# Statistics and the Medical Literature

Dave Harrington

September 12 & 19, 2016

# General remarks

# Questions received so far

*Ike Swetlitz:*

How important is effect size when determining whether to publish, as opposed to (or in addition to) statistical significance?

- Very, but important effect sizes vary with disease, type of study

What do we look for in the power of a study?

- Generally want it the power (probability of detecting important effect) $\geq 80\%$

Do we have standards for minimizing false positive and false negative results?

- Yes, but they may vary with type of studies.
- Generally false positive rate should be $< 5\%$
- False negative rate should be $< 20\%$

# Questions

*Sharon Begley:*

How do we tell if a study is sufficiently powered to determine important effect sizes?

- Design before study execution, confidence intervals after study is completed
- Is it possible to examine power as a function of effect size?
  - Yes, but should be done before study is launched
- How to interpret dropouts?
  - With difficulty ...

# What does a statistical reviewer look for?

We do not expect designs or analyses to be perfect

We look for

- Accuracy, but this is difficult to check in detail in almost all cases.
- Transparency
- Following principles of scientific investigations: formulating a hypothesis, then testing it
- Use of 'state of the practice' methods
- Sensitivity analyses: do different approaches lead to the same qualitative conclusion

# Statistical review at NEJM

5 statistical reviewers, spanning range of subject areas

We participate in the weekly AE sessions where papers are chosen for a closer look.

We provide a detailed statistical review for any paper that the associate editors wish to evaluate further.

- We guarantee a review in 2 weeks for routine papers
- 6 calendar days for fast track papers.

Approximately 15 - 20% of papers are rejected on statistical grounds

- But our focus is on trying to improve the statistics (if needed) in papers with strong science.

# Overview

Techniques of statistics designed to

- 'Minimize' the chance of a false positive (but it can never be 0)
- 'Maximize the chance of a true positive (but it can never be 1)
- Acknowledge uncertainty in results, conclusions

Techniques and theory developed for settings much simpler than studies with human subjects and with far less 'noise'.

Considerable progress over the last 25 years, but methods for interpreting large, complex studies still less than perfect

- Design sometimes requires compromise
- Interpretation requires judgment

# Interpreting the design

# Main elements of the design – RCTs and observational studies

Plan to minimize bias?

- RCT: Randomization, blinding, identical evaluation/follow-up schedules
- Observational studies: record important confounders, plan for post-hoc adjustment, perhaps include matching in recruitment

Chance of detecting important effects/associations? (Adequately powered, at leat 80%)

Plan to minimize false positive results if many endpoints/subgroups examined?

- More difficult with large observational studies, especially genomics/genetics

# Power and type I error

Type I error (alpha error)

- Probability that trial will report a false positive, i.e., claim a significant result when there is no treatment effect.
- Typically set no larger than 5%
- Depends on method of analysis, does not depend on sample size

Power

- Probability that the trial will report a true positive, i.e., claim a significant result when there is a treatment effect.
- Should be 80% or greater
- Depends on sample size, method of analysis and size of treatment effect.
- Power calculations relevant when study is designed.
- Power calculations have little value after a study is complete.
  - Precision measured through confidence intervals

# Design of SPRINT

From the methods section of the paper:

> *We planned a 2-year recruitment period, with a maximum follow-up of 6 years, and anticipated a loss to follow-up of 2% per year. With an enrollment target of 9250 participants, we estimated that the trial would have 88.7% power to detect a 20% effect with respect to the primary outcome, assuming an event rate of 2.2% in the standard-treatment group.*

