SCT

SOCIETY FOR CLINICAL TRIALS

Society for Clinical Trials 34th Annual Meeting

Workshop P7
Practical Statistical Reasoning in Clinical Trials for Non-Statisticians

Sunday, May 19, 2013

1:00 – 5:00 PM

Fairfax B Room

34th Annual Meeting
Boston, MA, USA

Practical Statistical Reasoning in Clinical Trials for Non-Statisticians

Michele Melia, ScM, Jaeb Center for Health Research
Paul Wakim, PhD, National Institute on Drug Abuse, NIH

Pre-Meeting Workshop P7
19 May 2013

Michele Melia, ScM

Senior Statistician
Jaeb Center for Health Research

- 25+ years experience with multicenter clinical trials in ophthalmology
- Senior statistician for 2 National Eye Institute-sponsored clinical research networks
    - Pediatric Eye Disease Investigator Group
    - Diabetic Retinopathy Clinical Research Network

Acknowledgment: Chris Coffey & Elizabeth Lazar

Paul Wakim, PhD

Senior Mathematical Statistician
National Institute on Drug Abuse, NIH

11+ years experience with multicenter clinical trials on treatments of substance use disorders in the Clinical Trials Network

11+ years experience in teaching Mathematics and Statistics to non-Math/Stat majors at both undergraduate and graduate levels

Outline

1. Introduction (Michele & Paul)
2. Trial Design (Michele)
3. Analysis Plan (Paul)
4. Trial Monitoring and Interim Analyses (Paul)
5. Primary Analysis (Michele)
6. Subgroup Analyses (Michele)
7. Publication of results (Paul)

References are listed at the end of these slides

2. Trial Design

What is the first step and most important part of trial design?

The primary research question

Formulating the primary research question

- Feasible
- Interesting
- Novel
- Ethical
- Relevant*

When the clinical trial is completed and the data analyzed, will the answer to the primary research question (regardless of positive or negative) advance scientific knowledge and/or clinical practice?

*Hulley et al. 2007

What is the next most important part of trial design?

Choosing a primary outcome

- Research question → what you want to show
- Primary outcome → how to show it

Choosing a primary outcome

- Rigorously defined
- Relevant to study goals
- Reproducible
- Assessable in all groups to be evaluated or compared
- Unbiased (minimize bias)
- Chosen in design phase (before data collection)
- Anticipates data analysis methods/needs
- Used to determine sample size
  - Different outcomes will require different sample sizes

Primary research question

- Example - COMET2
  - Does near correction for reading prevent myopia in school-aged children with accommodative lag and near esophoria?
- Using primary research question as the guide, then need to define:
  - Primary outcome
  - Interventions
  - Population to be studied
  - Time period

Primary research question - example

- Primary outcome - Incidence of myopia
  - Myopia (refractive correction) ≤-1.00 D
  - Interventions - None (control) vs near correction
- Population to be studied - School-aged children with accommodative lag and near esophoria
  - Ages 8 to <12 years
  - Accommodative lag ≥1.0 D
  - Near esophoria ≥2.0 Δ
- Time period - 3 years

Other basic trial design features

- Randomization design → determined by study question and scientific & practical considerations
  - Parallel group
  - Factorial
  - Crossover
  - Cluster
- Comparative type → dictated by study question
  - Superiority (Efficacy/Effectiveness)
  - Equivalence
  - Non-inferiority

Defining Randomization Design

Randomization Design	**Unit of Randomization**	**Treatment assignment**
Parallel group	Subject*	A or B
Crossover	Subject	A then B, or B then A
Cluster	Group of subjects, e.g. by site	A or B
Factorial	Subject	A or B *and* C or D (A&C, A&D, B&C, B&D)

Comparative Type

The comparative type of the trial is determined by the hypothesis that will be tested. Note that the alternative is what we hope to prove.

Type of Test	**Null Hypothesis**	**Alternative Hypothesis**
Efficacy/Effectiveness (frequently called ‘superiority’):		
2-sided	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$
1-sided	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 < 0$ (or $\mu_1 - \mu_2 > 0$)
Equivalence (2-sided)	$\mid \mu_1 - \mu_2 \mid \geq M$	$\mid \mu_1 - \mu_2 \mid < M$
Non-inferiority (1-sided)	$\mu_1 - \mu_2 \leq -M$	$\mu_1 - \mu_2 > -M$
Superiority (1-sided)	$\mu_1 - \mu_2 \leq +M$	$\mu_1 - \mu_2 > +M$

Efficacy/Effectiveness (Superiority) Hypothesis: Example

Does performing near activities while patching for amblyopia affect improvement in visual acuity in children age 3 to < 7 years (as compared to distance activities)?

Primary outcome: Treatment group difference in mean visual acuity at 8 weeks

- H_o: No treatment difference ($\mu_N - \mu_D = 0$)
- H_a: Treatments are different ($\mu_N - \mu_D \neq 0$)

Equivalence Hypothesis: Example

Do patching and atropine eye drops have similar effectiveness with respect to improvement in visual acuity for treating amblyopia in children 7 to <13 years old?

Primary outcome: Treatment group difference in mean visual acuity at 8 weeks

- H_o: Treatments not equivalent ($|\mu_P - \mu_A| \geq M$)
- H_a: Treatments are equivalent ($|\mu_P - \mu_A| < M$)

M = equivalence margin = 5 letters

Non-inferiority hypothesis: Example

Is Bangerter filter as good as patching with respect to improvement in visual acuity for treatment of amblyopia in children 3 to <10 years old?

Primary outcome: Treatment group difference in mean visual acuity at 8 weeks

- H_o: Bangerter is inferior (μ_B - $\mu_P \leq$ -M)
- H_a: Bangerter not inferior (μ_B - $\mu_P >$ -M)

M = non-inferiority margin = 4 letters

Significance Level (Type I error rate or α)

- Significance level (α) – Probability of erroneously rejecting the null hypothesis
  - Determined prior to initiating study
  - Frequently, $\alpha = 0.05$ (5% risk of erroneously rejecting the null hypothesis)
  - Choice of significance level may be more ($\alpha = 0.01$) or less conservative ($\alpha = 0.10$) based on study factors

Test Statistic

- Test statistic – A quantity computed from the data used to measure the plausibility of the alternative hypothesis relative to null hypothesis
  - E.g. t-score = (Sample Mean – Hypothesized Population Mean) / std(mean)

Does performing near activities while patching for amblyopia result in more improvement in visual acuity (VA) among children age 3 to < 7 yrs as compared to distance activities?

- H_o: difference in mean VA between groups = 0
- Mean VA difference (Near - Distance) = -0.03 Std = 0.16

t-score = (-0.03 – 0) / 0.16 ≈ -0.19

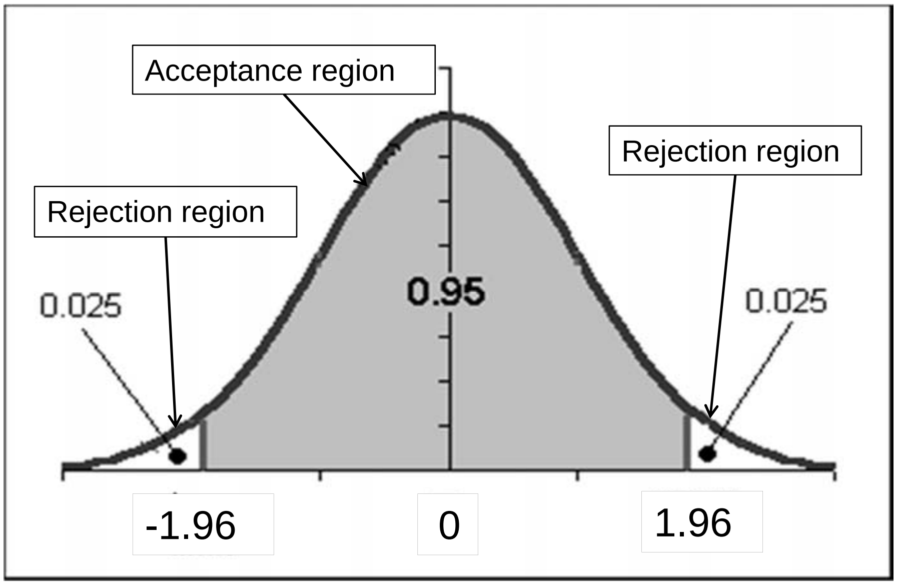
Acceptance and Rejection Regions

Acceptance Region: The range of test statistic values for which H_0 is not rejected

Rejection Region: The range of test statistic values for which H_0 is rejected

The test statistic must fall into one of these regions.

Evaluating the statistical significance of the test statistic



Rejecting the Null Hypothesis

- If the test statistic falls into the rejection region, the test is said to be statistically significant
- If we don’t reject H_0, we can’t claim to ‘accept (or prove) H_0’
  - Suppose one makes a statement ‘all swans are white’
  - To examine this statement, a sample of swans is drawn
  - Two things can happen:
    a) All swans in the sample are white
    b) At least one swan in the sample is not white
  - The event (b) establishes the falsehood of statement
  - However, the event (a) does not prove the statement!

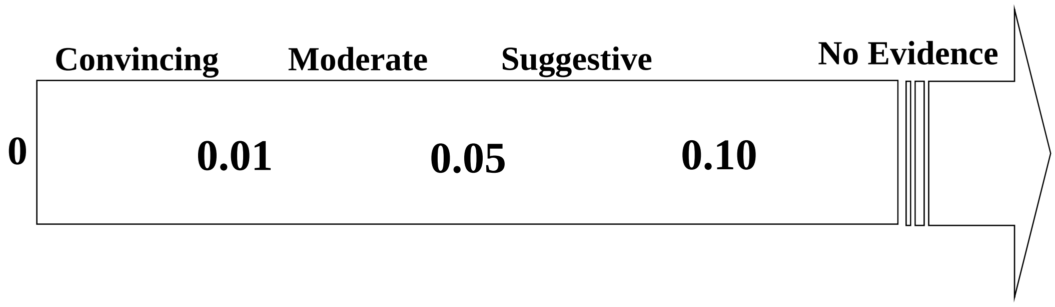
Acknowledgment to Chris Coffey

Quantify significance of results

- P-value: Measures the credibility of the null hypothesis
  - The probability of obtaining the observed test statistic or more extreme values if the null hypothesis is true
    - Small p-values suggest that observed results are not likely under the null hypothesis
  - Compare p-value from observed test statistic to the significance level (α)
    - If p-value < α → Reject Ho; otherwise fail to reject Ho

P-values and Hypothesis Testing

- Need to evaluate size of the p-value to judge strength of the evidence against null hypothesis
  - Degree of evidence may differ despite same conclusion (p=0.045 vs p=0.001)
  - Nearly identical p-values (p=0.051 vs p=0.049) may lead to different conclusions (α = 0.05)



Convincing Moderate Suggestive No Evidence

0 0.01 0.05 0.10

Quantify significance of results

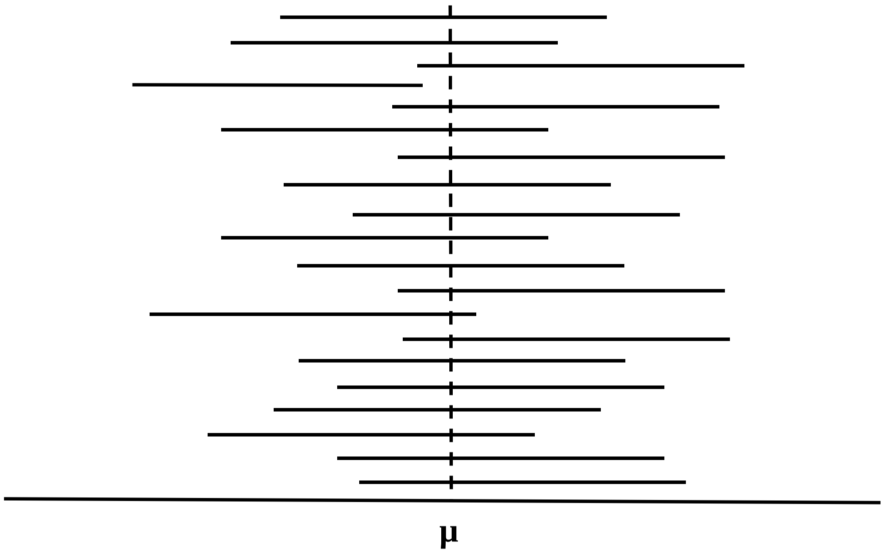
- P-value: Indicates whether results are *statistically* significant but no information on *clinical* significance

COMET2: Determine whether progressive-addition lenses (PALs) relative to single-vision lenses (SVLs) slow the progression of low myopia in children with high accommodative lag and near esophoria

Results: PALs were found to slow myopia progression by 0.28 D (over 3 years) compared to SVLs (2-sided P=0.04).

What is a 95% Confidence Interval (CI)?

