Fundamentals of Interventional Trials: "Statistics for the Masses"

Ajay J. Kirtane, MD, SM

Columbia University Medical Center
The Cardiovascular Research Foundation

CARDIOVASCULAR RESEARCH FOUNDATION

COLUMBIA UNIVERSITY MEDICAL CENTER

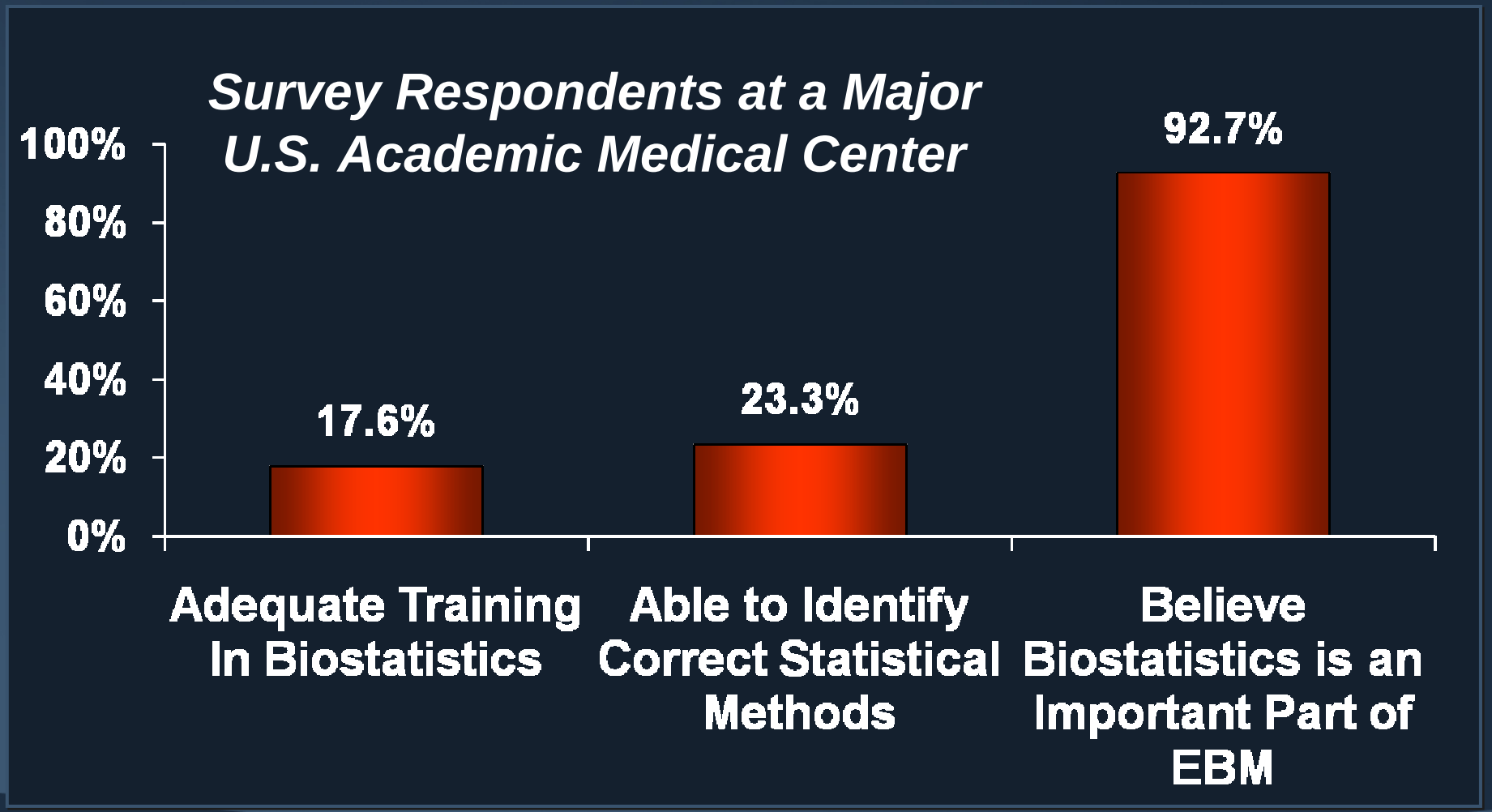
The Problem with Statistics

- "You can drag a p-value out of a stone…"
- "98% of all statistics are made up"
- "Statistics can be made to prove anything – even the truth"
- "Facts are stubborn things, but statistics are more pliable"
- "Statistics are no substitute for judgment"

  (Henry Clay)

The Bigger Problem

Statistics are Here to Stay!



CARDIOVASCULAR RESEARCH FOUNDATION

West and Ficalora, *Mayo Clin Proceedings* 2007; 82(8):939-9.

COLUMBIA UNIVERSITY MEDICAL CENTER

Outline of this Talk

- Significance Testing
- Basic and Advanced Clinical Trial Design
  - Types of Studies
  - Sample Size and Power
- Endpoints in Interventional Studies
- The Impact of Routine Angiographic Follow-up and The Oculostenotic Reflex

Significance Tests and P-values

- Clinical trial comparing two treatments
- Control BIAS by randomization, blinding (whenever possible)
- How strong is the evidence that an observed treatment difference is real?
  - Could it be due to chance?
  - Perform a significance test
  - Get the P-value
  - Small P means strong evidence

Significance Tests - Types of Data

- Binary outcome →
  - Parametric: Chi-Squared test
  - Non-parametric: Fisher's Exact test
  - Adjusted: Logistic Regression
  - e.g. target lesion revascularisation (yes/no)
- Time to event outcome →
  - Logrank test
  - Adjusted: Cox Regression
  - e.g. time to death

Significance Tests - Types of Data

- Quantitative outcome →
  - Parametric: t-test
  - Non-parametric: Wilcoxon Rank Sum
  - Adjusted: Linear regression, ANOVA
  - e.g. late loss (mm)

Binary Outcome (Example)

- TYPHOON trial [NEJM 14 Sept 2006]
- DES vs BMS in Primary PCI

	SES	BMS
Total Number of Patients	355	357
Number with Endpoint	26	51
Percentages (Frequency)	7.3% (26/355)	4.3% (51/357)

Binary Outcome (Estimation)

- **TYPHOON trial: 7.3% vs 14.3%**
- **Relative Risk is the ratio:** $\frac{7.3}{14.3} = 0.51$

  Relative Risk = 1 no difference
  Relative Risk >1 new treatment worse
  Relative Risk <1 new treatment better
- **Relative Risk Reduction**
  = 100 x (1–relative risk) = 49%
- **Absolute Risk Reduction = 14.3% - 7.3% = 7.0%**
- **Number needed to treat =** $\frac{100}{\text{absolute \% reduction}}$ **= 14**

Binary Outcome (Estimation)

- **Odds ratio =**

$$\frac{7.3}{100-7.3} \div \frac{14.3}{100-14.3} = 0.47$$

- **If percentages are small, odds ratio and relative risk are similar**

Binary Outcome (Chi-squared test)

- Null Hypothesis: both stents are equally effective
- If null hypothesis is true:
  - What's the probability (P) of getting a difference 7.3% versus 14.3% or bigger?
  - Answer: P = 0.004
- Strong evidence that drug-eluting stent reduces risk of primary endpoint (i.e. we reject that the null hypothesis is true)