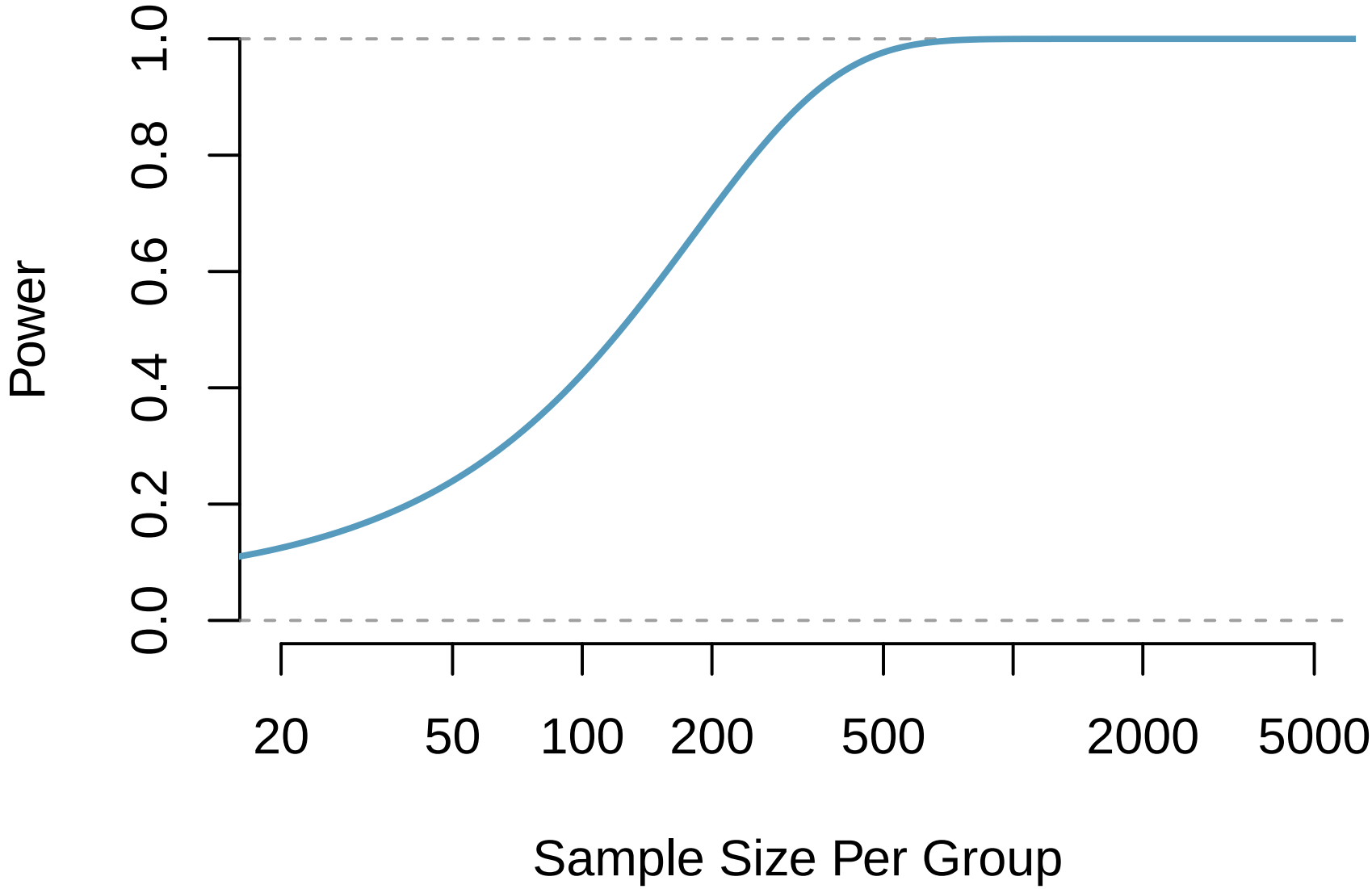
# Calculating Power

Next two slides show graphs of power in hypothetical study of blood pressure lowering med.

Two medications, experimental vs control, using $\alpha = 0.05$

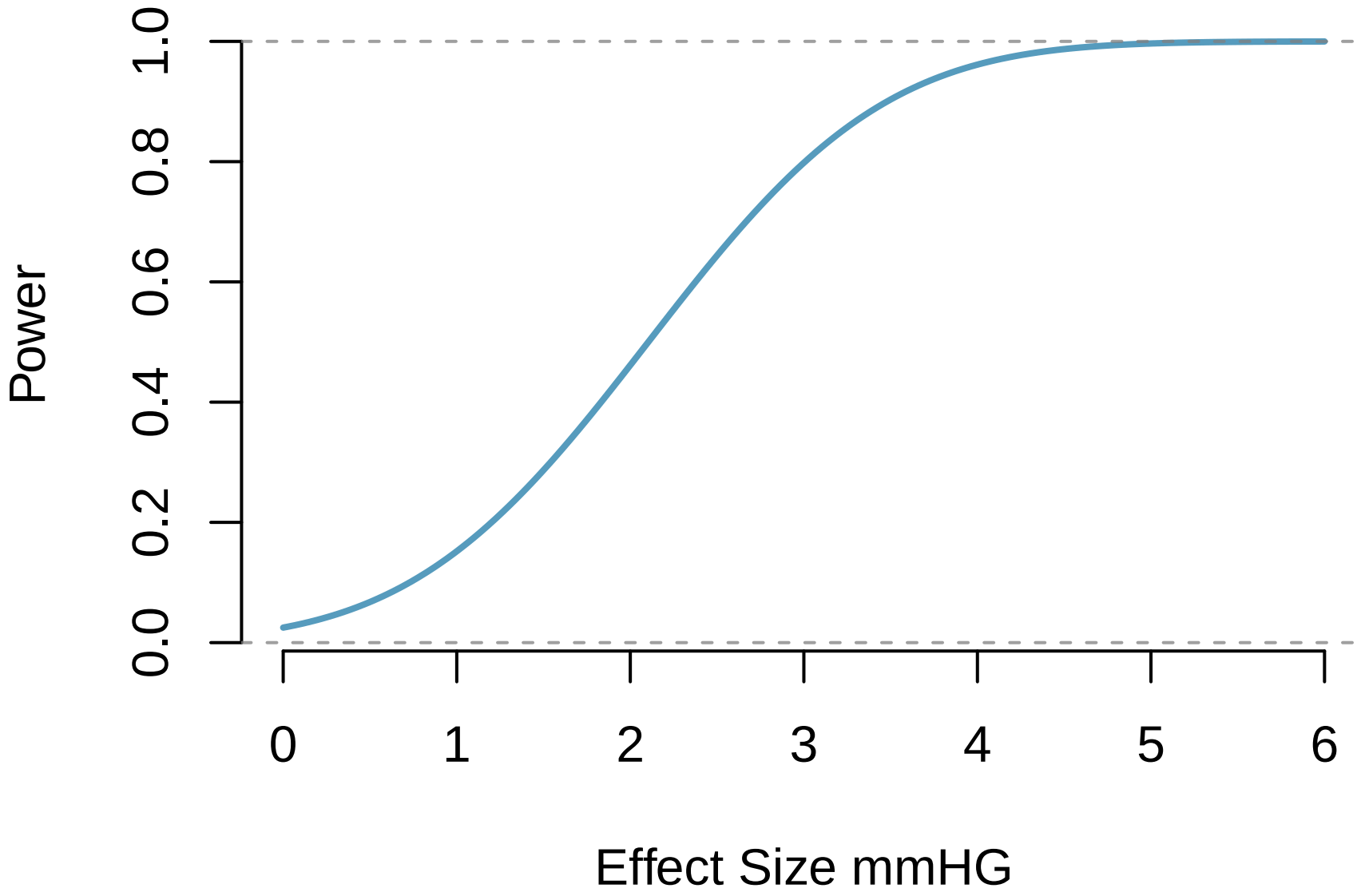
- First, power as a function of sample size if true effect is 3mmHg reduction
- Second, power as a function of reduction in bp (effect size) for sample size 250 per group.

# Power vs. sample size




More than about 250 to 350 per group doesn't provide much additional value.

# Power vs. effect size, 250 participants per group



# Confidence intervals

Confidence intervals are the preferred way to summarize outcome data.