A 95% CI is generated by a statistical procedure that captures the population parameter (μ) in 95% of its applications



Confidence Intervals

- Confidence Intervals:
  - Provides information about uncertainty in the estimate of the population parameter (Ex: Mean difference in VA) by including lower/upper bounds around the sample estimate
    - COMET2: 0.28 D (95% CI: 0.01 – 0.55 D)
  - Express how certain (confident) we are that the procedure used to generate the interval includes the population parameter
    - Increasing length of confidence interval (90% → 95%) improves likelihood of capturing the population parameter

Relationship with Hypothesis Testing

What is the relationship between confidence intervals and hypothesis testing?

- Decision to reject/fail to reject the null hypothesis depends on whether the confidence interval includes values consistent with the null hypothesis
  - If CI includes null hypothesis → Fail to reject
  - If CI excludes null hypothesis → Reject

Relationship with Hypothesis Testing

ATS6: Determine whether performing near activities while patching for amblyopia affects improvement in visual acuity among children age 3 to < 7 years

H_o: $\mu_N - \mu_D = 0$ H_a: $\mu_N - \mu_D \neq 0$

Construct 2-sided 95% CI on treatment difference in mean VA (logMAR lines) at 8 weeks

- Results: 0.0 (95% CI, -0.3 – 0.3)
- 95% CI includes 0, so we fail to reject null hypothesis (no difference between treatments) at α=0.05 level

Sample Size

- Now that we have defined:
  - study question
  - primary outcome
  - test hypothesis
  - randomization design
- It’s time to think about sample size

Basis of sample size determination

- In all clinical trials, we are selecting a sample from a target population
- The possibility exists that the sample we select will not be representative of the outcome rate or treatment effect
- Goal is to choose sample size to:
  - ensure high chances of getting the correct answer

  AND

  - enroll as few subjects as possible

What information is needed?

- Basic trial design features:
  - Comparative type (efficacy, equivalence, non-inferiority)
  - Randomization design (parallel group, crossover, cluster)
- Number of treatment groups
- Primary outcome & outcome rate or variance in controls
- Size of treatment difference to be detected
- Risk we are willing to take that study will “miss” a true treatment difference (β=type II error; 1-β=study power)
- Risk we are willing to take that study will erroneously conclude treatments are different (α=type I error)

Effect of basic trial design elements on sample size

Randomization Design	Relative Effect on Sample Size
Parallel Group	(Reference)
Crossover	Smaller
Cluster	Larger
Factorial	2 for price of 1*
Comparative Type	
Efficacy/effectiveness	(Reference)
Equivalence	Larger
Non-inferiority	Smaller or larger

*Assuming no treatment interaction.

2x2 Factorial Trial

Treatment A	Treatment B		Total
	Yes	**No**	
Yes	n	n	2n
No	n	n	2n
Total	2n	2n	4n

Assuming that effect of A is the same with or without B, and vice versa, this design permits testing of effect of A *and* effect of B with the same sample size required for testing either treatment alone.

This assumption is known as ‘no interaction’.

Outcome variable

- What is the expected proportion with the outcome (or variance of the outcome if continuous) in the control and treatment groups?
  - Continuous outcomes usually have smaller sample size than a proportion using the same measurement, but may be less clinically interpretable
  - E.g. ATS1: N for mean ΔVA outcome = 400; N for proportion ($\geq$20/32 or improved 3+ lines) $\approx$ 1000+
- Accurate estimate of outcome in control group is key

The smaller the treatment effect to be detected, the larger the required sample size

Defining type I and II errors Efficacy/Effectiveness Study

Study conclusion	**Truth**	
	Treatments not different ($\Delta=0$)	Treatments differ ($\Delta\neq0$)
Treatments not different ($\Delta=0$)	True negative	False negative Type II error (β)
Treatments differ ($\Delta\neq0$)	False positive Type I error (α)	True positive

Defining Type I and II Errors Generally

Study Conclusion	**Truth**	
	H_o true	**H_a true**
H_o true	True negative	False negative (Type II error, β)
H_a true	False positive (Type I error, α)	True positive

How to determine alpha (α) and beta (β)

Although α often is set at 0.05 and β at either 0.10 or 0.20, they should be study-specific

- Seriousness of disease and impact of treatment
- Public health importance of disease and treatment
- Availability of other treatments
- Cost of treatment

Sample size increases as α and β decrease

Example – ATS17 (Levodopa for Amblyopia)

- Type I error – conclude that levodopa is effective when in truth it is not effective
  - Effective treatment options for residual amblyopia are limited, so many children may receive levodopa
  - Children are unnecessarily exposed to risks of drug
  - Treatment costs are increased for no benefit
- Type II error – conclude that levodopa is not effective when in truth it is effective
  - Children with residual amblyopia will not receive an effective treatment
  - Other effective treatment options are limited

Example – Collaborative Ocular Melanoma Study Large Tumor Trial

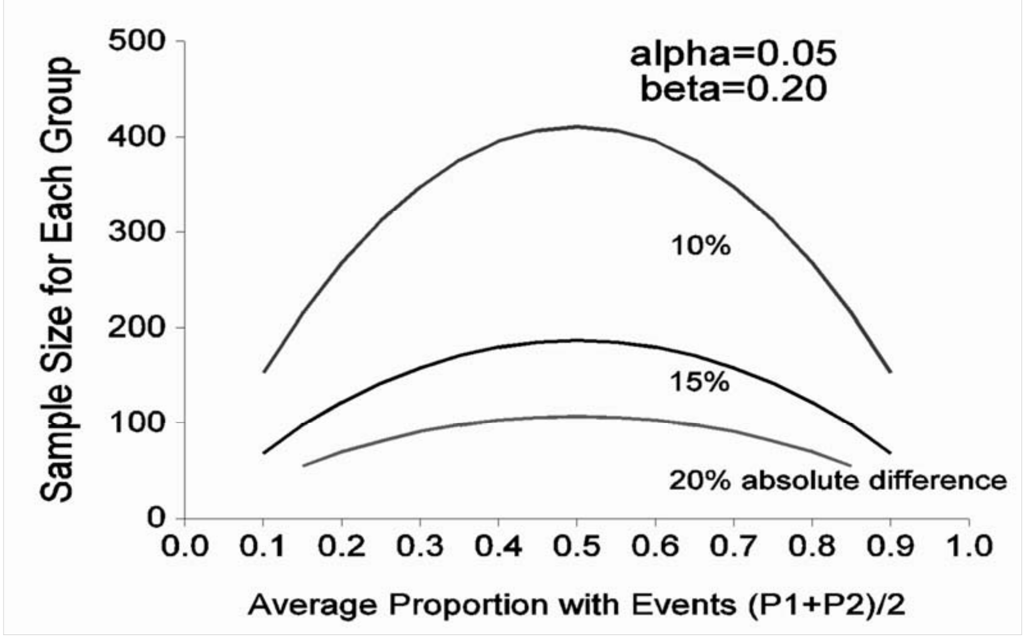
- Type I error – conclude that external beam radiation prior to enucleation improves 5 year survival when in fact it does not
  - Patients are unnecessarily exposed to radiation
  - Treatment costs are increased with no benefit
- Type II error – conclude that external beam radiation prior to enucleation does not improve 5 year survival when in fact it does
  - Patients do not receive an effective treatment for a highly fatal disease (5 year all-cause mortality≈40-50%)

Effect of Variance of Outcome on Sample Size

Larger variance➔larger sample size

- For a proportion, variance is function of P*(1-P)—largest for P=0.5
- P is the average of the control and treated group outcome proportions

Sample size according to difference in proportion of events between treatments



Adjustments to sample size estimate

- Losses to follow up
- Treatment group crossovers
- Poor treatment adherence
- Ineligible patients enrolled
- Misclassification of outcome

Presence of any of these increases sample size

Your statistician can adjust the sample size for these, BUT this does not affect the potential for bias

Unequal treatment group sizes

- Equal size groups are statistically optimal
  - Maximizes power for a given sample size
- Reasons to consider unequal group sizes:
  - More info is needed on effect of new treatment (e.g. adverse effects)
  - Subjects unwilling to be randomized if chance of control is too high
  - Reduce study cost when 1 treatment is more expensive

Number of treatment groups

- Increasing the number of treatment groups will increase the sample size
  - The per group sample size will be larger than that for the corresponding 2 group study
  - For example, if the 2 group study requires 50 subjects per group, a 3 group study will require more than 50 subjects per group
- Increase depends on # of specific group comparisons planned
  - 2 group study has 1 comparison
  - 3 group study has 2 or 3 comparisons → higher chance of type I error with larger # of comparisons; controlled by increasing sample size
- Increase depends on sizes of the detectable treatment effects

Effect of Number of Treatment Groups on Sample Size: Convergence Insufficiency Treatment Study

No. of groups	**2**	**3**	
No. of comparisons	1	2	3

List of comparisons	Pc=0.3 v Ps=0.1	Pc=0.3 v Ps=0.1
		Pc=0.3 v Pn=0.15
			Pc=0.3 v Ps=0.1
			Pc=0.3 v Pn=0.15
			Pn=0.15 v Ps=0.1

Sample size ratio 1:1 (1:1:1) 2:1 (2:2:1)	 180 204	 603* 505	 >2400 >2700

Pc=success proportion in computer group; Pn=near target pushup group; Ps=sham group; *Pc v Ps comparison has >99% power with 1:1:1 ratio.

Ways to decrease sample size (for dichotomous outcome)

- Increase magnitude of treatment effect to be detected
- Increase the number of events in control group (assuming # events is proportional between groups, e.g. 2:1)
  - Lengthen follow up
  - Change primary outcome or widen outcome criteria
  - Switch to a surrogate outcome
  - Limit enrollment to higher risk patients
- Minor: increase alpha or beta; change to one-sided

Ways to decrease sample size (for continuous outcome)

- Reduce variance of outcome measure
  - Change to more precise measurement method
  - Limit enrollment to patients with less variance
  - Use mean or median of multiple measurements

- Increase magnitude of treatment effect to be detected

- Minor: increase alpha or beta; change to one-sided hypothesis test

Sample size summary

- Sample size and other scientific demands usually must be balanced with practical limitations of available funds and number of eligible patients
- Finding a satisfactory balance frequently involves modifying aspects of the study design
- Given the close link between study design and sample size, it is advisable to evaluate sample size requirements as early as possible in the planning process

3. Analysis Plan

Statistics rarely offers a single “right” way of doing anything.

Wheelan 2013

Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

ITT vs. Per-Protocol Analysis

ITT

Intention-to-treat (or intent-to-treat)
= include in the analysis all participants who were randomized
= "once randomized, analyzed"

Per-protocol

Include in the analysis a select subgroup "as stated in the protocol", e.g. those who took at least 80% of their medicine, or those who attended at least 75% of psychotherapy sessions

Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

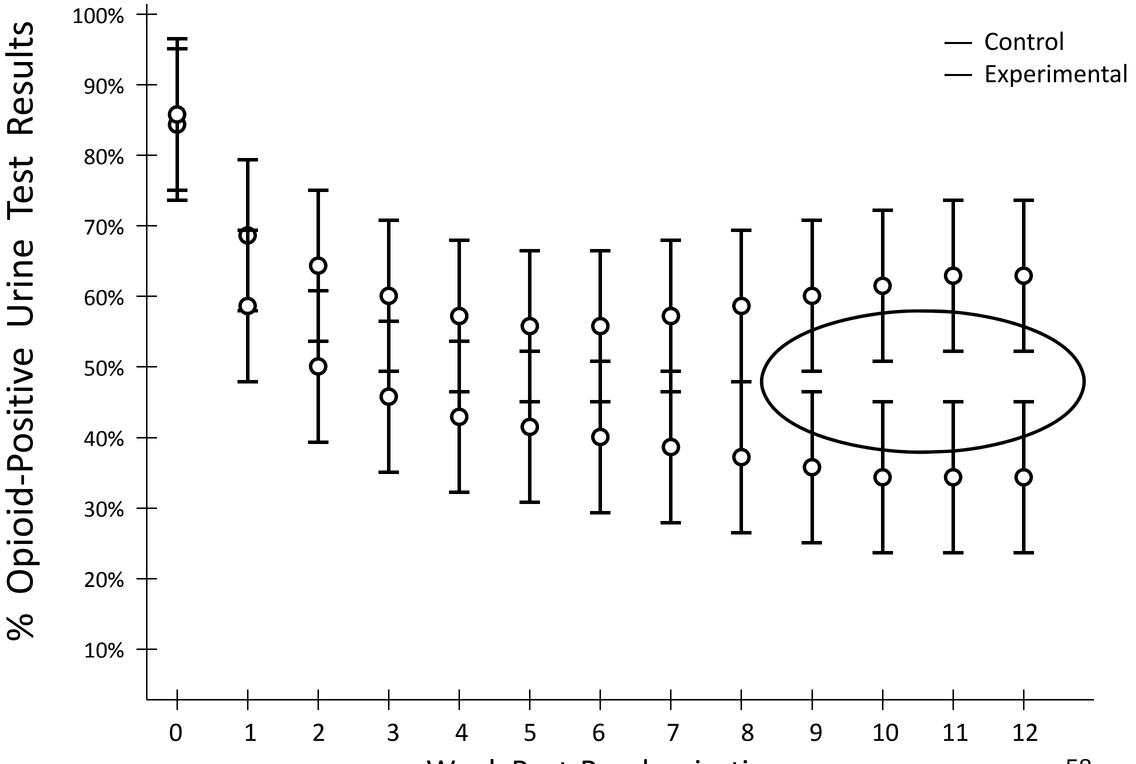
Statistical Test

- Which statistical test?
- At what time point? (primary outcome measure)