Significance Testing

Alpha and the Magical <0.05 Threshold

- Researchers are trained to have an endorphin surge when they see the text "P<0.05"!!!
- But this is somewhat arbitrary… this just means that we accept that there is a 5% or less chance that the results observed (for example showing a difference between two stents) could be due to chance alone
  - Alpha is the "False positive" rate

Significance Testing

- P-values measure the strength of evidence against the null hypothesis
- P < 0.05, statistically significant at 5% level
  - does not mean PROOF of a treatment difference (just evidence)
  - it's an arbitrary guideline
- P > 0.05, "not statistically significant"
  - does not mean no difference exists
  - maybe the study was too small

Binary Outcome

- Confidence Interval - Expresses uncertainty in the estimate
- TYPHOON trial:
  - Observed relative risk = 0.51
  - 95% confidence interval is 0.33 to 0.80
  - 95% sure true relative risk is in this interval
  - 5% chance true relative risk is outside the confidence interval
- Bigger study means tighter confidence interval
- To halve the width, need 4 times the trial size

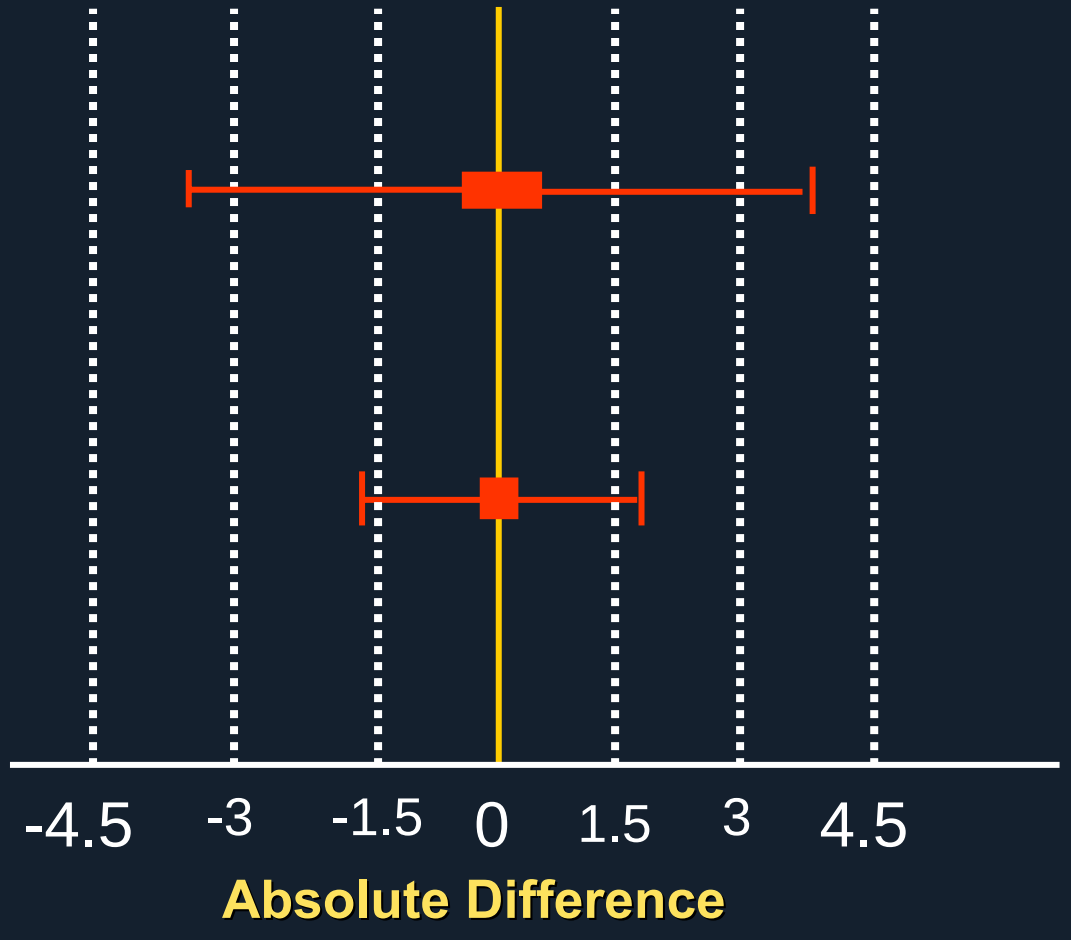
Confidence Intervals of Difference in ST rates between DES A and DES B

In this small trial:

Though there is no difference, we cannot exclude a truly larger difference!

In a larger trial (with 10X as many patients and the same event rates):

There might still be a difference, but we feel more certain of the rates!



Link Between P-value and Confidence Interval

- P < 0.05 means

  95% Confidence Interval for relative risk/odds ratio does not include 1

- P > 0.05 means

  95% Confidence Interval includes 1

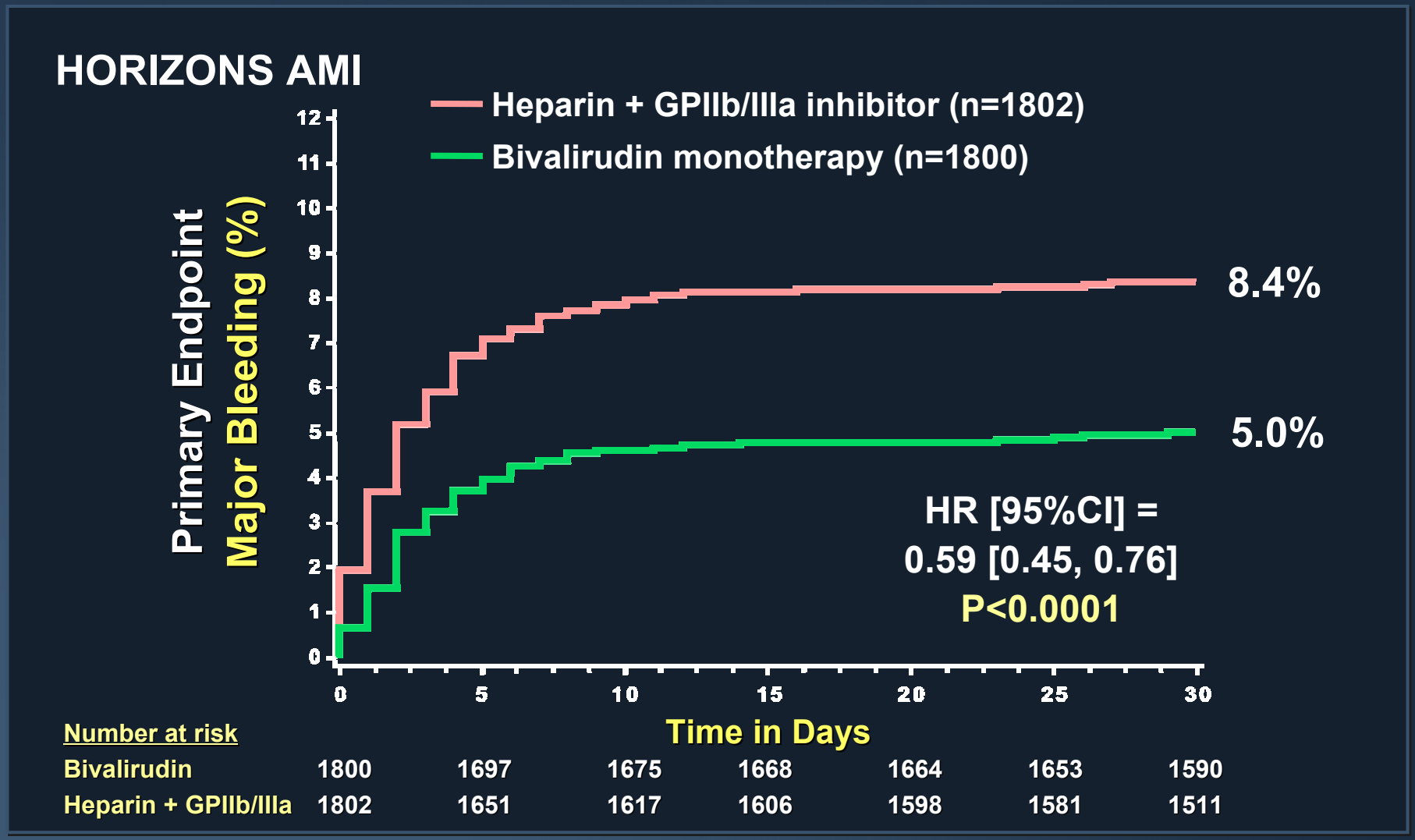
Time to Event Outcome (Survival)