Easiest definition:

- Confidence interval provides a single estimate with a 'margin of error'.
- The size of the margin is determined by the variability in the data and the 'confidence coefficient'
  - Confidence coefficient is an estimate of the likelihood that the interval is correct
  - 95% is typically used for confidence coefficient

# Measuring precision after study completion

From *Postmenopausal estrogen use and progestin use and the risk of cardiovascular disease,* NEJM 15 August 1996

> *We observed a marked decrease in the risk of major coronary heart disease among women who took estrogen with progestin, as compared with the risk among women who did not use hormones (multivariate adjusted relative risk, 0.39; 95 percent confidence interval, 0.19 to 0.78) . . .*

# Measuring precision after study completion

*However, there was no significant association between stroke and use of combined hormones (multivariate adjusted relative risk, 1.09; 95 percent confidence interval, 0.66 to 1.80) . . .*

# Controlling type I error

The more tests one does, the more likely it is that at least one will be a false positive.

Suppose each test is done at level $\alpha = 0.05$.

| Number of Comparisons | Experimentwise Error |
|---|---|
| 1 | 0.05 |
| 2 | 0.10 |
| 3 | 0.14 |
| 5 | 0.23 |
| 10 | 0.40 |
| 100 | $>$0.90 |

# Are fMRI results reliable?

The New York Times | http://nyti.ms/2bYUMQm

SundayReview | NEWS ANALYSIS

## Do You Believe in God, or I Software Glitch?

By KATE MURPHY AUG. 27, 2016

# Controlling type 1 error

Assume target experimentwise error 5% ($\alpha = 0.05$)

Bonferroni approximation, no order specified for comparisons

- Divide significance level by number of planned tests
- 5 comparisons, use p = 0.01
- Not practical when many comparisons planned, especially in genetics studies

# Controlling type 1 error...

Holm's method, no order specified

- Order the p-values from smallest to largest
- Stop testing as soon as a p-value is too large
- 5 comparisons:
  - Compare smallest p-value to $0.05/5 = 0.01$.
  - Compare next smallest to $0.05/4 = 0.0125$.
  - Next smallest to $0.05/3$
  - etc

# Controlling type 1 error...

Holm-Bonferroni: one primary and several secondary endpoints

- Test primary outcome at $\alpha = 0.05$. Stop testing if not significant
- If significant, test secondary outcomes using Holm's method

The most important detail is sometimes hidden

- Was the number of comparisons planned or reported the same as the number done?

# Anticipating many tests: Duty Hours

NEJM 25 Feb 2016

> *Because one midpoint interim analysis was performed for data and safety monitoring purposes, the level of statistical significance for our final analyses of only patient outcomes was adjusted to 0.04 in order to maintain an overall significance level for the entire trial of 0.05.*

# Duty hours . . .

*In the context of a hypothesis of no difference [non-inferiority] in outcomes across study groups, correction for multiple comparisons was not a conservative approach for reducing the false discovery rate; thus, we report non–Bonferroni-corrected P values for all estimates. Bonferroni adjustment of P values for patient outcomes entails lowering the value from 0.04 to 0.004 (adjustment for 11 tests), whereas adjustment of P values for resident outcomes entails lowering the value from 0.05 to 0.0015 (adjustment for 34 tests).*

# Controlling type 1 error in genomic studies

Genomic studies typically look at associations between a characteristic and thousands of genes.

Two common methods:

- Use $\alpha = 0.0000001$, or smaller.
- Set an acceptable threshold for the proportion of discoveries that will be false, called the false discovery rate (FDR).
  - Use sophisticated methods to estimate FDR, then keep it in bounds.
  - Not discussed in these slides

# Special topic: Non-inferiority (NI) randomized trials

The NI design has one explicit and one implicit goal.

- Explicit goal: demonstrate that experimental treatment $T$ is as effective, or nearly as effective, as best available therapy, C.
- Implicit goal: demonstrate that $T$ is better than placebo or no treatment (labeled $P$ for placebo).

Ordinarily, both must be true for $T$ to be a therapeutic option.

# Goals of NI Designs

The ideal study design would be a three-arm design, with $P$, $C$, and $T$.

But a placebo or no treatment arm is usually unethical

The non-inferiority trial provides a direct comparison of $T$ to $C$, but it does not provide a direct comparison of $T$ to $P$.

# NI designs

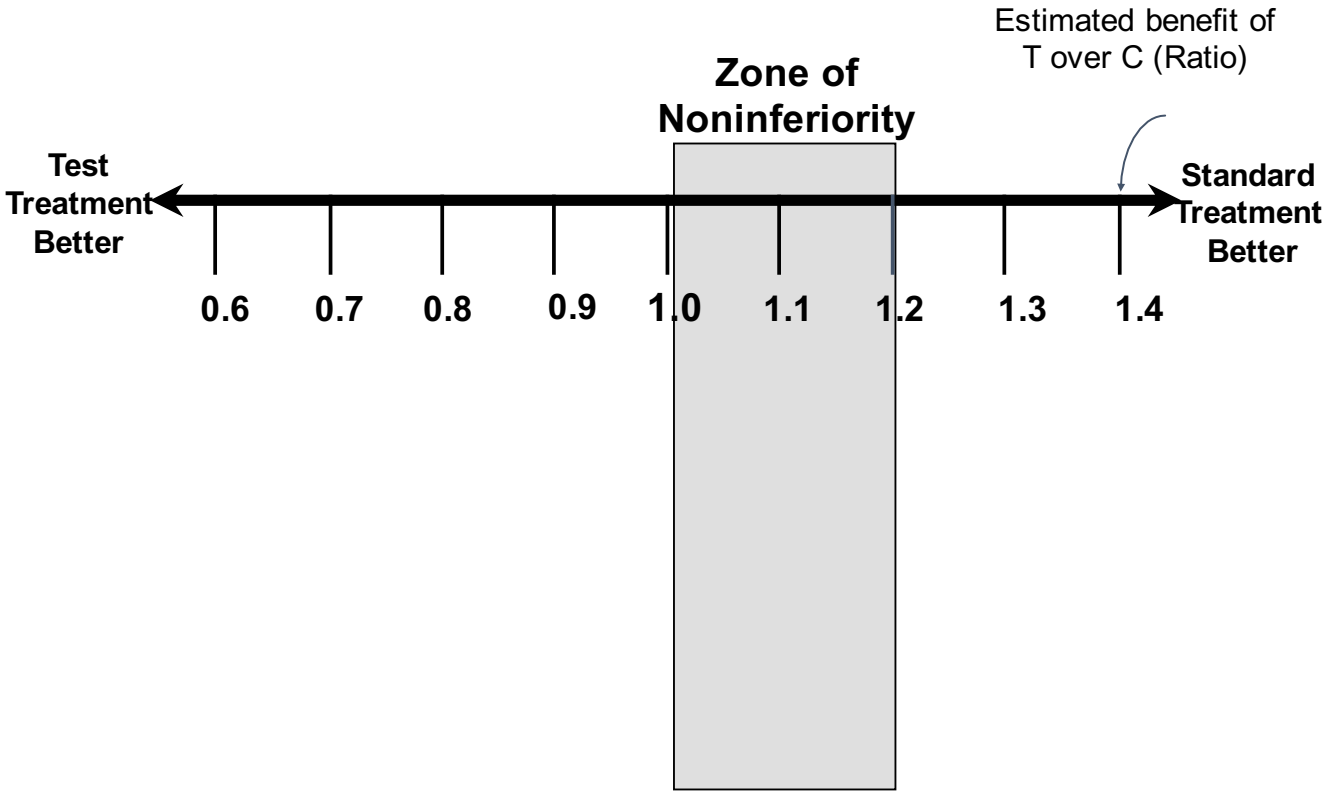
The role of null and alternative hypotheses are reversed in NI trials.

