endpoints in Clinical Trials. <https://www.fda.gov/downloads/drugs/guidancecomplianceinformation/guidances/ucm536750.pdf>
  - E17 General Principles for Planning and Design of Multiregional Clinical Trials. <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM519603.pdf>
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, Washington, DC
- Temple R, Pledger G. The FDA's critique of the anturane reinfarction trial. *N Engl J Med* 303:1488-92, 1980.
- Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med* 303:1038-41, 1980.
- Second International Study of Infarct Survival Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 2(8607):349-60.
- Wedel H, DeMets D, Deedwania P, et al. Challenges of subgroup analyses in multinational clinical trials: Experiences from the MERIT-HF trial. *Am Heart J* 142:502-11, 2001
- Snyder PJ, Bhasin S, Cunningham GR et al. Effects of testosterone treatment in older men. *N Engl J Med* 374:611-24, 2016.