- Kaplan-Meier plots
  - Estimates the population survival curve
  - Allows estimation of survival over time, even when patients drop out or are studied for different lengths of time
  - Be careful of apparent large late differences
    - CONFIDENCE INTERVALS?
      - not usually presented
    - CHECK THE SIZE OF THE RISK SET!

Time to Event Outcome

- Logrank test
  - Compares survival curves by comparing observed to expected at each event time
- Hazard Ratio
  - Similar to relative risk
  - Instantaneous relative risk averaged over time using a Cox proportional hazards model
  - If event rate is low, hazard ratio ~ relative risk

Time to Event Outcome (Example)



CARDIOVASCULAR RESEARCH FOUNDATION

COLUMBIA UNIVERSITY MEDICAL CENTER

Quantitative Outcome

- TAXUS IV trial [NEJM 15 Jan '04]
- Late loss (mm)

	PES	BMS
Total Number of Patients	292	267
Mean Late Loss	0.23 mm	0.61 mm
Standard Deviation	0.44 mm	0.57 mm
Standard Error of Mean	0.026 mm	0.035 mm

- Standard deviation measures spread
- SEM = SD/$\sqrt{N}$ measures precision of the mean

Quantitative Outcome

- Mean treatment difference = 0.38mm
- Standard error of Difference (SE) =

$$\sqrt{.026^2 + .035^2} = 0.044 \text{ mm}$$

- Two-sample t-test

  t = 0.38 / 0.044 = 8.72

- P < 0.0001
- 95% CI = mean difference $\pm$ 1.96 x SE

  = 0.29 to 0.47 mm

Basic Clinical Trial Designs

Prospective vs. Retrospective

- **Prospective Studies:**
  - Patients are followed *forward* in time, and data is gathered at baseline and as events happen
- **Retrospective Studies:**
  - All events have happened already, and patients' information is collected from the past (through chart review, phone calls, etc).
- **Hybrids: for example, database queries from prospectively conducted studies**

Basic Clinical Trial Designs

Cohort vs. Case-Control

- **Cohort Studies:**
  - **The vast majority of studies are cohort studies – a group of patients followed until they have events.**
- **Case Control studies can be useful for analyzing low-frequency events:**
  - **Patients with events of interest (e.g. stent thrombosis) are identified and then matched to patients without events to determine predictors of events**

Basic Clinical Trial Designs

Randomized vs. Observational

- **Randomized Studies:**
  - **Patients are allocated to a treatment (e.g. DES vs. BMS) randomly, thus reducing bias**
- **Observational Studies:**
  - **Data is analyzed based upon what treatment was received (e.g. DES vs. BMS)**
  - ***The reasons for treatment received may be subject to "confounding" or bias***

Basic Clinical Trial Designs

Confounding in Observational Studies

- The problem with DES vs. BMS observational studies is that treatment may be influenced by other factors…
  - Hospital A usually treats "sick patients" with DES and "less sick patients" with BMS
  - Right before the abstract deadline, data is analyzed, and the "unadjusted" (crude) rate of death is twice as high with DES compared to BMS
  - I sense a late-breaking trial!!!!

Basic Clinical Trial Designs

Confounding in Observational Studies

- **BUT… If the data is analyzed after "adjustment" for whether patients were deemed "more sick" or "less sick" by the treating physicians, one might find no differences**
- **Two fundamental problems:**
  - **How can one effectively capture "sick" status in the database?**
  - **If it isn't captured, YOU CANNOT RISK-ADJUST**

Basic Clinical Trial Designs
Are Randomized Trials the Answer?

- Generally, RCTs will provide higher quality data than observational studies
  - Randomization can reduce treatment biases by simply "flipping a coin"
- However, there are several caveats:
  - Patients in RCTs are highly selected (e.g. COURAGE)
  - Conduct in RCTs is not always representative
    - Specifically, is the control group really reflective of practice?
  - Blinding (at all stages) is very important

Randomized Trials vs. Registries:

Each has strengths and weaknesses

RCTs	Registries
Equal distribution of measured and unmeasured confounders	
High-level study processes (event reporting, monitoring, adjudication)	
Regulatory body oversight / potentially higher level of quality control	
Ability to study complex or high risk cohorts	
More generalizability / less selection bias	
Robust sample size	
Less potential for “artificial” study processes (e.g. routine angiographic f/u)	

Basic Clinical Trial Designs

Types of Comparisons

- Superiority vs. Non-Inferiority
  - Superiority aims to show that one therapy is definitively different ("better") than another
  - Non-Inferiority aims to show that one therapy is no worse than another

Basic Clinical Trial Designs
A Superiority Trial

- 100 patients randomized to DES A vs. BMS
  - 10 clinical restenosis events with DES A, 20 with BMS
- Basic superiority question: What is the probability that this difference could have occurred by random chance?
  - *If this is <5%, then DES A is superior to BMS*

Basic Clinical Trial Designs
A Superiority Trial

- In this case, we would be comparing 10% vs. 20%, which seem quite different
- BUT the number of patients (total number of events), not just the percentage, matters

Also, remember that if there have been 19 BMS events instead of 20, there would have been no statistically significant difference between the two groups!