Suppose on relative risk scale, the hazard ratio of Treatment vs control should be no larger than 1.20

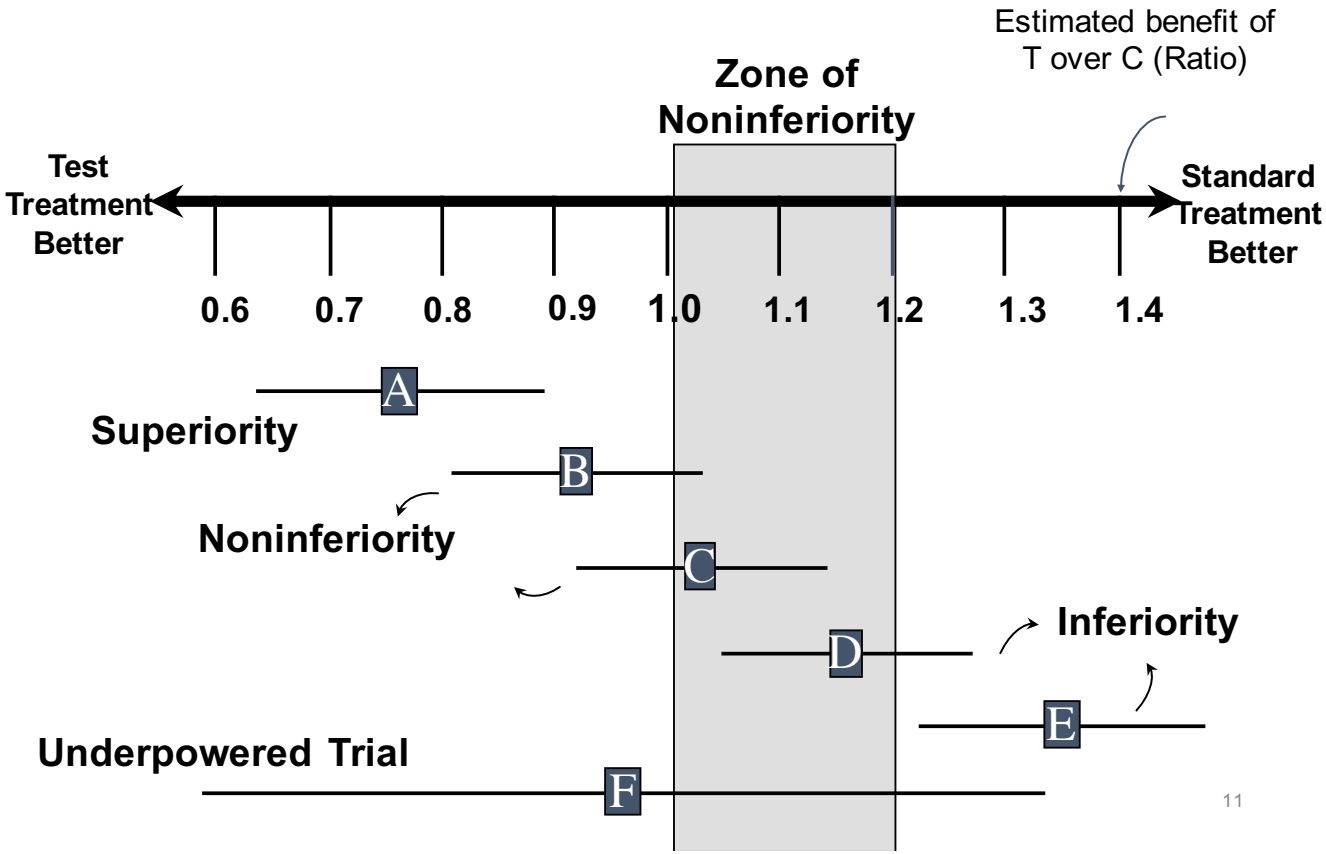
- 1.20 is an example of a non-inferiority margin
- The null hypothesis ($H_0$) is: Relative Risk $\geq 1.20$
- The alternative hypothesis ($H_A$) is: Relative Risk $< 1.20$