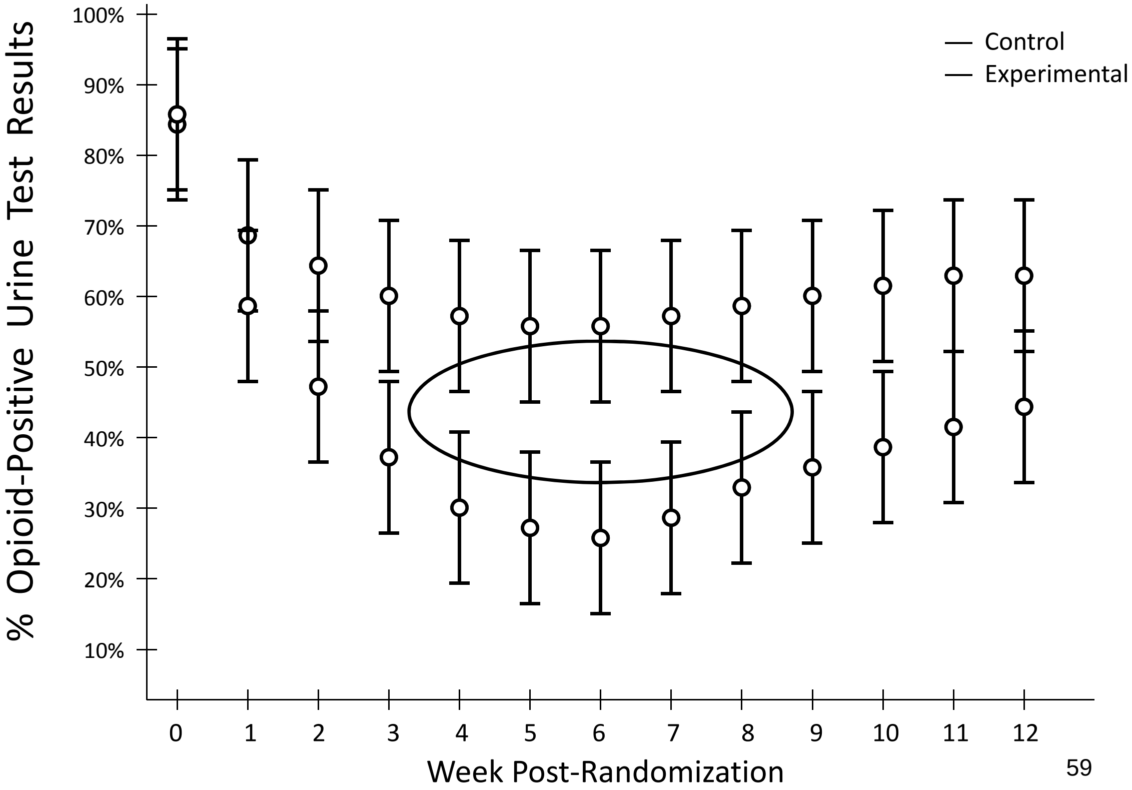
Statistical Model

- Simple vs. complex model
- Parameter in the model that corresponds to the primary hypothesis
- Stratification variables (FDA 1998, CPMP 2004)
- Using covariates (baseline vs. post-randomization)
- Site effect (random vs. fixed)
- Interactions (e.g. treatment-by-site)
- Longitudinal (repeated measures) model

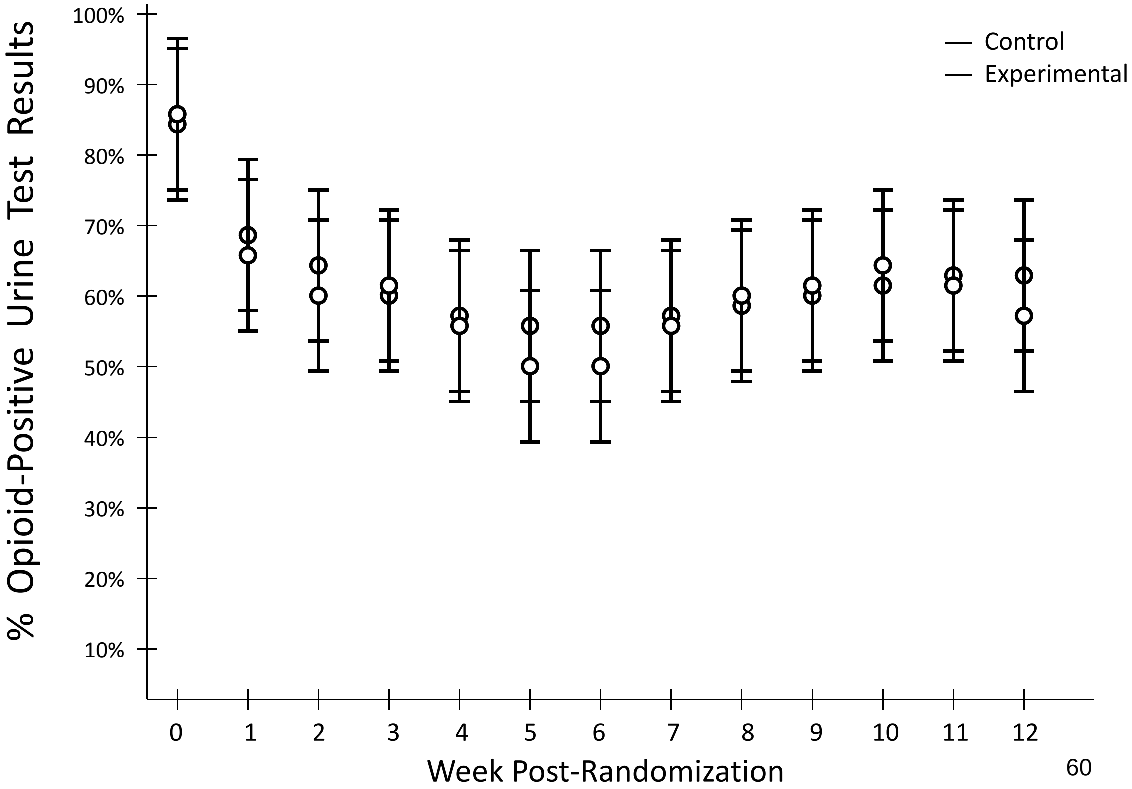
Statistical (Longitudinal) Model



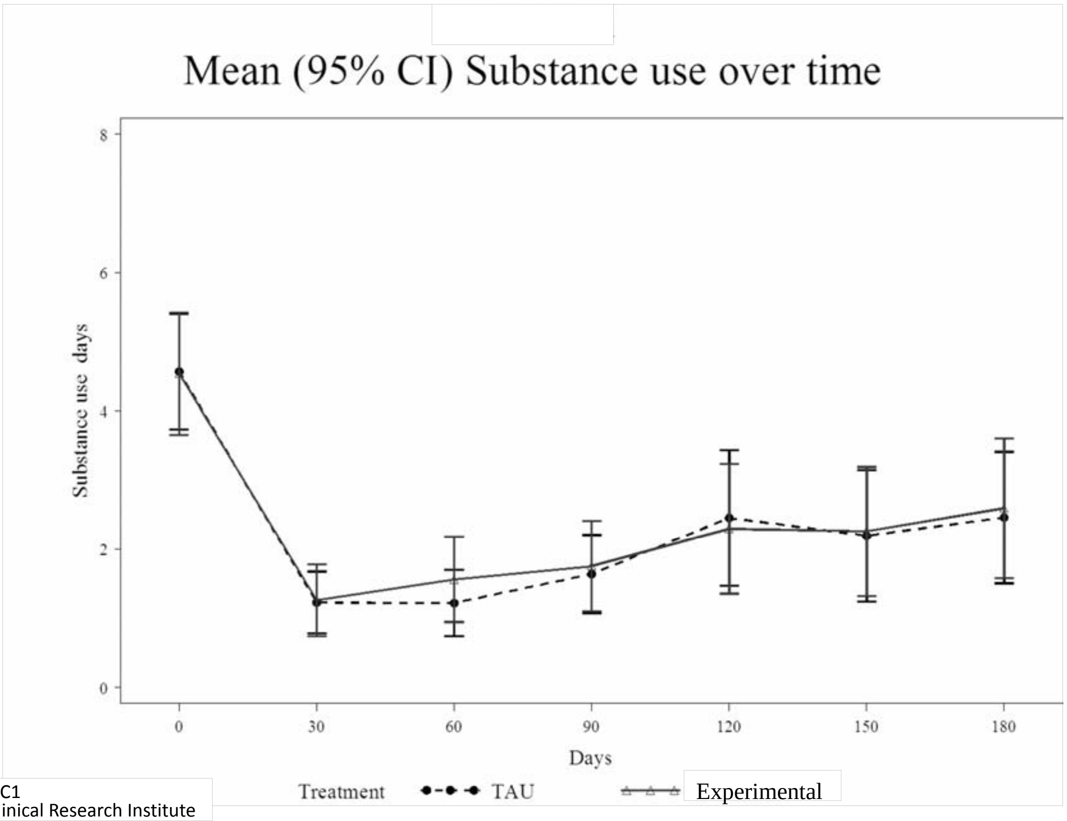
Statistical (Longitudinal) Model



Statistical (Longitudinal) Model



Statistical (Longitudinal) Model



Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

Multiplicity Adjustment

What is it?

When to do it?

Why do it?

What are the options?

When not to do it?

Multiplicity Adjustment: What is it?

Multiplicity adjustment is a way to control for false positive conclusions, i.e. to control the "family-wise" or "study-wise" rate of false positive conclusions

Multiplicity Adjustment: When to do it?

Formally, whenever there is more than one primary endpoint (or primary hypothesis), more than two treatment conditions, more than one dose vs. placebo, or more than one time point

Informally, whenever there is more than one secondary analysis, including subgroup analyses

Multiplicity Adjustment for Primary Analyses – an example of when it is needed –

From the Abstract:

Purpose/Objectives: *To test the effectiveness of two interventions compared to usual care in ①decreasing attitudinal barriers to cancer pain management, ②decreasing pain intensity, and ③improving functional status and ④quality of life (QOL).*

Thomas et al. 2012

Multiplicity Adjustment: Why do it?

If you calculate many P values, some are likely to be small just by random chance. Therefore, it is impossible to interpret small P values without knowing how many comparisons were made.
… It is easy to be fooled by these small P values.

Motulsky 2010

Multiplicity Adjustment: Why do it?

Recall:

Alpha = Type I error = chance of finding a statistically significant result when the null hypothesis is true (e.g. no difference)

Alpha = 0.05 is most commonly used

Multiplicity Adjustment: Why do it?

Example:

Two endpoints: drug use and retention

Each endpoint is tested at the 5% alpha level

The experimental treatment is considered beneficial if *either or both* endpoints are significant

Without multiplicity adjustment, the chance of the treatment being found beneficial *when it is not* can be as high as 10% (not 5%)

Multiplicity Adjustment: What are the options?

1) Basic Procedures
2) Stepwise Procedure (*pre-specified* testing sequence)
3) Stepwise Procedure (*data-driven* testing sequence)
4) Other more complicated methods

Dmitrienko 2011

Multiplicity Adjustment: What are the options?

1) **Basic Procedures, e.g. Bonferroni:**

P-values are compared to a pre-specified *fraction* of the alpha level (0.05).

Example: 3 tests → new alpha = 0.05/3 = 0.017

Pros: simple

Cons: least powerful (most conservative)

Dmitrienko 2011

Multiplicity Adjustment: What are the options?

2) **Stepwise Procedure (pre-specified testing sequence) e.g. fixed-sequence procedure.**

Hypotheses are ordered a priori, typically reflecting clinical importance

Testing begins with the first hypothesis, and each test is carried out without a multiplicity adjustment as long as significant results are observed in all preceding tests, i.e. the testing stops when the first non-significant result is observed

Dmitrienko 2011 & Dmitrienko 2009

Multiplicity Adjustment: What are the options?

3) Stepwise Procedure (data-driven testing sequence), e.g. Holm, Hochberg & Hommel

Start with the lowest p-value (Holm) or highest p-value (Hochberg & Hommel) and follow a sequence of steps

Hommel's is more powerful than Hochberg's, which is more powerful than Holm's

Dmitrienko 2011

Multiplicity Adjustment: When *not* to do it (i.e. not necessary)?

When *all* the primary endpoints have to be statistically significant in order to claim treatment benefit, e.g. to get FDA approval

EMEA/CPMP 2002

Example: the experimental treatment is considered beneficial only if *both* drug use *and* retention are found to be statistically significant

Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

Reference

The Prevention and Treatment of Missing Data in Clinical Trials
Panel on Handling Missing Data in Clinical Trials
National Research Council of the National Academies
July 2010

8 recommendations on minimizing missing data

12 recommendations on statistical approaches

Extent of the Issue (in the CTN)

Based on the first 24 multi-site clinical trials on substance abuse conducted between 2001 and 2010 in NIDA's Clinical Trials Network (CTN), the percent of missing data for the *primary outcome measure* ranged from 2% to 60% (median=25%).