Basic Clinical Trial Designs

A Non-Inferiority Trial

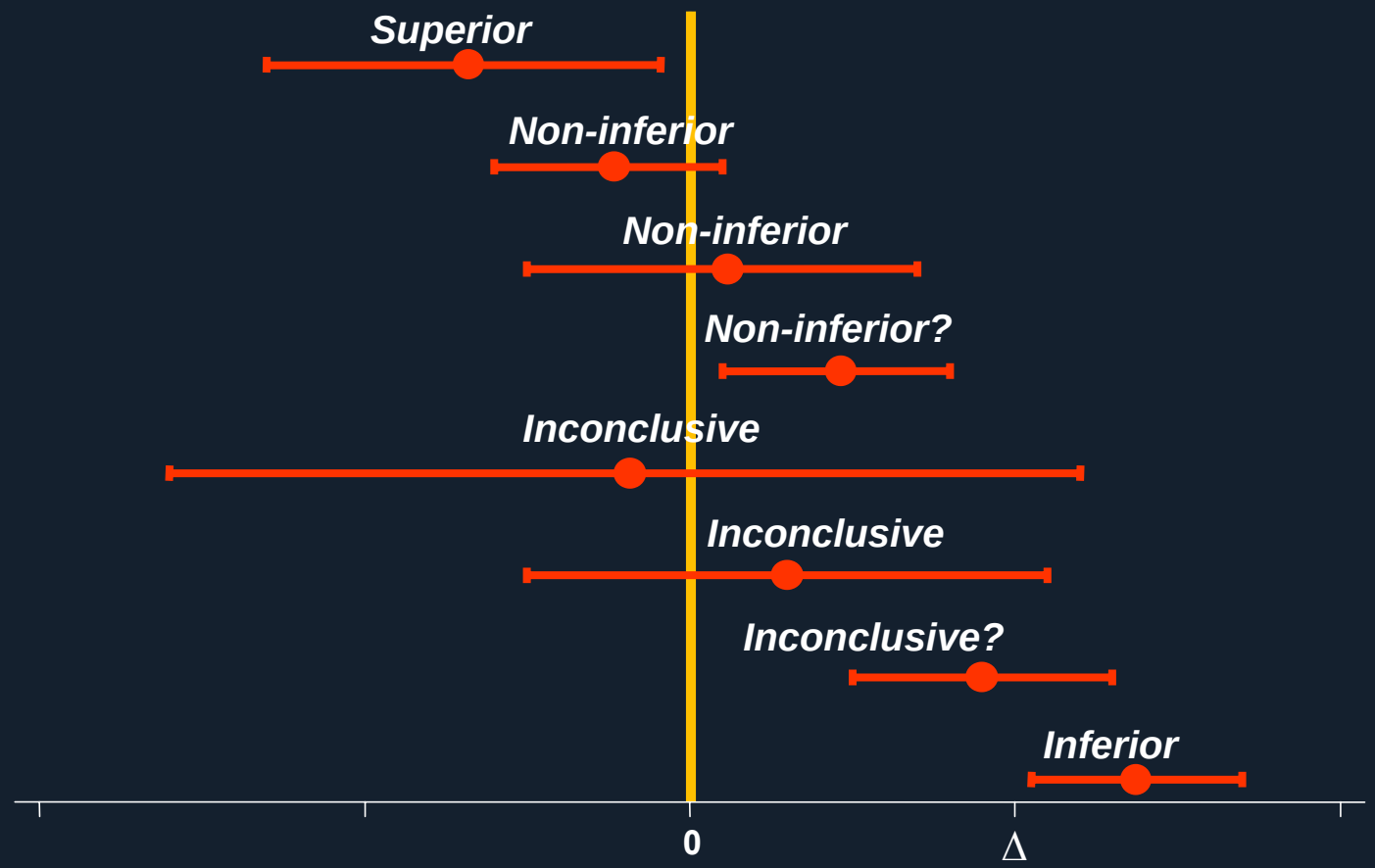
- 1000 patients randomized to DES A or DES B
  - 70 clinical restenosis events with DES A; 50 with DES B
- Basic non-inferiority question: What is the probability that DES A could be worse than DES B?

Basic Clinical Trial Designs

Non-Inferiority Explained

- In order to define "worse", we first need to figure out how much "worse" we are willing to accept
  - This is defined as the non-inferiority "delta" or "margin" and is almost always ARBITRARY
  - Commonly set at 20%, but can vary greatly
- Thus, we want to be sure (or at least 95% confident) that DES A is not 20% worse than DES B

Possible Non-Inferiority Results



Basic Clinical Trial Designs

Non-Inferiority

- Increasingly common as it gets harder and harder to demonstrate incremental gains over successful therapies
- Tempting to choose a delta too large to reduce sample size
- Risk of accepting less and less efficacious therapies = “creep”
  - If I am not inferior to a “straw man” comparator with a broad delta, does that make me the best?

Basic Clinical Trial Designs

Superiority vs. Non-Inferiority?

- A larger sample size will generally be needed to demonstrate superiority over non-inferiority, particularly when differences in therapies are small
  - DES vs. DES for example
- The only exception to this is when the magnitude of effect is large
  - DES vs. BMS for example

Basic Clinical Trial Designs

Superiority vs. Non-Inferiority?

- Two stents are being compared in a trial
  - Stent A: Expected rate of TVF: 10%
  - Stent B: Expected rate of TVF: 15%
- Numbers of patients for typical study (at 80% power)
  - Superiority: 686 patients per arm (1:1 randomization)
  - Non-inferiority:
    - 479 patients per arm, accepting 1% delta
    - 354 patients per arm, accepting 2% delta

Basic Clinical Trial Designs

Pros/Cons of Unbalanced Randomization

- Typical Randomization is 1:1
  - 1:1 affords greatest ability to detect differences between two treatments (highest efficiency)
- But do we really want to treat so many patients with a control stent and “reinvent the wheel”?
  - What if we randomized 2:1, or 3:1 and in doing so were able to gather more data about a new stent?
  - Regulatory bodies often require large bodies of data on a new device

Basic Clinical Trial Designs

Pros/Cons of Unbalanced Randomization

- Back to our example – Head-to-head stent trial
  - Stent A: Expected rate of TVF: 10%
  - Stent B: Expected rate of TVF: 15%
- For a typical superiority design (80% power at $p<0.05$)
  - 1:1 needs 686 patients with each stent (total 1372 patients)
  - 2:1 needs 1004 with Stent A; 502 with Stent B (total 1506 patients)
  - 3:1 needs 1322 with Stent A; 441 with Stent B (total 1763 patients)
- ***While we treat proportionally less patients with the control stent, the overall study size increases (so does cost!)***

Basic Clinical Trial Designs

The importance of "Power" in Statistics

- Power Calculations are *critical* when designing studies
  - A RCT is not intrinsically better than an observational study if not adequately powered… in fact, it can be more misleading!
- Power is defined as the ability to be able to statistically detect a difference *when one is truly present*

Basic Clinical Trial Designs

Statistical Power

- **Power is typically set at 80% (or higher)**
  - ***Thus, we accept a 1 in 5 possibility ("fall of the cards") that even if there is an actual difference between 2 stents, we will not be able to find a difference!!!***
  - ***1-power equals the "False Negative" rate***
- **When there is a lot riding on a trial, how much risk can you assume???**
  - **A 10% increase in power (to 90%) will increase sample size!!!**

Basic Clinical Trial Designs

The Importance of Power Calculations

- If I flip a coin twice and it comes up heads once and tails once, does it definitively mean that the coin is fair (or has a 50/50 chance of heads)?

- On the other hand, if I flip a coin twice and it comes up heads twice, does that mean that it will never come up tails (or that it will come up heads twice as often)?

Basic Clinical Trial Designs

The Importance of Power Calculations

- **Underpowered studies:**
  - **When they are negative:**
    - **Can make two therapies seem similar when in fact differences might exist**
  - **Confidence intervals can help clarify the picture and determine how certain one can be with the results**

Hypothetical Underpowered Trial

- DES A vs. DES B with 500 patients randomized (250 per group)
- 30 day rate of stent thrombosis:
  - 4 events (1.6%) for DES A
  - 4 events (1.6%) for DES B
- Does this mean there are truly no differences between DES A and DES B?