Small $p$-values lead to rejection of $H_0$

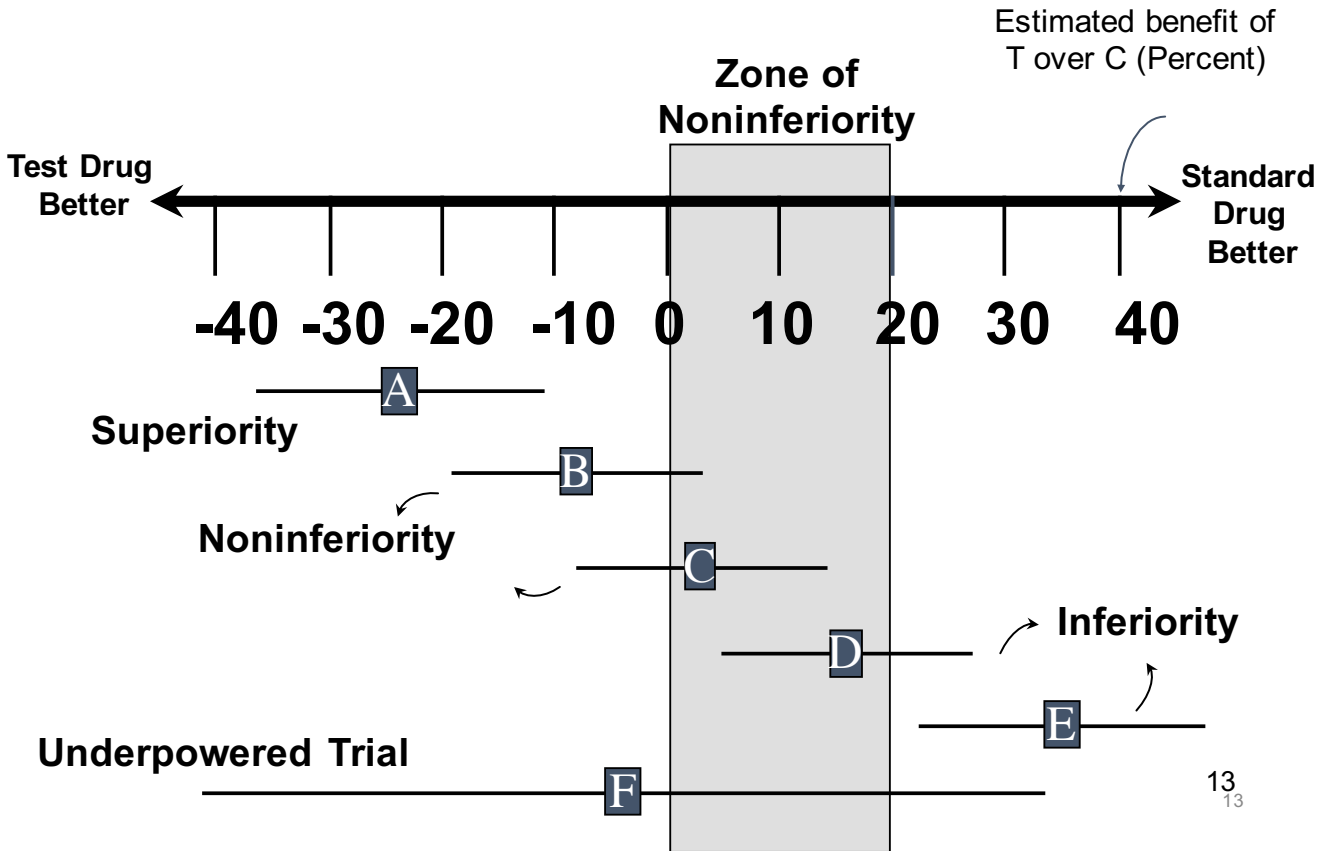
## Ratios of Event Rates :
## Test Drug/Standard Drug



## Ratios of Event Rates :
## Test Drug/Standard Drug



## Difference in Event Rates :
## Test Drug - Standard Drug




13
13

# Threats to Validity of NI Designs

Non-inferiority designs have some intrinsic limitations that make them more difficult to design and more vulnerable to problems than the superiority design.

The following are the most important issues:

- Assay Sensitivity: possibility that $T$ and $C$ are ineffective, possibly from lack of adherence
- Assay Constancy: $C$ is still as effective as in historical trials
- Dropouts can make two treatments seem more similar than they are.

Some Questions . . .

# Explaining false positives

Denise G asked for clarification on how false positives arise when looking at several endpoints.

Here is a different explanation than the one I gave last week

- $p$-values are calculated assuming null hypothesis is true, i.e., no treatment effect.
- Setting a $p$-value threshhold of 0.05 (the type I error) means that 5% of the time, a p-value will be significant when there is no treatment effect.
- Extreme example: if one examines 100 subgroups, 5 of 100 comparisons (5%) will be significant even when there is no treatment effect on any subgroups.
- So a chance spruious finding will be very high.

# Propensity scores

Best explained verbally, with reference to the example of confounding found later in these slides.

# Important issues in an analysis

# Main issues in analysis

- Should the primary analysis of RCT be intent to treat (ITT) or per protocol (PP)?
- Did the randomization produce balance?
- Confounding minimized in observational studies? Any inappropriate claims of causality?
- Substantial missing data or attrition that may affect results?
- What are the summary statistics for the outcome?
- What does the primary analysis show?
    - Focus on confidence intervals, not p-values

# Results of an RCT

- Is the effect size important?
- Was the event rate in the control group close to the design specification?
- Are the trial results consistent? Across subgroups, across secondary endpoints?
- Are the conclusions in the discussion supported by results? Do they match the conclusions in the abstract?
- Is the biological rationale for the outcome plausible?

# Effect size important?

Varies with context

- No prior progress in a disease? Smaller effect size may be important
- Public Health problem? Vaccine studies
- Other, less toxic treatmens available?

For RCT, we sometimes look at Number Needed to Treat (NNT).

- See *Can this treatment help me?*, Frakt and Carroll, NYT 26 Jan 2015
- Briefly, number of people who need to be treated for one patient to derive benefit.

# What happens when there is attrition?

Attrition in an RCT can cause several problems

- Random attrition reduces effective sample size
- Missing data from non-random attrition can cause bias

Two strategies in general use:

- Intent-to-treat (ITT)
- Per protocol (PP)

Neither is perfect

# Intent-to-treat (ITT) vs. per-protocol (PP)

ITT: analyze according to assigned treatment, not treatment received.

Main justification:

- p-values are calculated assuming no treatment difference (the null hypothesis)
- Under that assumption, assigned treatment does not affect outcome.
- p-values will be correct (valid) when comparing the two groups according to treatment assignment.

Example may help make this clear.

# Simple trial, success vs failure outcome, no difference, non-random crossover

Suppose two treatments ($A$ and $B$) are equally effective.

100 participants randomized to each treatment.