Wakim et al. 2011

What's the big deal?

We need N=450 (based on power analysis)

And we expect 25% missing

So we set the initial N=600

So that the final (analyzed) N=450

Technical terms that we can’t escape…

Missing at random (MAR)

Missing completely at random (MCAR)

Missing not at random (MNAR)

Ignorable

Non-ignorable

… but what do they mean?

Missing Completely at Random (MCAR)

(Non-technical) Definition:
The fact that Y is missing has nothing to do with its unobserved *value*, or with other measured variables

Therefore:
The set of participants with complete data can be regarded as a simple random (or representative) sample of all participants

What to do?
Ignore the missing data and analyze the available data (“complete case” or “pairwise deletion” method)

Missing at Random (MAR)

(Non-technical) Definition:
The fact that Y is missing *can* be explained by other values of Y, or by other measured variables

Therefore:
The observed data can be used to account for the missing data

What to do?
Use Maximum Likelihood or Multiple Imputation approach, and include in the model the other measured variables that explain missingness

Missing Not at Random (MNAR)

(Non-technical) Definition:
The fact that Y is missing *cannot* be explained by other values of Y, or by other measured variables

Therefore:
The observed data cannot be used to account for the missing data; and outside information is needed

In simple English:
We have a problem – need more sophisticated and novel methods

In Summary…

	Missingness (i.e. whether the data are missing or not)	
	is related to	is not related to
MCAR		observed or unobserved data
MAR	observed data	unobserved data
MNAR	unobserved data	

Based on Graham 2009

Bottom Line

MCAR: No big deal

MAR: Use available collected data to “explain” missing mechanism, and use existing statistical methods

MNAR: Need outside information to “explain” missing mechanism

Ignorable & Non-Ignorable

(roughly speaking)

Ignorable (available data is sufficient):

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)

Non-Ignorable (need outside information):

- Missing Not At Random (MNAR)

Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

Handling of Outliers

What is an outlier?

How do outliers arise?

How are outliers identified?

Is it legitimate to remove outliers?

How should outlier removal be reported?

Motulsky 2010

What is an outlier?

An outlier is a value that is so far from the others that it appears to have come from a different population.

The presence of outliers can invalidate many statistical analyses.

Motulsky 2010

How do outliers arise?

Incorrect value

- Invalid data entry
- Experimental mistakes

Correct value

- Biological diversity
- Random chance
- Wrong assumption

Motulsky 2010

How are outliers identified?

- Statistical tests
- Single vs. multiple outliers
- Ultimately a subjective exercise

Motulsky 2010

Is it legitimate to remove outliers?

When is it “cheating” and when is it the responsible thing to do?

It’s all about pre-specification and disclosure

How should outlier removal be reported?

- Keep the outlying observations in the database, with a flag
- Show a graph with *all* values, and the outliers identified/marked
- Report how many outliers were excluded from the primary analysis, and the criteria used to identify the outliers
- Consider reporting the results in two ways: with and without the outliers

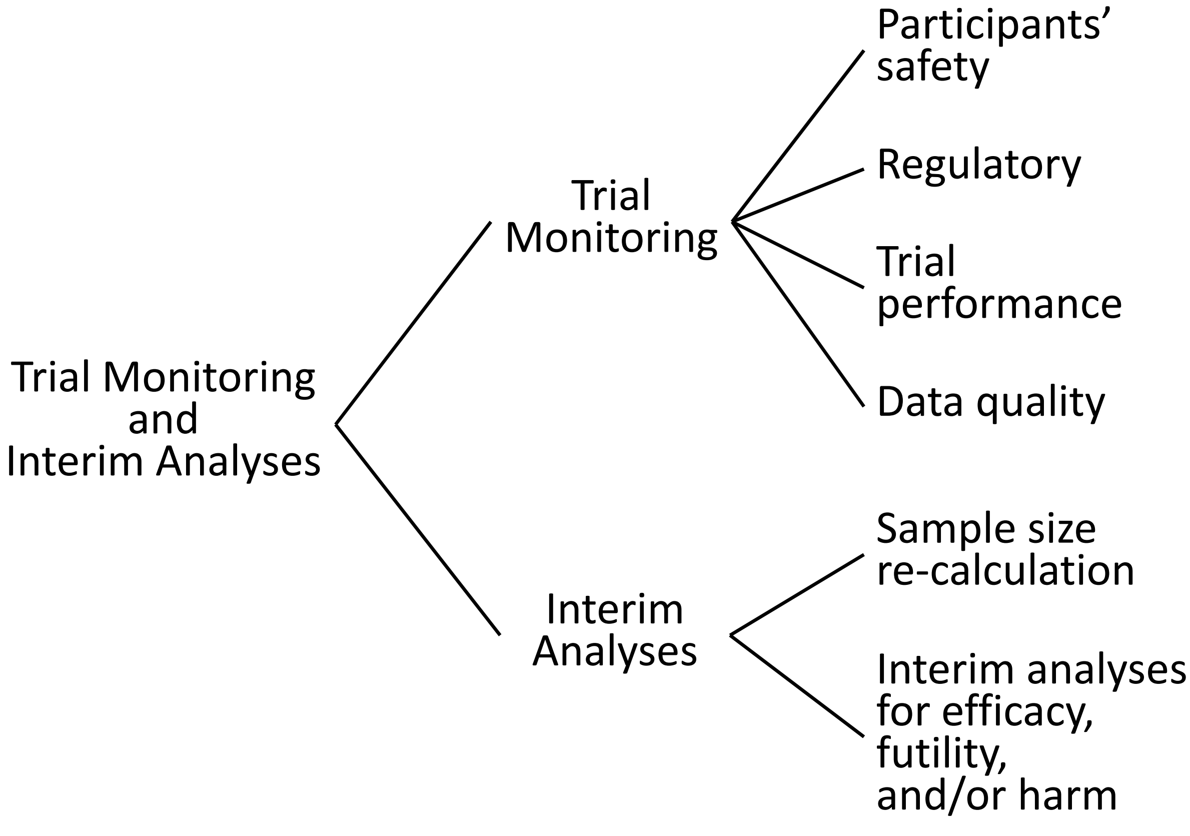
Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

Main Components of the Analysis Plan

1) ITT vs. per-protocol analysis
2) Statistical test or model
3) Multiplicity adjustment
4) Handling of missing data
5) Handling of outliers
6) Interim analyses
7) Sensitivity analysis
8) Secondary and subgroup analyses

4. Trial Monitoring and Interim Analyses



Trial Monitoring and Interim Analyses

Trial Monitoring

- Participants' safety
- Regulatory
- Trial performance
- Data quality

Interim Analyses

- Sample size re-calculation
- Interim analyses for efficacy, futility, and/or harm

Why are trial monitoring and interim analyses important?

1) Participants’ safety and well-being
2) Trial integrity
3) Optimal use of resources
4) Ethical considerations

Trial Monitoring

What to monitor?

1) Adverse events (AEs) and Serious Adverse Events (SAEs)
2) Regulatory compliance
3) Recruitment
4) Availability of primary outcome
5) Treatment exposure
6) Retention (follow-up visits)
7) Data quality

Interim Analyses

4 Main Points About Interim Analysis

1. It is a statistical analysis of the response variables performed while the trial is proceeding.

2. It is used to decide whether the study has come to an *early conclusion* without the need to either randomize unnecessarily additional participants, or expose them senselessly to a therapy that is proving to be inferior.

Based on Motulsky (2010), Friedman et al. (2010) & Moyé (2006)

4 Main Points About Interim Analysis

3. Because repeated examination of accumulating data increases the probability of declaring a treatment difference even if there is none, statistical adjustments have to be made.

4. None of the statistical techniques available for interim analyses should be used as the sole basis in the decision to stop or continue the trial.

Based on Proschan et al. (2006) & Friedman et al. (2010)

Possible reasons for terminating a trial earlier than scheduled

1) Serious adverse effects
2) Greater than expected beneficial effect
3) Improbable statistically significant difference by the end of the trial
4) Severe uncorrectable logistical, data quality or recruitment problems
5) Primary research question answered elsewhere or no longer sufficiently important

Friedman et al. 2010

Interim Analyses

- Sample size re-calculation (or re-estimation)
- Interim analyses for efficacy, futility and/or harm

Sample Size Re-Calculation

- Based on nuisance parameters *only* (no statistical penalty)
- Based on nuisance parameters *and* observed treatment effect (statistical penalty)

Proschan et al. 2006

Sample Size Re-Calculation Based on Nuisance Parameters *Only*

Are the values of variances, correlations, drop-out rate, or proportion of events in the control group, that we assumed at the beginning of the trial consistent with what we actually see so far?

And consequently, is the sample size we calculated initially still adequate based on these values?

Sample Size Re-Calculation Based on Nuisance Parameters *Only*

Result	**Decision**
Current N is adequate	Keep N the same
N should be higher	Increase N
Lower N is adequate	Keep N the same or decrease N?

Sample Size Re-Calculation Based on Nuisance Parameters *Only*

Result	**Decision**
Lower N is adequate	Keep N the same

Pros:

- Insures adequate power for primary analysis (just in case)
- Helps in interaction and safety analyses
- Helps in secondary and sub-group analyses

Cons:

- May unnecessarily subject participants to risk
- May waste resources that could be spent on other research
- May unnecessarily delay publishing important results

Sample Size Re-Calculation Based on Nuisance Parameters *Only*

Result	Decision
Lower N is adequate	Decrease N

Pros:

- Ends the trial and publishes results sooner
- Saves resources

Cons:

- Not enough power for primary analysis (just in case)
- Less data for interaction and safety analyses
- Less data for secondary and sub-group analyses

Sample Size Re-Calculation Based on Nuisance Parameters *and* Observed Treatment Effect

Should the sample size be changed based on the values of the nuisance parameters *and* the treatment effect observed so far?

This is controversial. Criticism has been about potential bias, loss of efficiency, and the possibility of increasing the sample size to detect clinically meaningless differences.

Proschan et al. (2006) & Proschan (2009)

Interim Analyses for Efficacy, Futility and/or Harm

(statistical penalty)

What's the general question?

Based on the data observed so far, is the experimental treatment:

- clearly beneficial (better than control); or
- clearly futile with no hope of efficacy; or
- clearly inferior (worse than control)?

If so, may stop the trial for ethical reasons and to save resources.

Interim Analyses for Efficacy, Futility and/or Harm

(statistical penalty)

Sequential designs (aka group sequential tests or repeated significance tests):

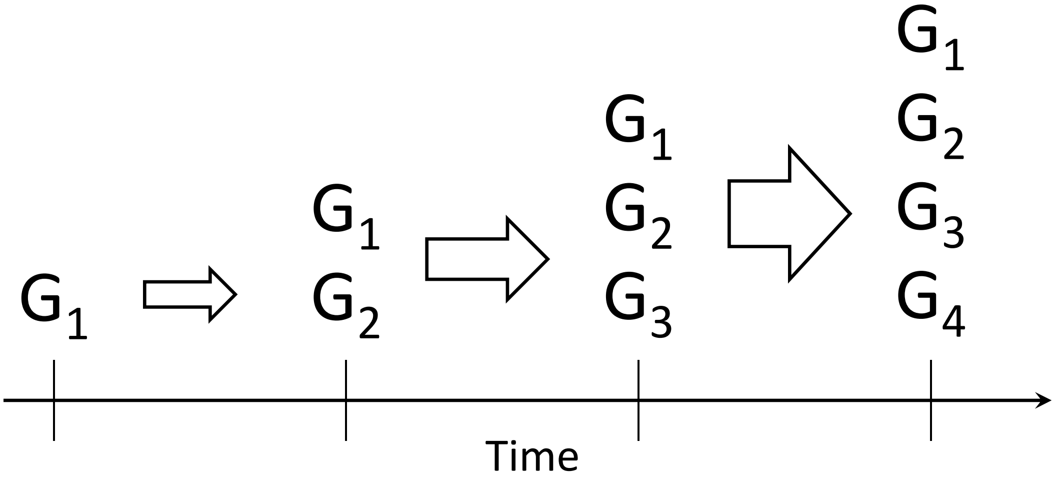
- Group sequential methods
- Flexible group sequential (alpha-spending) methods

Stochastic curtailment tests:

- Conditional power tests (frequentist)
- Predictive power tests (mixed Bayesian-frequentist)
- Predictive probability tests (fully Bayesian)

Dmitrienko et al. (2005)