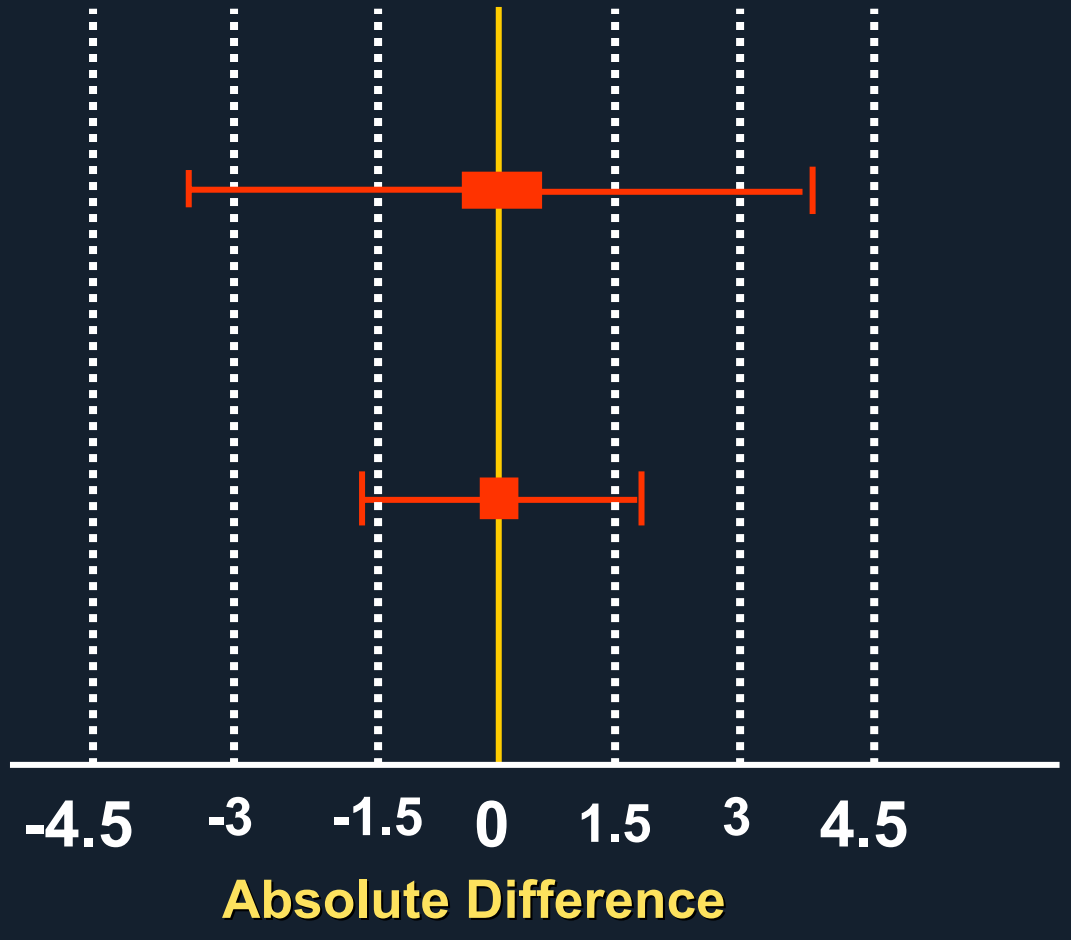
Confidence Intervals of Difference in ST rates between DES A and DES B

In this small trial:

Though there is no difference, we cannot exclude a truly larger difference!

In a larger trial (with 10X as many patients and the same event rates):

There might still be a difference, but we feel more certain of the rates!



Confidence intervals don't lie!

CARDIOVASCULAR RESEARCH FOUNDATION

COLUMBIA UNIVERSITY MEDICAL CENTER

Basic Clinical Trial Designs

The Importance of "Power" in Statistics

- Underpowered studies:
  - When they are "positive":
    - May only get published if results are statistically significant
    - Typically exaggerate treatment effects
    - Even when differences are "statistically significant", the absolute and relative differences are usually overstated

Basic Clinical Trial Designs

Look at Number of Events Too

- Percentages alone can be very misleading – especially when sample size / events are low
- What if the rate of restenosis is 5% with DES A vs. 10% with DES B? This seems like a big difference, but might not be if there were not many overall patients in the study!

Baseline Rate	Total N	Number of Events	95% Confidence Interval
5%	20	1	[0.1%,24.9%]
5%	100	5	[1.6%,11.2%]
5%	1000	50	[3.7%,6.5%]

Basic Clinical Trial Designs

Beware of "Relative Risk"

- For every relative risk increase (or reduction) the baseline risk will determine the absolute risk increase / the number needed to harm

Baseline Rate	Excess Rate (Relative Risk of 2)	Absolute Risk Increase	Number Needed to Harm
0.5%	1%	0.5%	200

Basic Clinical Trial Designs

Relative vs. Absolute Risk

- Even if the relative risk were twice as great, it is important to consider not only relative risk, but *absolute risk* as well
  - If you sell one share of a $1 stock and it then doubles, are you as upset as if you sold one share of Berkshire Hathaway Series A at $134,100 and *it* then doubled?

Basic Clinical Trial Designs

The Importance of Power Calculations

- An extreme example:
- Late stent thrombosis: 0 events with BMS vs. 5 events with DES (p=0.02)
  - The calculated relative risk is *infinite* – is this biologically plausible?
  - Do we think that the rate of late stent thrombosis is truly 0% with BMS?

Basic Clinical Trial Design

Sample Size Calculations 101

- What do I need?
  - Baseline assumptions:
    - Event rate in treatment group
    - Event rate in control group
  - Other parameters
    - Balanced or Unbalanced Randomization (1:1 or other)
    - Superiority or Non-Inferiority Hypothesis
    - Power (usually 80% or greater)
    - Alpha (almost always 0.05 for two-sided, 0.025 one-sided)
  - A computer program to crunch the numbers

Basic Clinical Trial Design

Audience Poll

- As overall event rates go down, overall sample size goes…. — ***UP***
- As the difference between groups increases, sample size goes…. — ***DOWN***
- As randomization goes from balanced to unbalanced, overall sample size goes…. — ***UP***
- As power goes up, sample size goes.. — ***UP***
- As alpha goes down, sample size goes…. — ***UP***

Intention to Treat Analysis in RCTs

- Usually Primary Analysis (non-inferiority trials may be exception, but not always)
- All patients are analyzed as randomized (enrolled)
  - Eliminates bias
  - Represents treatment strategy
  - Includes withdrawals!!
    - Large withdrawal percentage may indicate more uncertainty in results than indicated by standard p-values and confidence bounds

Intention to Treat Analysis in RCTs

Issues raised by withdrawals:

- Outcome information is usually not available
- Exclusion from analyses
- Dealing with withdrawals in an ITT analysis?
  - Design trial to minimize withdrawal
  - Use alternative source of outcome information when possible (e.g. death registries)
  - Analytic approaches exist – do a sensitivity analysis

Clinical Endpoints in DES Studies

- Death (Cardiac, Non-Cardiac)
- Myocardial Infarction (Q-Wave, NQWMI)
  - But what is the threshold?
  - But what is the assay?
- Target Lesion Revascularization
- Target Vessel Revascularization
- Stent Thrombosis (subcategories)

Clinical Endpoints in DES Studies

- Composite Endpoints
  - Cardiac death/MI
  - Device-oriented (TLF): cardiac death/MI attributable to the target lesion/TLR
  - TVF: cardiac death/MI attributable to the target vessel/TVR
  - MACE: a mixed bag (and varies from study to study!)