ITT table:

| Response | Treatment $A$ | Treatment $B$ |
|---|---|---|
| Success | 40 | 40 |
| Failure | 60 | 60 |

Now assume, after randomization:

- 10 participants with good prognosis (future responders) switch from $A$ to $B$
- 10 participants with bad prognosis (future non-responders) switch from $B$ to $A$

# Simple trial, but with selective crossovers.

Two treatments still equally effective.

Table for the as-treated groups

| Response | Treatment $A$ | Treatment $B$ |
|---|---|---|
| Success | 30 | 50 |
| Failure | 70 | 50 |

An as-treated analysis would imply $B$ more effective than $A$

# ITT can be biased when there is a real treatment effect (random crossovers)

Suppose $B$ is more effective than $A$, so for 100 in each group:

| Response | Treatment $A$ | Treatment $B$ |
|---|---|---|
| Success | 30 | 50 |
| Failure | 70 | 50 |

Assume 10 randomly chosen participants from each group switch treatments, after randomization.

- 10 $A \to B$, 5 respond, 5 do not
- 10 $B \to A$, 3 respond, 7 do not

# Table with just patients who do not switch

| Response | Treatment $A$ | Treatment $B$ |
|---|---|---|
| Success | 27 | 45 |
| Failure | 63 | 45 |

Attrition did not change measured success rates

- but it does reduce the effective sample size

# ITT Table with assigned treatment, real response patients

*A* gets 5 responders (who received B)

*B* gets 3 responders (who received A)

| Response | Treatment *A* | Treatment *B* |
|---|---|---|
| Success | 32 | 48 |
| Failure | 68 | 52 |

Apparent success rate:

- *A* 32% vs. 30% before crossover
- *B* 48% vs. 50% before crossover

Response proportions have moved closer together.

Non-random attrition can also cause bias in the analysis because of missing data

# Missing data

Some data are almost always missing

- Some measurements might not be taken, some forms might not have been completed.
- Almost never malicious, almost always unavoidable.

An example

- Patients experiencing side effects from a drug stop taking the drug.
- Response to the drug probably not available for those patients
  - Such patients may be systematically different

# Methods of analysis with missing data

Drop the cases with some missing observations.

- Usually the worst option.

Pharma often used Last Observation Carried Forward (LOCF) in the past.

- Also not recommended.

Model the missingness mechanism.

- Can work well if the modeling is done carefully
- Many methods, many favor multiple imputation

# Attrition: some general remarks

Attrition is only a problem when it is substantial.

- Statisticians begin to worry when more than 10% of participants are lost
- Substantial attrition turns a randomized study into an observational study.
- Methods for observational studies are discussed later.

# Endpoints

Clearly defined, biological rationale, measurement minimizes bias, clinically relevant?

- Measurement bias can be subtle
  - Different evaluation schedules on two treatments can lead to bias.
  - Lack of blinding caused by different side effects

Endpoints best summarized using both single value estimates and confidence intervals

# Outcome summary measures

Differences in average outcome

Response rates

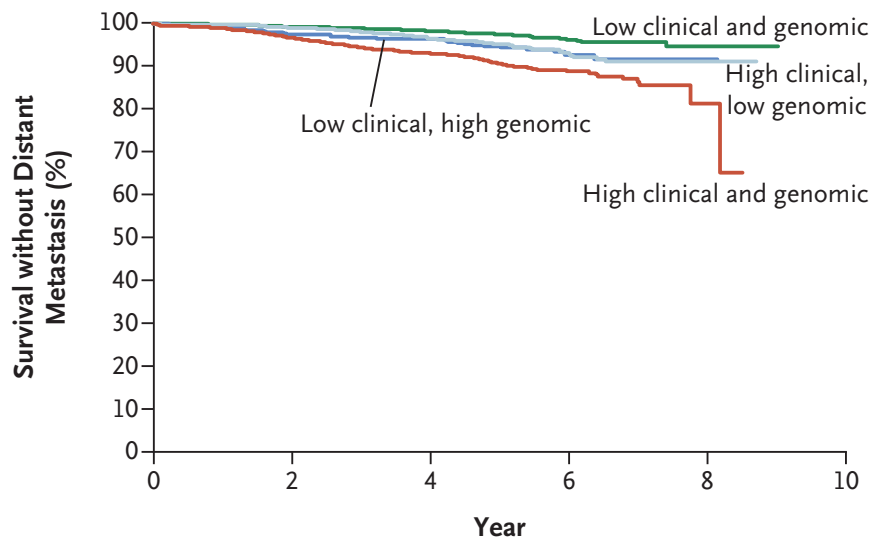
- Relative Risk
- Odds Ratio

Time-to-event studies

- Hazard Ratios
- Median times to an event

# Time-to-event - Genomics and Breast Cancer, 25 Aug NEJM

## Survival curves: survival without distant mets




**No. at Risk**

| | | | | | |
|---|---|---|---|---|---|
| Low clinical and genomic risk | 2745 | 2628 | 2331 | 735 | 33 |
| Low clinical, high genomic risk | 592 | 550 | 484 | 136 | 2 |
| High clinical, low genomic risk | 1550 | 1457 | 1317 | 311 | 9 |
| High clinical and genomic risk | 1806 | 1689 | 1462 | 395 | 11 |