Group Sequential Methods



Moyé 2006

Group Sequential Methods

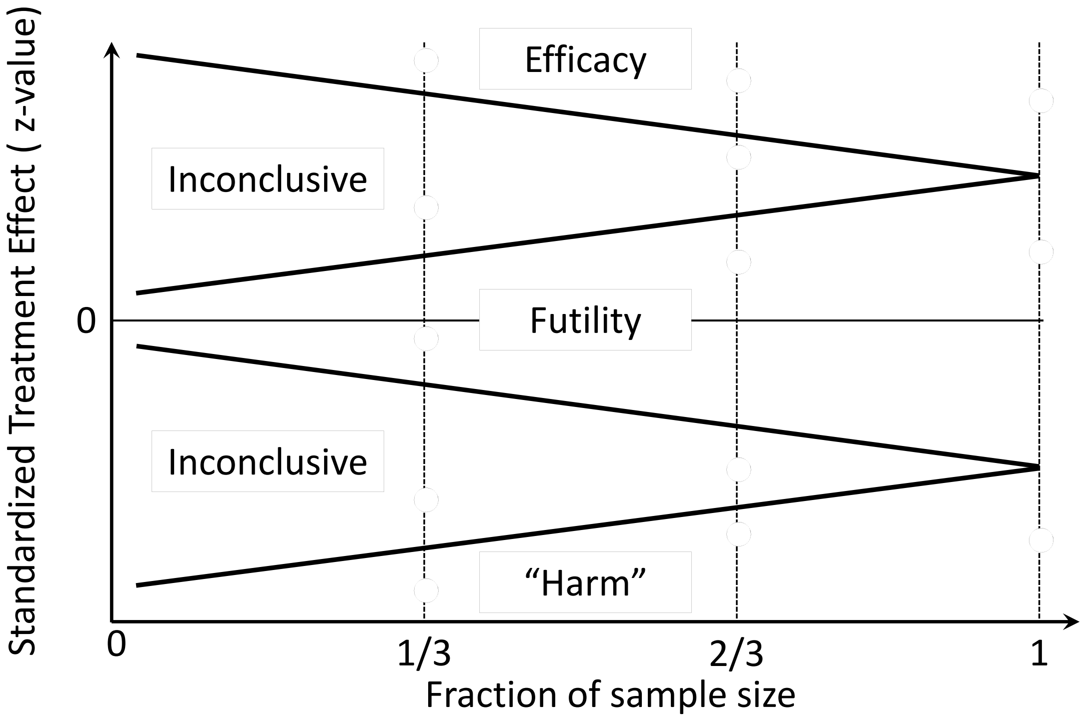
Group sequential procedures are simply processes that analyze groups of patients sequentially.
… each group’s data is added to the data that has been collected and is already available from the previous groups.

Moyé 2006

Group sequential design enables early trial stopping if there is harm, suggestion of futility, or overwhelming evidence of efficacy.

Zhu et al. 2011

Group Sequential Method
Two-Sided Stopping Boundaries



Based on Jennison & Turnbull (2000) and CTN DSC1-Duke Clinical Research Institute

Flexible Group Sequential (Alpha-Spending) Methods

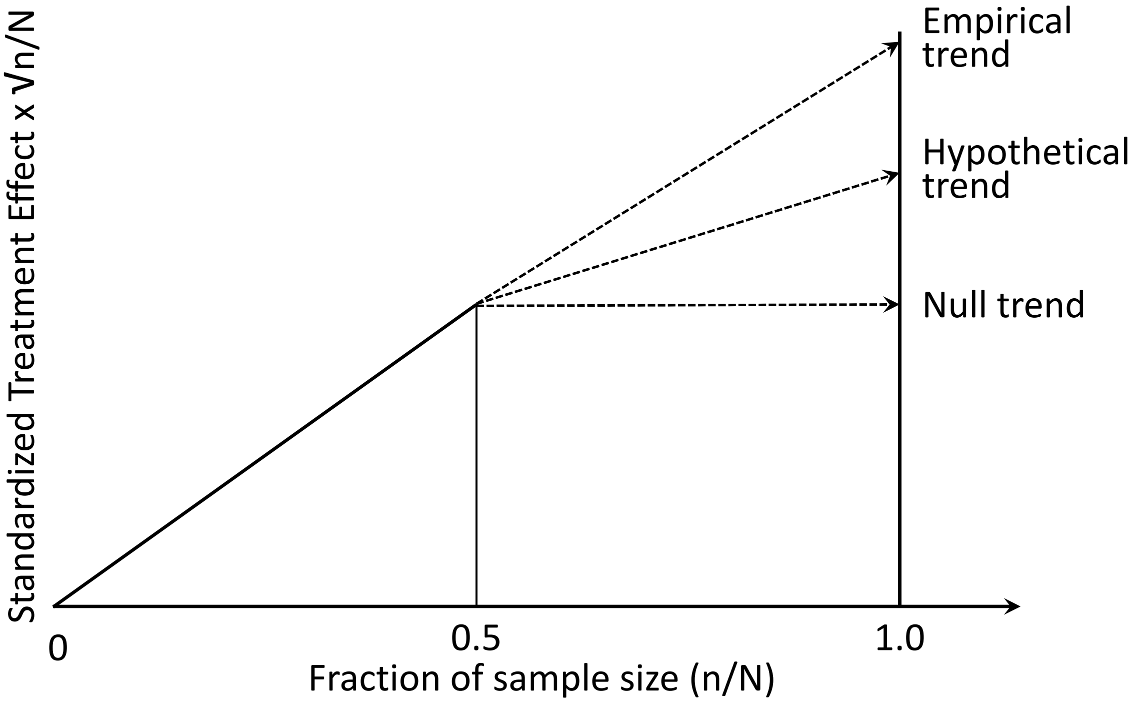
(e.g. Lan-DeMets method)

Same as group sequential methods, but without pre-specifying the number or spacing of interim looks.

- Allow for unplanned and unequally-spaced interim looks
- Provide flexibility on how to "spend" the Type I error (or alpha) during the course of the trial
- Guarantee that at the end of the trial, the overall Type I error will be the pre-specified value of alpha

Based on Friedman et al. (2010), Dmitrienko et al. (2005) & Zhu et al. (2011)

Stochastic Curtailment Tests


Standardized Treatment Effect x √n/N
Empirical trend
Hypothetical trend
Null trend
0
0.5
1.0
Fraction of sample size (n/N)

Based on Lan & Wittes (1988)

Conditional Power Tests

(frequentist approach)

Conditional power (CP) is the probability that the final study result will be statistically significant, given the data observed thus far
and a specific assumption about the pattern of the data to be observed in the remainder of the study, such as assuming the original design effect,
or the effect estimated from the current data,
or under the null hypothesis.

Lachin (2005)

Predictive Power Tests

(mixed Bayesian-frequentist approach)

They average the conditional power over the posterior distribution of the treatment effect, which is itself based on its prior distribution and the data observed so far.

Based on Dmitrienko et al. (2005)

Predictive Probability Tests

(Bayesian approach)

They are completely based on the posterior probability of a clinically important treatment effect (rather than statistical significance) given the already observed data.

Based on Dmitrienko et al. (2005)

One Cautionary Note

When performing a sample size re-calculation based on nuisance parameters only, without performing an interim analysis on futility, one may increase the sample size and extend the trial when in fact, an interim analysis would have revealed futility.
In other words, spend more money testing a futile treatment.

Another Cautionary Note

Because the decision to stop the trial may arise from catching the treatment effect at a random high, truncated RCTs (tRCTs) may overestimate the true treatment effect.

Briel et al. (2009)

Truncated RCTs were associated with greater effect sizes than RCTs not stopped early.

Bassler et al. (2010)

The Importance of Timing

Example

N=200
Interim analysis at 50% of sample size
Primary outcome assessed at 3 months post-rand.

The interim analysis is performed:

- NOT when 100 participants are randomized
- NOT when 100 participants have a non-missing primary outcome
- BUT when 100 participants have reached, or should have reached the 3-month time point

The Importance of Timing

Logistically:

- Decision needs to be made before the end of recruitment

Statistically:

- Too early: the results may not be robust enough
- Too late: recruitment may be completed

4. Primary Analysis

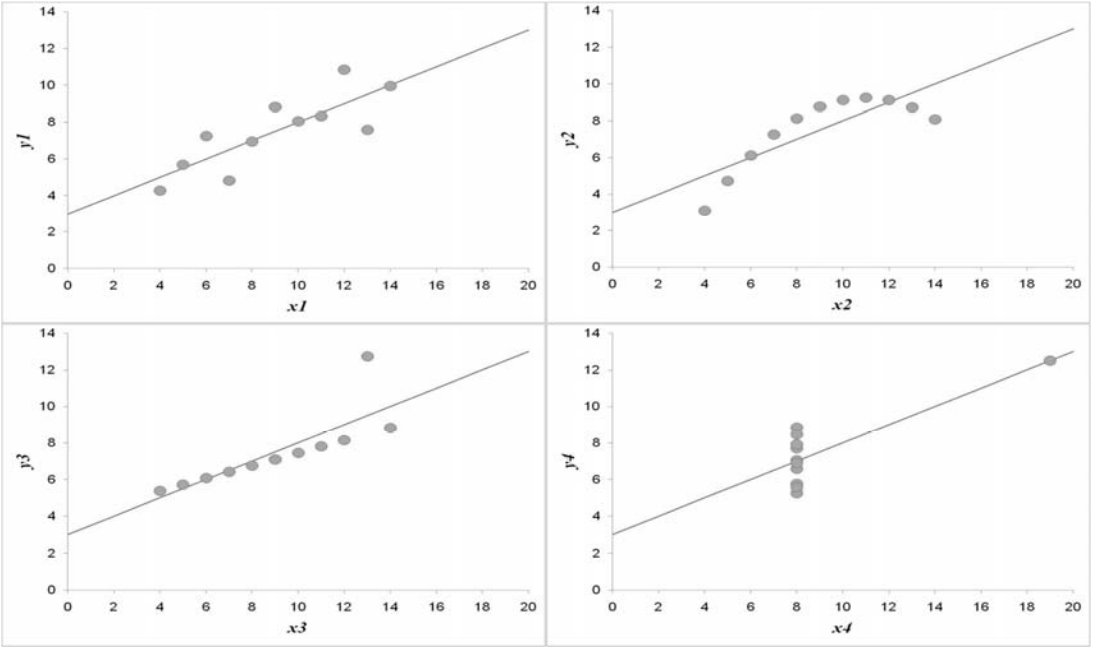
Anscombe’s Quartet

N=11 (*x*,*y*) data points produce the following statistical results:

Property	Value
Average (SD) of x variable	9.0 (3.16)
Average (SD) of y variable	7.5 (1.94)
Correlation between x and y	0.816
Regression line	y = 3 + 0.5x

From Wikipedia

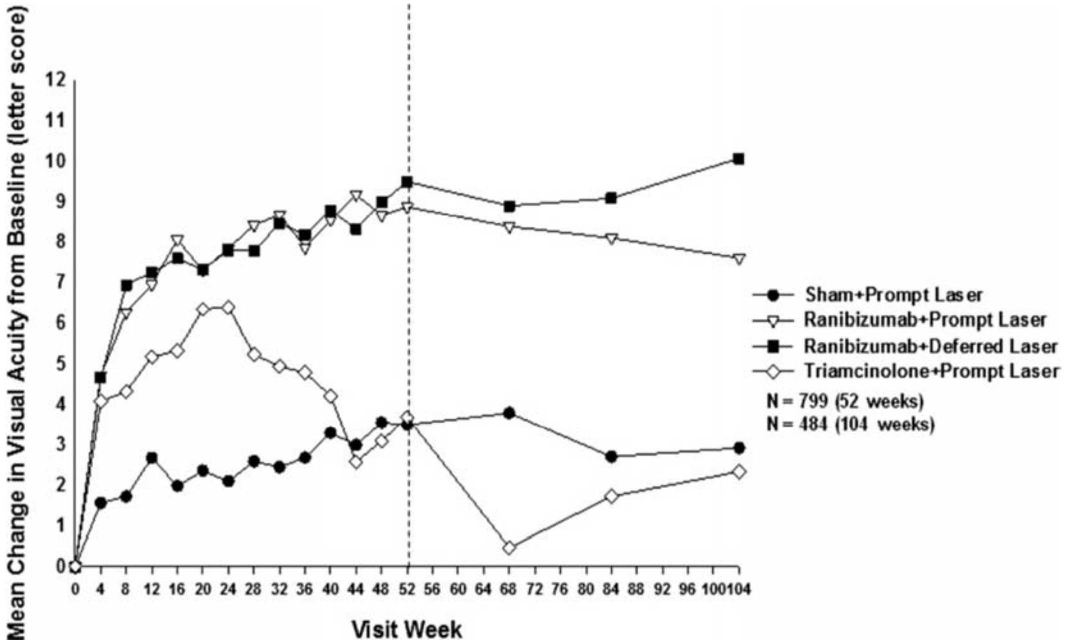
Anscombe’s Quartet



Principles for Primary Data Analysis

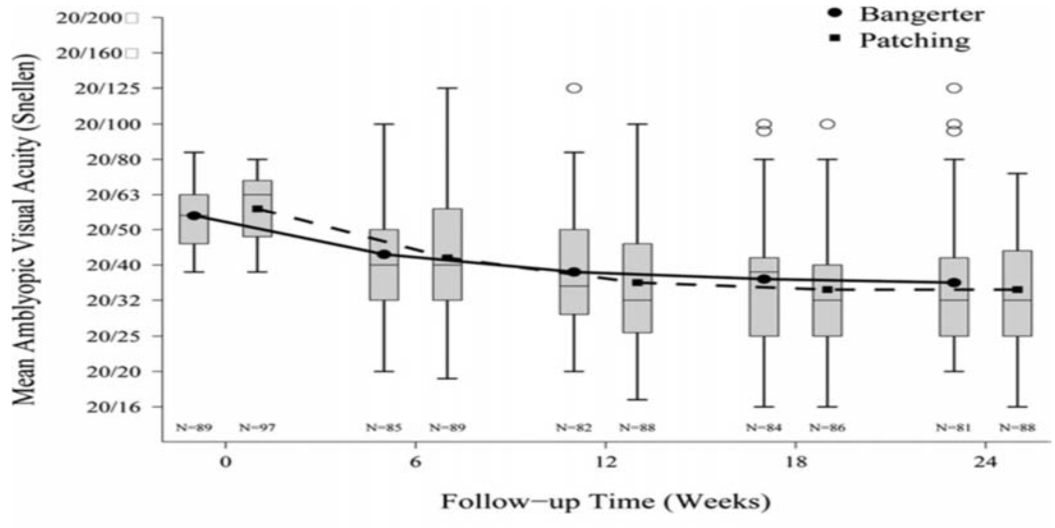
1. There is no substitute for a descriptive plot of the data

Visual Acuity with Anti-VEGF+Laser, Steroid+Laser, or Laser Alone for Diabetic Macular Edema (LRT for DME)



Elman et al. *Ophthalmology* 2010; 117:1064-77.