Clinical Endpoints in DES Studies

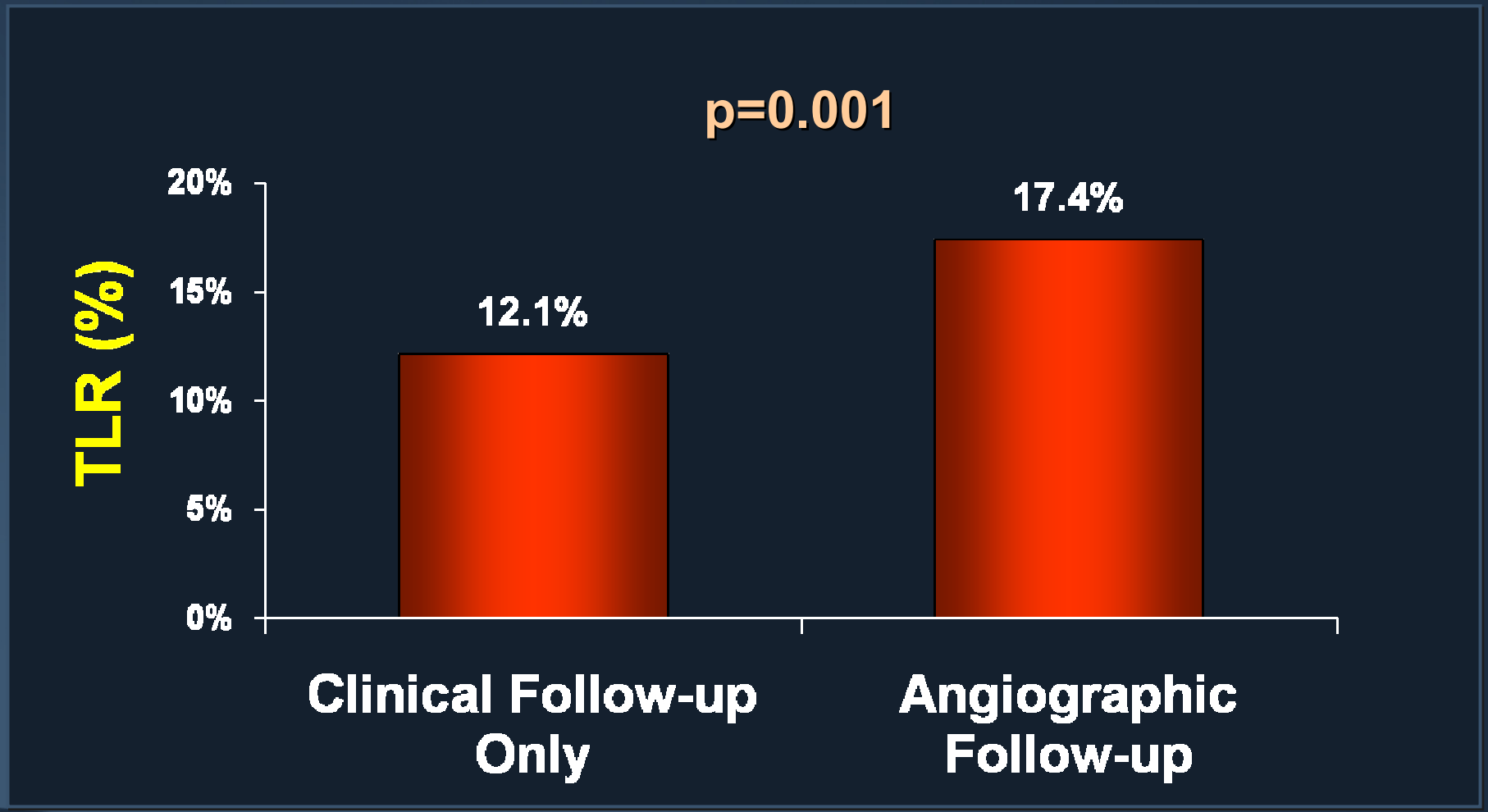
- Endpoint Adjudication is critical
  - Objective assessment of outcomes
    - Preferentially blinded to treatment
    - Minimizes site-specific differences
    - Minimizes conflict of interest concerns
    - Can allow integration of angiogram, QCA (Corelab) data, clinical data, lab data, EKG data

Angiographic Follow-Up In DES Studies

- Routine angiographic follow-up allows mechanistic observations to be made and additionally allows the use of efficacy surrogates
- But angiographic follow-up can bias results!!
  - Oculostenotic reflex
  - Reverse oculostenotic reflex

Impact of Routine Angiographic Follow-up

Rates of TLR from 6 BMS Studies



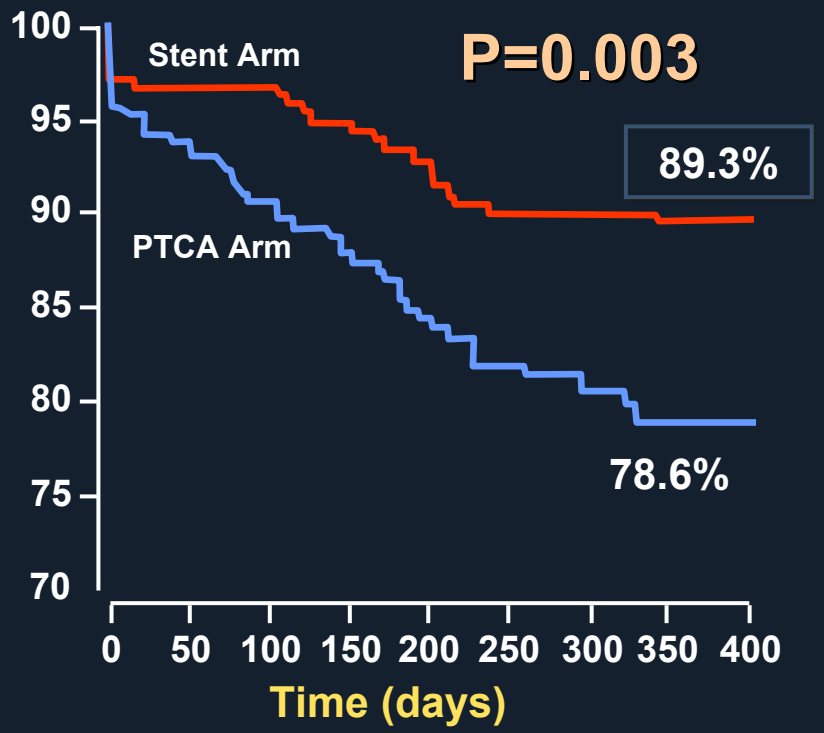
Cutlip et al, *J Am Coll Cardiol* 2002; 40: 2082-2089