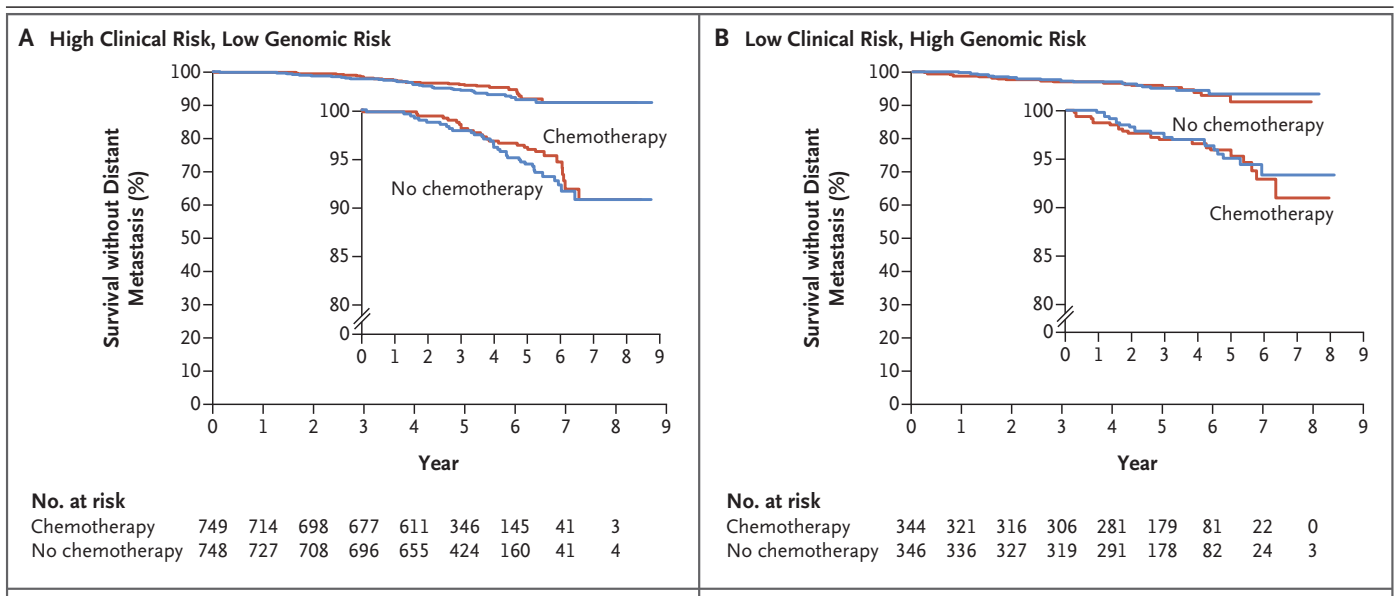
**Figure 3. Survival without Distant Metastasis in the Four Risk Groups.**

The analysis includes all enrolled patients, and the risk groups are based on corrected risk. The time-to-event curves were estimated by means of the Kaplan–Meier method.

# Time-to-event - Genomics and Breast Cancer, 25 Aug NEJM

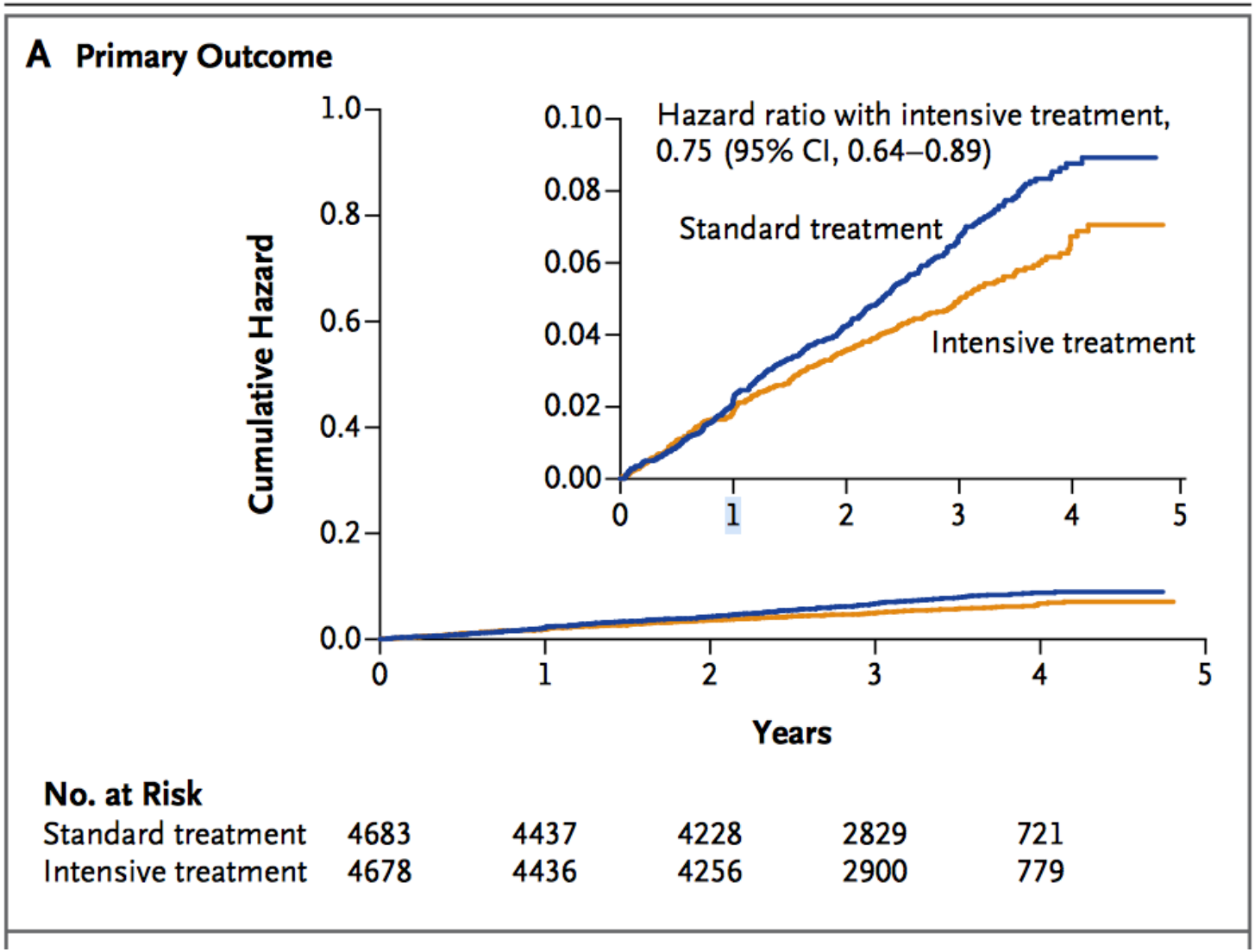
Confidence intervals for outcome: High clinical, low genomic risk:

- No chemo: 5-yr estimate: (92.5%, 96.2%)
- Hazard ratio, chemo vs no-chemo: (0.5, 1.21), p = 0.27