Bangerter Filter versus Patching for Moderate Amblyopia



Mean Treatment Group Improvement in Amblyopic Visual Acuity at Follow−up Visits (Lines)

Treatment Group:	6−wk Visit	12−wk Visit	18−wk Visit	24−wk Visit
Bangerter	1.1 Lines	1.6 Lines	1.8 Lines	1.9 Lines
Patching	1.4 Lines	2.1 Lines	2.3 Lines	2.3 Lines

Rutstein et al. *Ophthalmology* 2010; 998-1004.

Principles for Primary Data Analysis

1. There is no substitute for a descriptive plot of the data
2. The possible effects of chance on the observed data (treatment difference) must be quantified
   - This is the goal of the statistical analysis

Example – LRT for DME

Change in Visual Acuity - 1 Year (letters)	**Sham + Prompt Laser**	**Anti-VEGF + Prompt Laser**	**Anti-VEGF + Deferred Laser**	**Steroid + Prompt Laser**
Mean ± SD	+3 ± 13	+9 ± 11	+9 ± 12	+4 ± 13
Median	+5	+10	+9	+5
Difference versus sham +laser (95% CI)		+5.8 (+3.2 to +8.5)	+6.0 (+3.4 to +8.6)	+1.1 (-1.5 to +3.7)
P-value*		<0.001	<0.001	0.31

*From analysis of covariance adjusted for baseline visual acuity and correlation between study eyes.

Interpretation

- There is less than a 1/1000 chance that the observed difference (or a difference more extreme) between the anti-VEGF groups and the laser group would have occurred if anti-VEGF were not different from laser
  - Observed results very unlikely due to chance
  - Anti-VEGF (with prompt or deferred laser) is better than laser alone
  - Likely difference is about 6 letters, but could be as large as ~9 letters or as small as ~3 letters

Interpretation

- There is a 31/100 chance that the observed difference (or a difference more extreme) between steroid+laser and laser alone could have occurred if there were no difference between steroid+laser and laser alone
  - Observed difference could be due to chance
  - Can we conclude that steroid+laser and laser alone are not different? NO - but is it unlikely the difference is larger than -2 or +4
  - Is 0.31 the likelihood the results are due to chance? NO
  - 0.31 is the probability of getting 1.1 letter or larger difference given there is truly no difference

Principles for Primary Data Analysis

1. There is no substitute for a descriptive plot of the data
2. The possible effects of chance on the observed data (treatment difference) must be quantified
3. Use intention-to-treat

Intention to Treat (ITT)

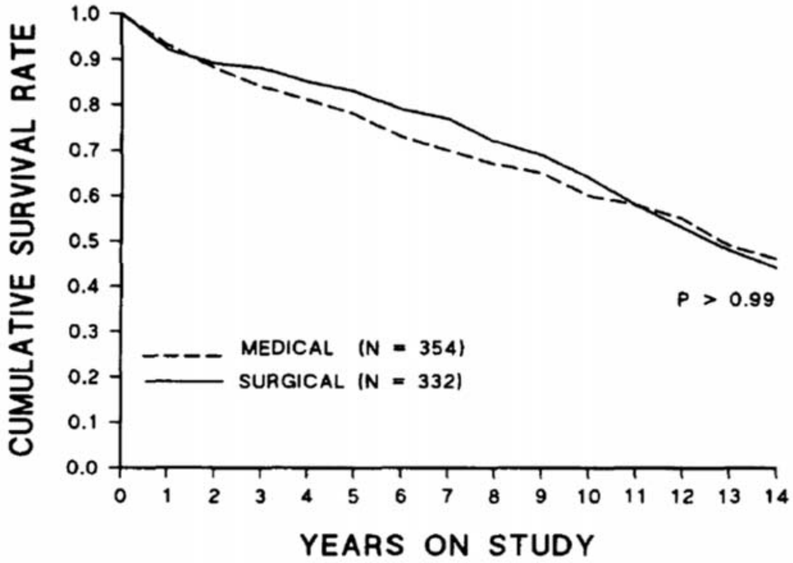
- ➢ Patients should be included in the group to which they were randomized for analysis, regardless of the treatment actually received
  - – Failure to adhere to or complete the assigned treatment is often due to side effects, perceived lack of efficacy, disease progression, i.e., it is at least partly an outcome of the assigned treatment
  - – Failure to attribute these outcomes to the assigned treatment can introduce bias into the treatment comparison

Example – Veterans Administration Cooperative Study of Coronary Artery Bypass Surgery*

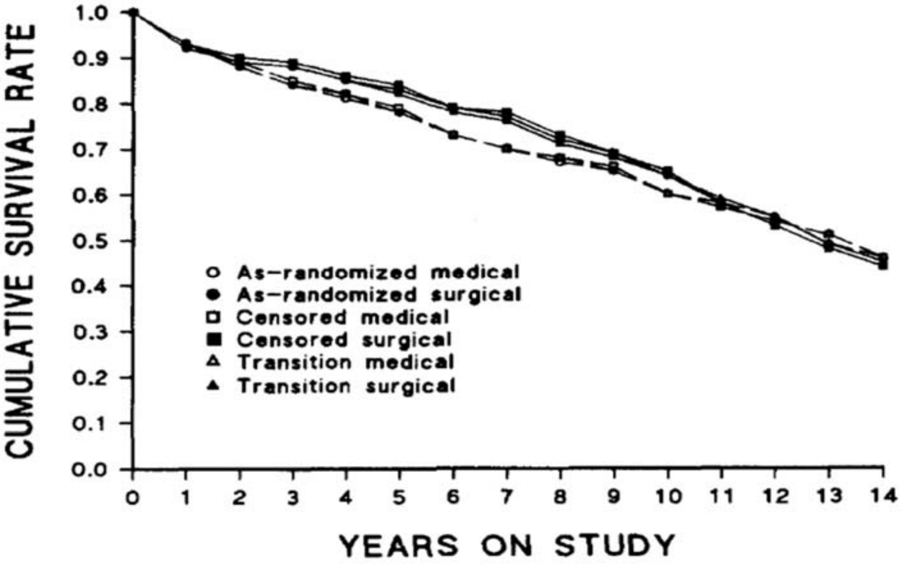
- ➢ Medical therapy versus bypass surgery for CAD
  - – 55% of medical therapy group received bypass surgery at some time during 14 years of follow up
  - – Small % of surgery group refused surgery
- ➢ Compare 5 analysis methods:
  1. ITT ('as-randomized')
  2. Exclude treatment crossovers from analysis ('adherers-only')
  3. Include crossovers in alternate group ('treatment-received')
  4. Censor crossovers at time of treatment change ('censored')
  5. Transfer crossovers to alternate group at time of treatment change ('transition')

*Peduzzi et al. *Stat Med* 1993;12:1185-95.

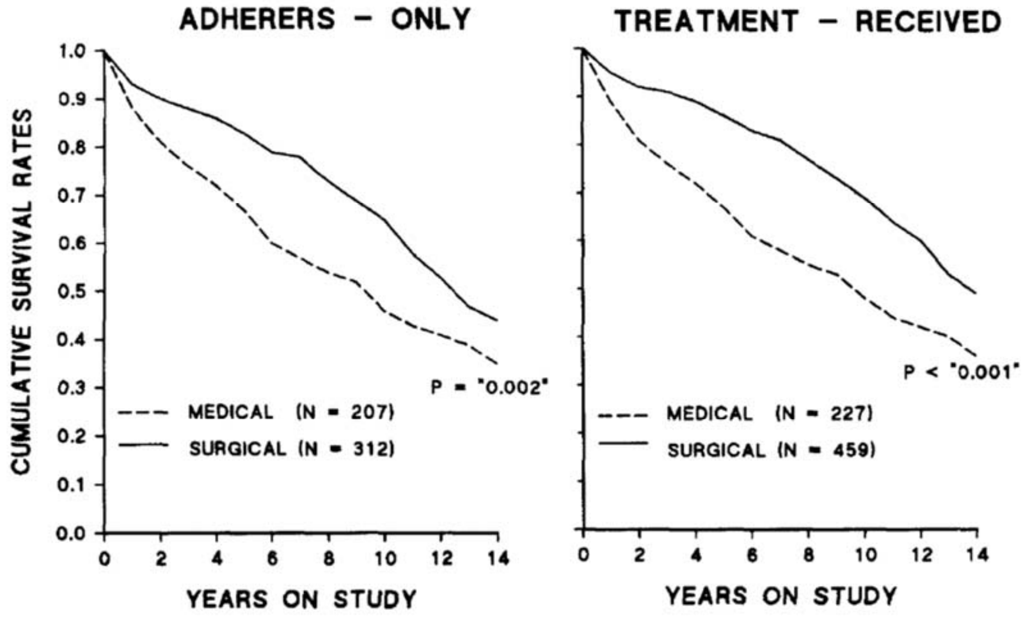
Results of ITT Analysis



Results of ITT, Censored, and Transition Analyses



Results of Adherers-Only and Treatment-Received Analyses



Conclusions

- Adherers-only and treatment-received analyses suffer from severe length-sampling bias
  - The longer the patient survives, the more chance to cross over to surgery
- Although designed to compare medical to surgical therapy, the trial ultimately compared two treatment strategies:
  - Treat with medical therapy until surgery warranted
  - Immediate surgery

Intention to Treat (ITT)

- ➢ ITT is the preferred analysis method in clinical trials as it avoids potential biases related to failure to adhere to assigned treatment
- ➢ ITT tests treatment strategy, rather than treatment received
  - Effect of following a treatment strategy is what is relevant when faced with a new patient
  - At time of initial treatment decision, it is unknown whether patient will adhere to treatment or whether other factors will intervene
  - Analyses based on knowledge of future events are not very relevant to current decision

Principles for Primary Data Analysis

1. There is no substitute for a descriptive plot of the data
2. The possible effects of chance on the observed data (treatment difference) must be quantified
3. Use intention-to-treat
4. Adjust for randomization stratification covariates and/or baseline level of outcome

Why Adjust for Randomization Stratification Variables?

- Generally strongly related to outcome
  - Stratification helps to ensure treatment groups are balanced on the variable
- Stratification variable relates to outcome regardless of treatment
- Adjustment for stratification variable helps to explain variability in the outcome, thereby reducing unexplained (error) variability
- Statistical power is increased
  - Same reasoning may be true for baseline level of outcome

Example – LRT for DME

Change in Visual Acuity - 1 Year (letters)	**Sham + Prompt Laser**	**Anti-VEGF + Prompt Laser**	**Anti-VEGF + Deferred Laser**	**Steroid + Prompt Laser**
Mean ± SD	+3 ± 13	+9 ± 11	+9 ± 12	+4 ± 13
P-value*		<0.001	<0.001	0.31

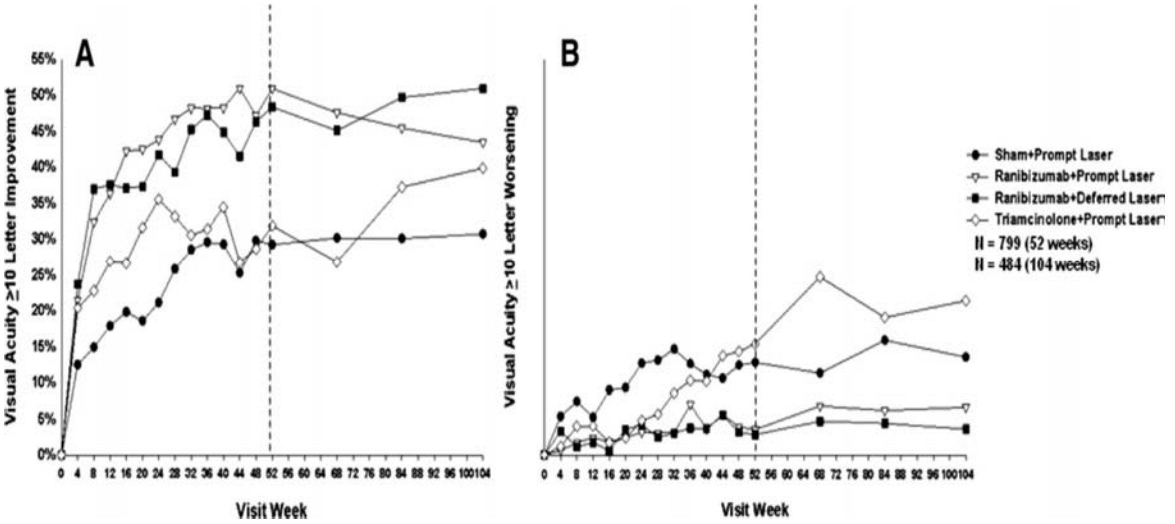
*From analysis of covariance adjusted for baseline visual acuity and correlation between study eyes.