CARDIOVASCULAR RESEARCH FOUNDATION

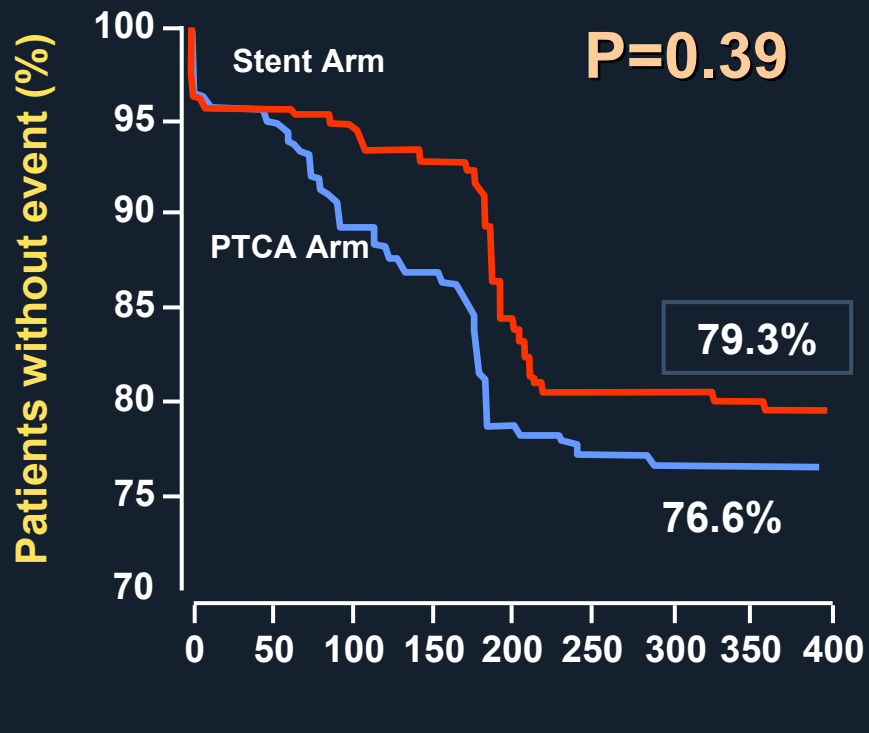
COLUMBIA UNIVERSITY MEDICAL CENTER

Benestent II: Angiographic vs. Clinical F/U

Clinical F/U only

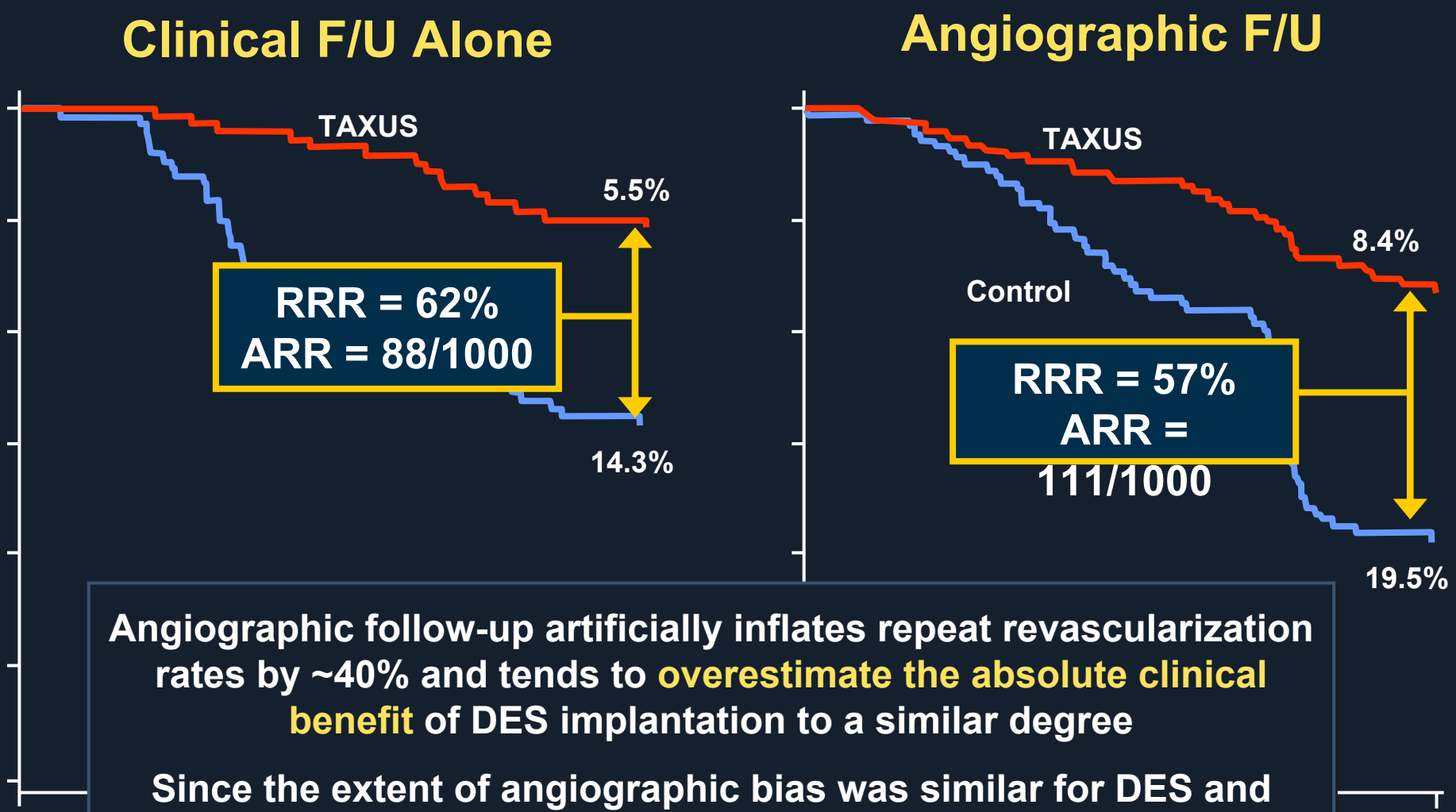


Angio + Clinical F/U



Event Free Survival

Impact of Routine Angiography in TAXUS IV

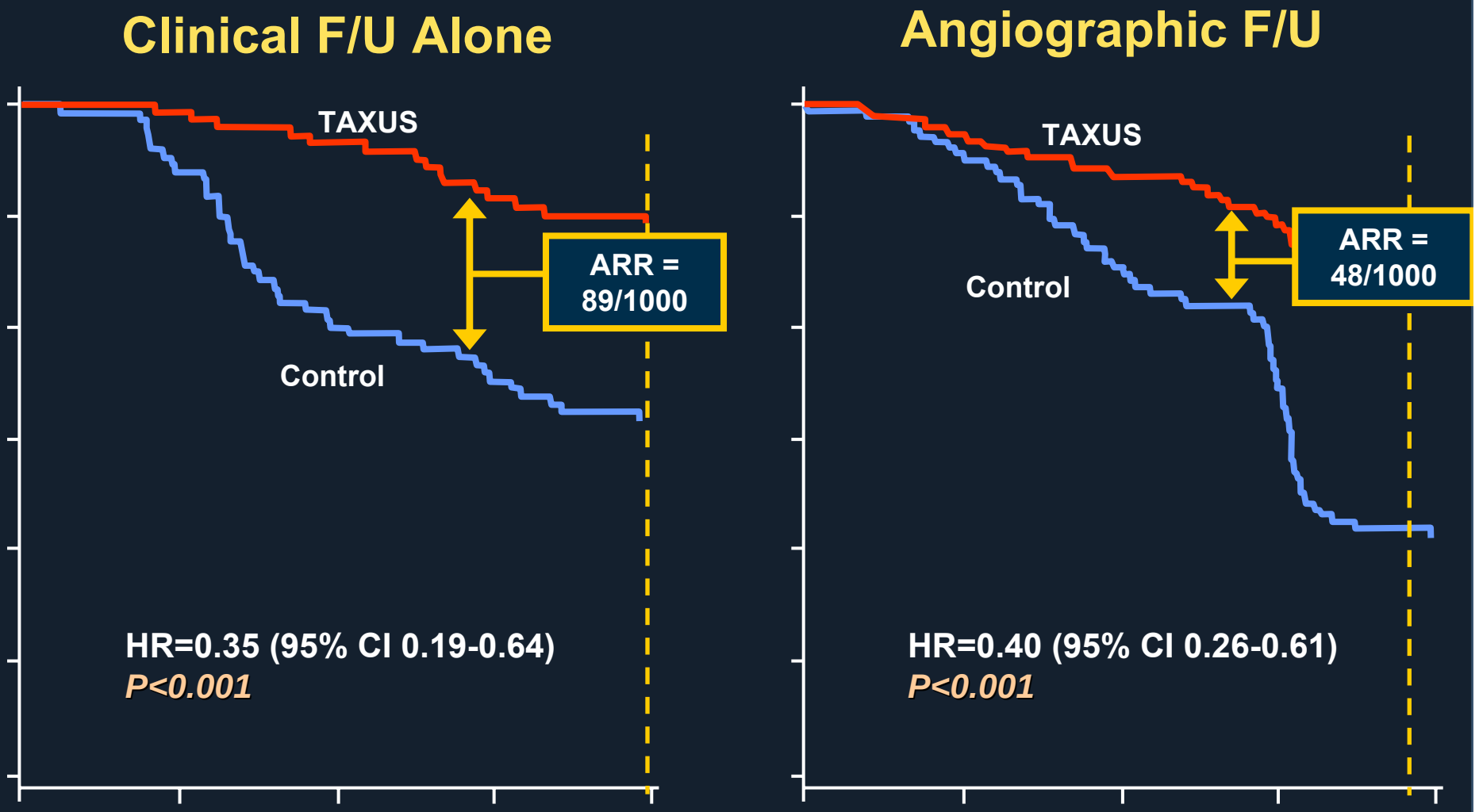


Angiographic follow-up artificially inflates repeat revascularization rates by ~40% and tends to overestimate the absolute clinical benefit of DES implantation to a similar degree

Since the extent of angiographic bias was similar for DES and BMS, however, the relative risk reduction is unaffected

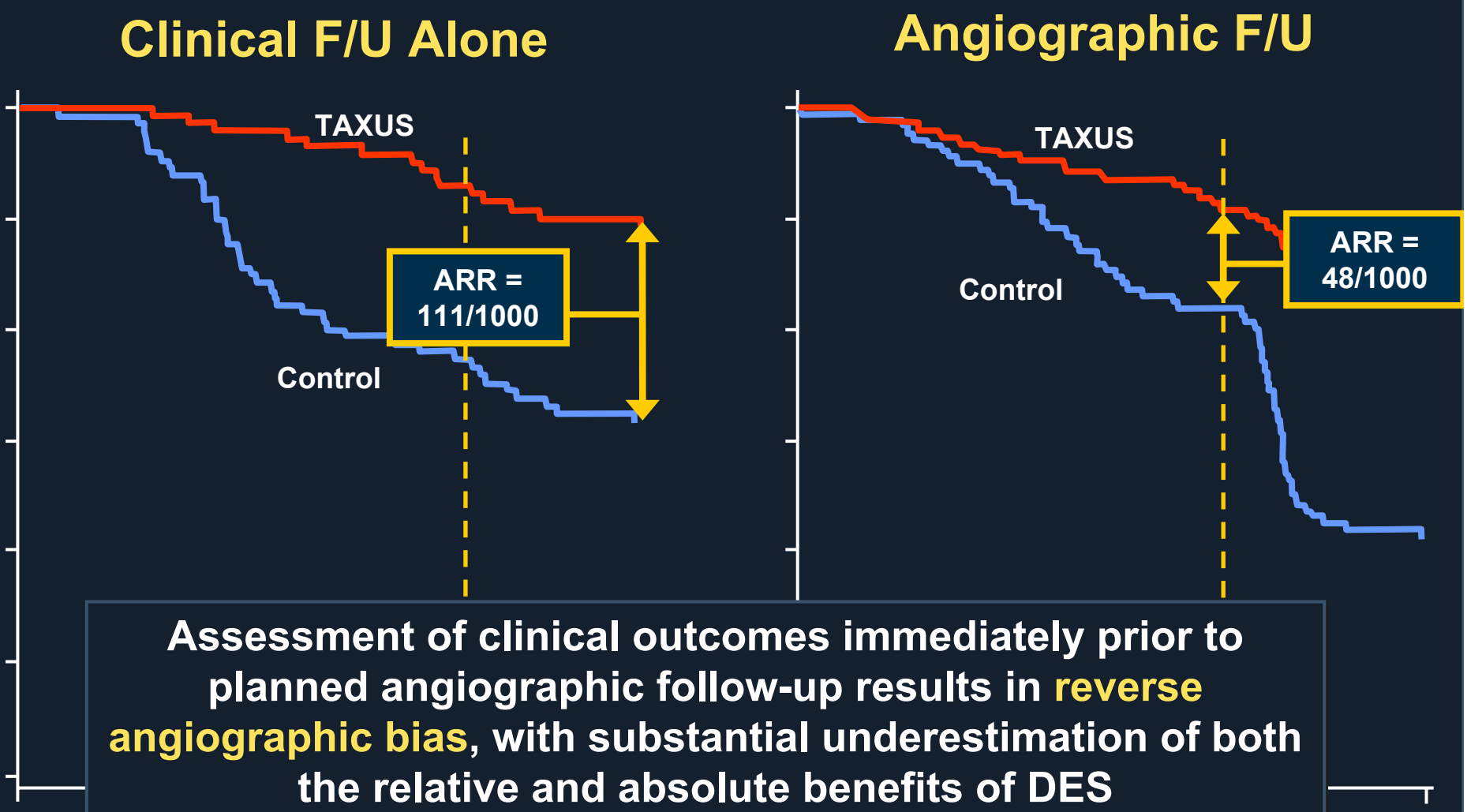
Pinto et al, *J Am Coll Cardiol* 2006; 48:32-36

Impact of Routine Angiography in TAXUS IV



Pinto et al, *J Am Coll Cardiol* 2006; 48:32-36

Impact of Routine Angiography in TAXUS IV



Assessment of clinical outcomes immediately prior to planned angiographic follow-up results in reverse angiographic bias, with substantial underestimation of both the relative and absolute benefits of DES

Pinto et al, *J Am Coll Cardiol* 2006; 48:32-36

Conclusions

- Statistics are very powerful tools, but like any tool, they can be misused
- Incomplete understanding and inappropriate uses of statistics can lead to faulty conclusions and mass hysteria (DES thrombosis)
- Always put the data in a clinical perspective
  - The combination of great clinical skills with a knowledge of statistical methodology (and limitations) is a formidable one