# Time-to-event outcomes - SPRINT

Cumulative hazards and hazard ratio



# Confounding in observational studies

# East Boston study of adolescents

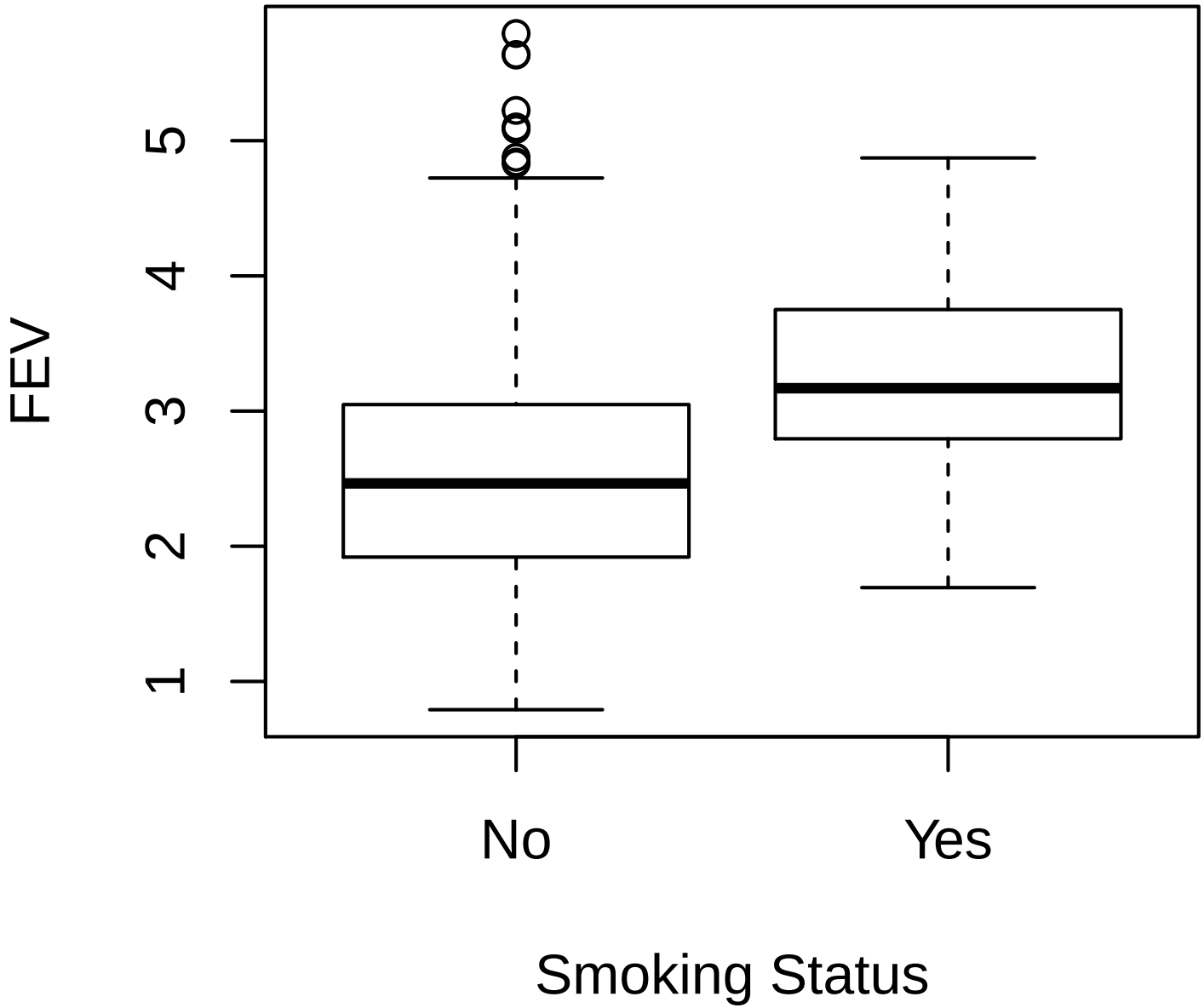
Data from a study examining lung function in 654 children from East Boston.

Among many variables, study recorded

- Age
- Smoking status (yes/no)
- Lung function, forced expiratory volume of air expelled in 1 second (fev)
    - Measured in liters

Tager, et. al *Am J Epi*, 1979

# Association between lung function and smoking



# Confounding

Previous plot suggests that smoking is associated with an average increase in lung function

- Smokers have (on average) an FEV that is higher by 0.7 liters/sec

Association can be caused by confounding.

A confounding variable is

- Associated with outcome
- Associated with exposure

Natural variable to examine here is age

# Adjusting for confounding

We use statistical models to estimate the simultaneous association

- Age and smoking status with FEV.
- Perhaps age, height and smoking status with FEV

If it works, it will show the association of smoking and FEV

- After adjusting for age and height.

# Model based adjustments

Association of smoking and FEV, after adjusting for

- age: decrease in FEV of approximately 0.30 liters/sec
- age and height: decrease of approximately 0.10 liters/sec
- age, height and age by height interaction: decrease of 0.17 liters/sec

Depending on model, estimated association varies

- But all adjustments suggest smoking is associated with decrease in lung function

# Some caveats with model based adjustments

Most datasets are far more complex than the East Boston study

We have no way of knowing which adjustment is correct

- Specific numerical values of adjusted associations less important than direction.

Statisticians discourage the use of the phrase

- 'Association of smoking and fev, after *controlling* for . . . '

Watch out for implied causal claims

# Other ways to adjust for confounders

Analyse separately by subgroups (stratify)

Match each smoker with a nonsmoker with similar age and height (matching)

- Requires a large dataset, difficult to do if it matching was not built into the design

Create a model for the probability of smoking, given participant characteristics

- The match smokers, nonsmokers with similar probabilities
- Called propensity score matching

All approaches less reliable when there are unmeasured confounders.

# Closing out

# These points bear repeating

We do not expect designs or analyses to be perfect

We look for

- Transparency
- Following principles of scientific investigations: formulating a hypothesis, then testing it
- Use of 'state of the practice' methods
- Sensitivity analyses: do different models lead to the same qualitative conclusion