- Baseline visual acuity is a randomization stratification variable
- Because visual acuity measurement is bounded there is limited room for improvement near the top and limited room for worsening at the bottom of the scale
  - Baseline VA is related to outcome (change)

Principles for Primary Data Analysis

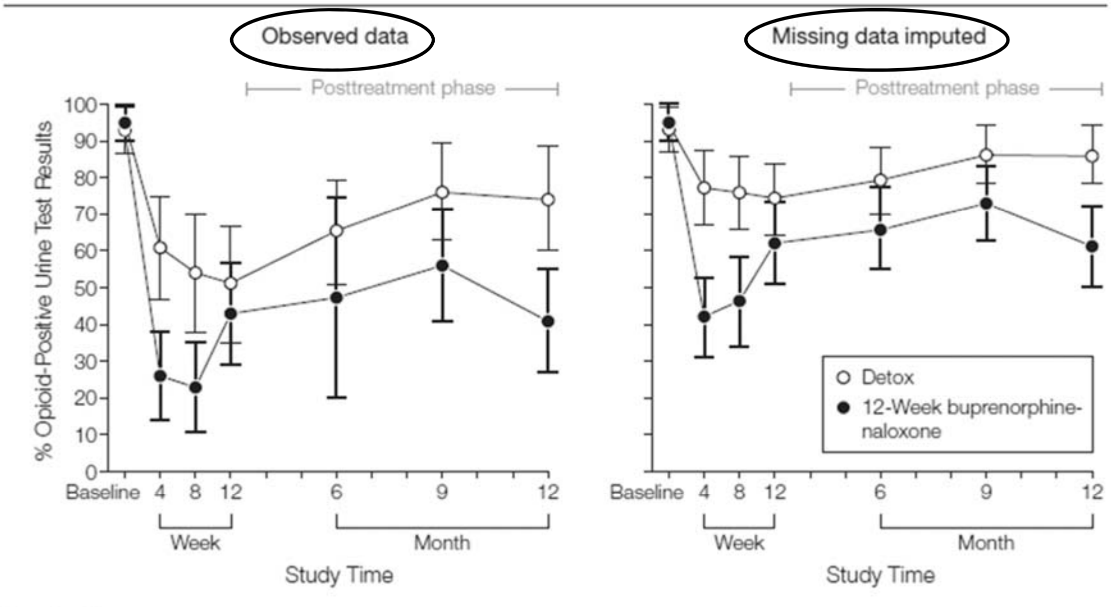
1. There is no substitute for a descriptive plot of the data
2. The possible effects of chance on the observed data (treatment difference) must be quantified
3. Use intention-to-treat
4. Adjust for baseline level of outcome and randomization stratification variables
5. Perform sensitivity analyses
   - Modified outcome or statistical model
   - Missing data assumption

LRT for DME



- Proportion of patients with 10 or more letter improvement is higher in anti-VEGF groups
- Proportion of patients with 10 or more letter worsening is higher in sham+laser and steroid+laser

Figure 2. Percentage of Opioid-Positive Urine Test Results at Baseline and Weeks 4, 8, and 12 and Follow-up Months 6, 9, and 12



No. of patients	Baseline	4	8	12	6	9	12
Detox	78	59	53	53	46	45	42
12-Week[a]	74	58	52	49	47	45	49

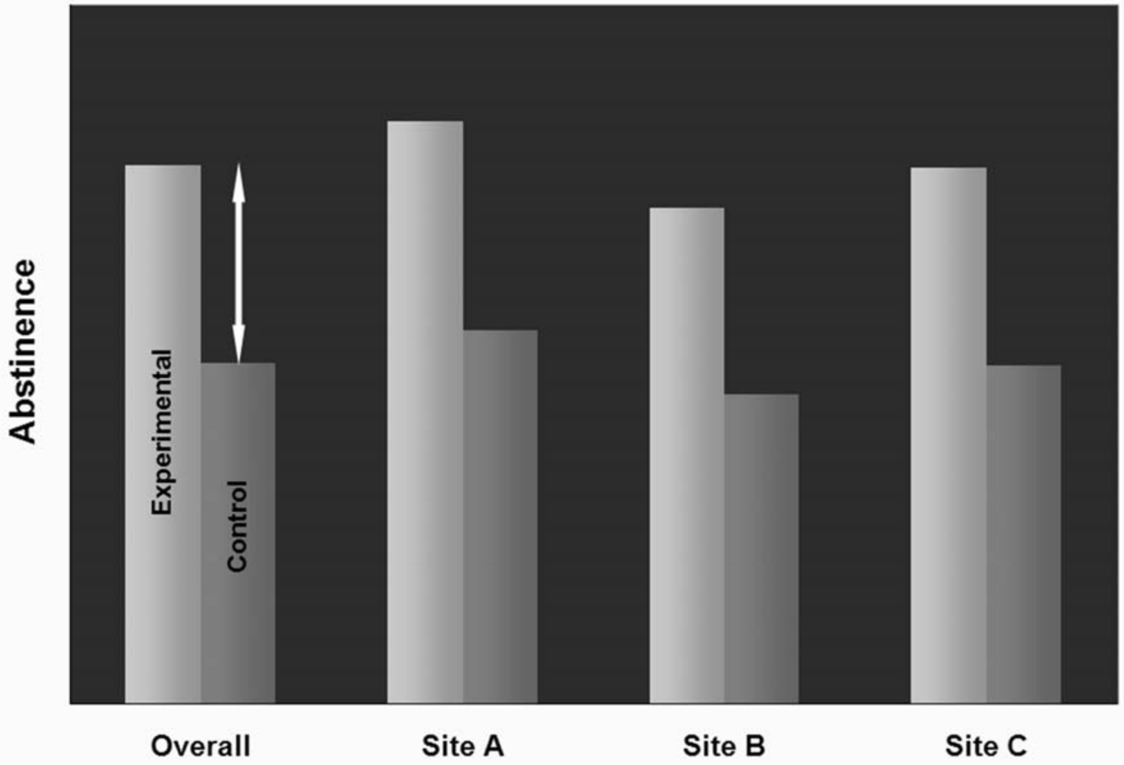
Woody et al. JAMA 2008

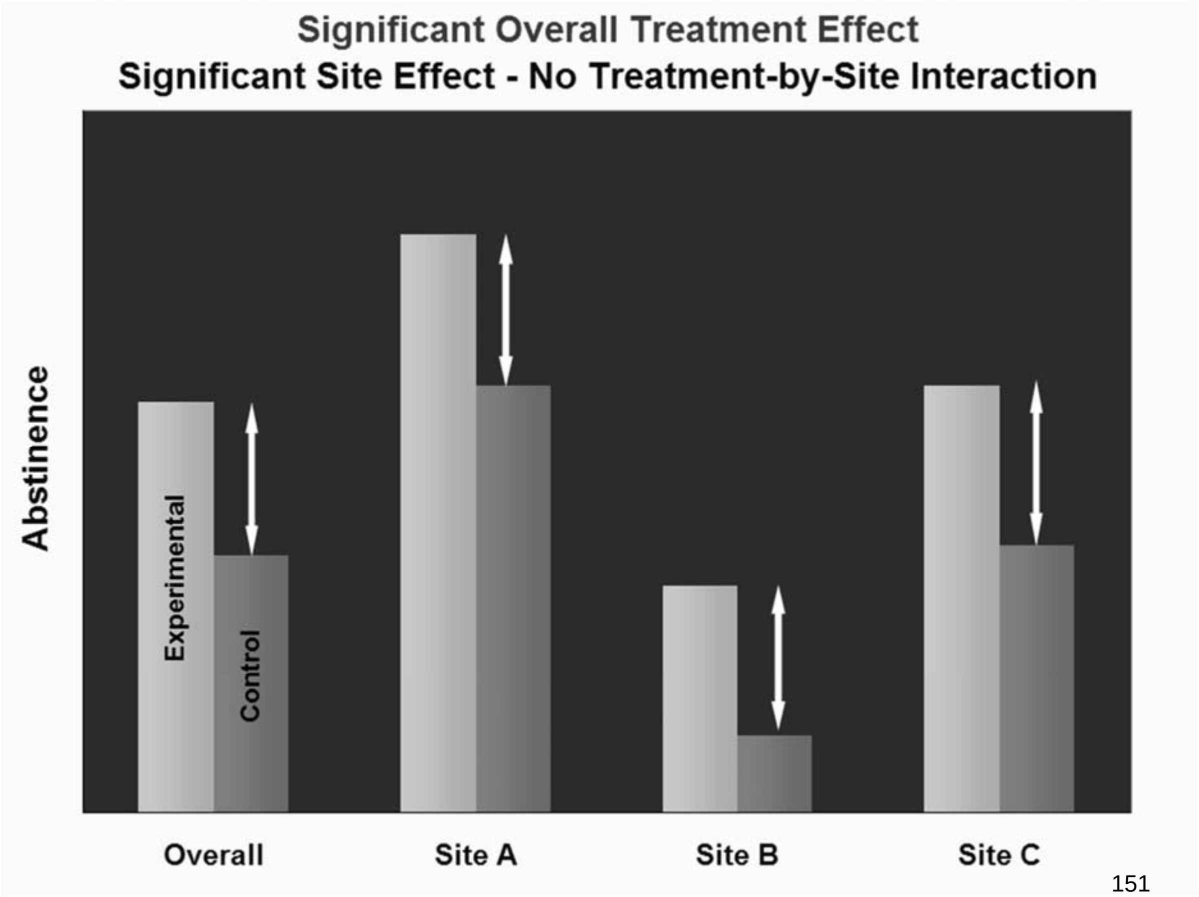
Detox indicates detoxification group. Error bars indicate 95% confidence intervals.
[a] 12-Week buprenorphine-naloxone group.

Principles for Primary Data Analysis

1. There is no substitute for a descriptive plot of the data
2. The possible effects of chance on the observed data (treatment difference) must be quantified
3. Use intention-to-treat
4. Adjust for baseline level of outcome and randomization stratification variables
5. Perform sensitivity analyses for missing data
6. Check for site effects and site*treatment interaction

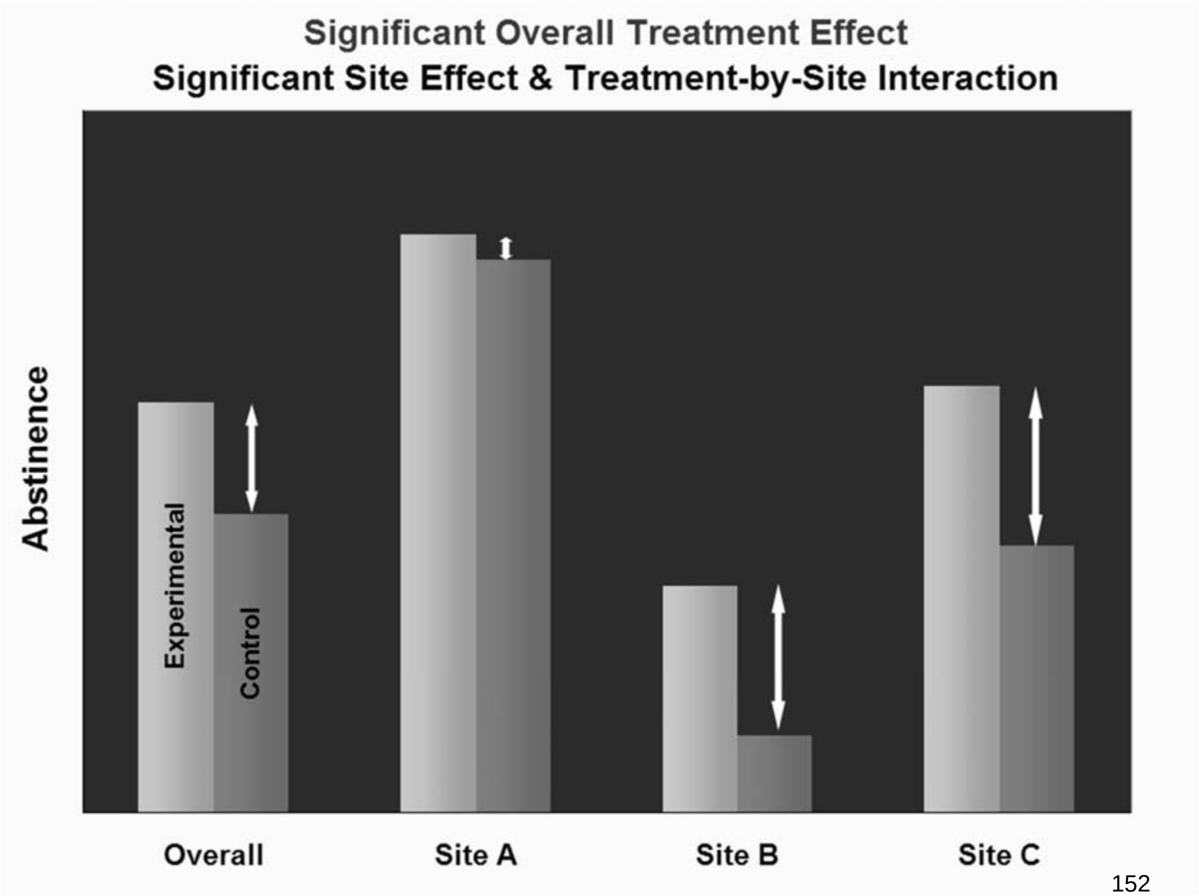
Significant Overall Treatment Effect
No Site Effect





Significant Overall Treatment Effect

Significant Site Effect - No Treatment-by-Site Interaction



Significant Overall Treatment Effect

Significant Site Effect & Treatment-by-Site Interaction

All combinations of significant *site effect* and *treatment-by-site interaction* are possible

Site Effect

- Based on the overall abstinence level (from both conditions) at each site
- Nothing to do with the treatment effect (the treatment difference) at each site

Treatment-by-Site Interaction

- Treatment effect varies by site

Summary – Site Effects

- Significant site effects are not surprising, and *do not affect* overall conclusion regarding treatment effect
- Significant site by treatment interaction *does affect* the interpretation of results and requires further investigation

5. Subgroup Analyses

What are subgroup analyses?

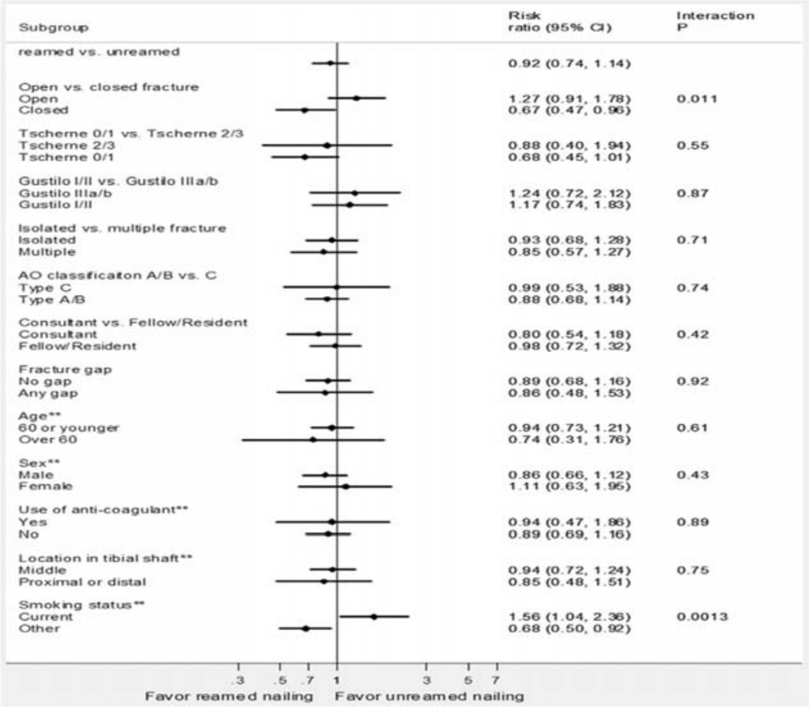
- Special type of secondary analyses that focus on differences in treatment effect among subgroups of trial participants, i.e. does treatment effect differ by:
  - Prognostic factor(s), such as disease severity
  - Comorbidity
  - Gender, age, racial/ethnic group
  - Genes/alleles associated with disease or treatment response

Reasons for Subgroup Analyses

- When overall analysis shows a treatment difference:
  - Show efficacy benefits hold across all clinically important subgroups, including those for whom there was reason to suspect less benefit
  - Identify subgroups with larger or smaller benefit
- Identify subgroups with significant benefit, when overall effect is NOT significant
- Generate hypotheses for future investigation

Adapted from Grouin et al 2005

Subgroup Analyses in SPRINT*



Subgroup	Risk ratio (95% CI)	Interaction P
reamed vs. unreamed	0.92 (0.74, 1.14)	
Open vs. closed fracture		
Open	1.27 (0.91, 1.78)	0.011
Closed	0.67 (0.47, 0.96)	
Tscherne 0/1 vs. Tscherne 2/3		
Tscherne 2/3	0.88 (0.40, 1.94)	0.55
Tscherne 0/1	0.68 (0.45, 1.01)	
Gustilo I/II vs. Gustilo IIIa/b		
Gustilo IIIa/b	1.24 (0.72, 2.12)	0.87
Gustilo I/II	1.17 (0.74, 1.83)	
Isolated vs. multiple fracture		
Isolated	0.93 (0.68, 1.28)	0.71
Multiple	0.85 (0.57, 1.27)	
AO classificaton A/B vs. C		
Type C	0.99 (0.53, 1.88)	0.74
Type A/B	0.88 (0.68, 1.14)	
Consultant vs. Fellow/Resident		
Consultant	0.80 (0.54, 1.18)	0.42
Fellow/Resident	0.98 (0.72, 1.32)	
Fracture gap		
No gap	0.89 (0.68, 1.16)	0.92
Any gap	0.86 (0.48, 1.53)	
Age**		
60 or younger	0.94 (0.73, 1.21)	0.61
Over 60	0.74 (0.31, 1.76)	
Sex**		
Male	0.86 (0.66, 1.12)	0.43
Female	1.11 (0.63, 1.95)	
Use of anti-coagulant**		
Yes	0.94 (0.47, 1.86)	0.89
No	0.89 (0.69, 1.16)	
Location in tibial shaft**		
Middle	0.94 (0.72, 1.24)	0.75
Proximal or distal	0.85 (0.48, 1.51)	
Smoking status**		
Current	1.56 (1.04, 2.36)	0.0013
Other	0.68 (0.50, 0.92)	

*Study to Prospectively Evaluate Reamed Intramedullary Nails in Tibial Fractures - Sun et al 2011

Interpretation of Subgroup Analyses is Controversial

Although subgroup analyses can provide new, provocative, and sometimes clinically relevant findings, this group of evaluations must be handled with extreme care.

Moyé 2012

Interpretation of Subgroup Analyses is Controversial

- *"Subgroups Kill People"*
- *… And Lack of Subgroup Analysis Kills People*
- *Researchers are thus criticized by policymakers for not doing enough subgroup analyses, and criticized by statisticians for doing too many: they are damned if they do, and damned if they don't.*

Petticrew et al. 2012

- "…if there are subgroup differences, only subgroup analyses can find them"

Berry 1990

Issues with Subgroup Analyses

- Analyzing many subgroups greatly increases the chance of type I (false positive) errors
  - Finding a significant subgroup effect in the context of multiple subgroup analyses cannot necessarily be considered conclusive evidence of a subgroup effect
- Trials are rarely adequately powered for subgroup analyses
  - Finding no significant effect of a subgroup cannot be considered conclusive evidence of no effect when the comparison is underpowered

Implications for Study Design

- Pre-specify the subgroup hypothesis(es), including expected direction of effect(s), in the protocol/statistical analysis plan
- Justify clinical importance and prior evidence, if any, supporting a subgroup effect
- Discuss place in the overall testing strategy
  - E.g., only test subgroup if overall effect is significant
  - Specify adjustments, if any, for multiple testing
- Evaluate *a priori* statistical power

Implications for Study Design

- Consider stratifying the randomization on important subgroup factor(s) if subgroup sample size is small
  - Improves chance of treatment group balance
  - Less important if subgroup factor not prognostic
- If a qualitative subgroup by treatment interaction is strongly suspected, consider powering the trial for subgroup analysis or design separate trials
  - Reporting an overall effect rarely makes sense in context of qualitative interaction

Guidelines for Analysis

- Consistency of treatment effect within subgroups is commonly assessed by adding the subgroup factor and subgroup factor x treatment interaction to the primary analysis model
  - This is preferred over performing a separate analysis of treatment effect within each level of the subgroup
- As the test for interaction generally lacks good statistical power, should also pay attention to size and direction of subgroup effects and apply clinical judgment regarding their importance

Guidelines for Analysis

- If primary analysis is adjusted for baseline factors, the subgroup analyses should adjust for same factors
  - Adjust for randomization stratification factors
- Subgroup analyses aimed towards showing a benefit in a subgroup when the overall treatment effect is not significant are highly problematic
  - If there was a hypothesis supporting beneficial effect in a subgroup but not overall population, it didn't make sense to do the trial in overall population
  - Raises level of concern that effect is due to chance
  - Do these analyses for hypothesis generation

Ten Criteria to Assess the Credibility of Subgroup Analyses

Design

1) *Was the subgroup hypothesis specified a priori?*

2) *Was the subgroup analysis one of a small number of subgroup hypotheses tested (≤5)?*

3) *Was the subgroup variable a baseline characteristic?*

4) *Was the subgroup variable a stratification factor at randomization?*

Sun et al. 2012

Ten Criteria to Assess the Credibility of Subgroup Analyses

Context

5) *Was the direction of subgroup effect correctly pre-specified?*
6) *Was the subgroup effect consistent with evidence from previous related studies?*
7) *Was the subgroup effect consistent across related outcomes?*
8) *Was there any indirect evidence to support the apparent subgroup effect—for example, biological rationale, laboratory tests, animal studies?*

Sun et al. 2012

Ten Criteria to Assess the Credibility of Subgroup Analyses

Analysis

9) *Was the test of interaction significant (interaction P<0.05)?*

10) *Was the significant interaction effect independent, if there were multiple significant interactions?*

Sun et al. 2012

Evaluation of Subgroup Findings in SPRINT*

Criterion	Subgroup Effect	
	Fracture Type	Smoking Status
Baseline characteristic	Yes	Yes
Hypothesis and direction of effect specified *a priori*	Yes	No
Number of subgroup hypotheses tested	7	12
Consistent with previous studies	No	No evidence
Biological rationale	Yes	No
Significant test for interaction	P=0.011	P=0.0013

*Sun et al 2011

Subgroup Analyses - Bottom Line

- They are important
- They should be pre-specified and based on clinically plausible hypotheses
- They are valuable for generating new hypotheses
- They should be limited to a handful
- They generally should *not* be interpreted as definitive
- Significant subgroup effect when there is no overall effect should be regarded with particular suspicion
- Their limitations should be clearly reported

7. Publication of Results

(preparing the primary manuscript)

Although the field of statistics is rooted in mathematics, and mathematics is exact, the use of statistics to describe complex phenomena is not exact. That leaves plenty of room for shading the truth.

Charles Wheelan

It's easy to lie with statistics, but it's hard to tell the truth without them.

Andrejs Dunkels

From Wheelan 2013

Friedman et al. 2010
(Chapter 19 – Reporting and Interpreting of Results)

To communicate appropriately, the investigators have to review their results critically and avoid the temptation of overinterpretation.

They are in the privileged position of knowing the quality and limitations of the data better than anyone else.

Therefore, they have the responsibility for presenting the results clearly and concisely, together with any issues that might bear on their interpretation.

References for General Guidelines

- CONSORT guidelines for general reporting (Altman et al. 2001*)
- Uniform requirements for manuscripts submitted to biomedical journals (International Committee of Medical Journal Editors 2010)
- Informative abstracts of clinical articles (Ad Hoc Working Group for Critical Appraisal of the Medical Literature 1987)
- Translating statistical findings into plain English (Pocock & Ware 2009)
- Statistical problems in the reporting of clinical trials (Pocock et al. 1987)

Friedman et al. 2010 (Chapter 19)

* See also Schulz et al. 2010 and Moher et al. 2010 for updated CONSORT Guidelines

References for Specific Guidelines

- Reporting of noninferiority and equivalence randomized trial results (Piaggio et al. 2006)
- Reporting systematic reviews and meta-analyses (Moher et al. 2009)
- Reporting of subgroup analyses (Wang et al. 2007)
- Extension of CONSORT Guidelines for reporting safety (Ioannidis et al. 2004)
- Reporting on randomization and baseline comparisons in clinical trials (Altman & Doré 1990)

Friedman et al. 2010 (Chapter 19)

Check List of Key Elements in Published Reports on Clinical Trials

- A clear and concise summary of the protocol, i.e. what was pre-specified:
  - The primary research question and why it is important
  - Precise and detailed description of the interventions
  - The primary and secondary outcome measures
  - Eligibility (inclusion/exclusion) criteria
  - Blinding (masking)
  - Sample size and power analysis
  - Randomization (stratification factors)
  - Planned statistical analysis
  - Planned interim analysis (if any)

Based on Altman et al. 2001

Check List of Key Elements in Published Reports on Clinical Trials

- CONSORT flow diagram

Based on Altman et al. 2001 and Friedman et al. 2010

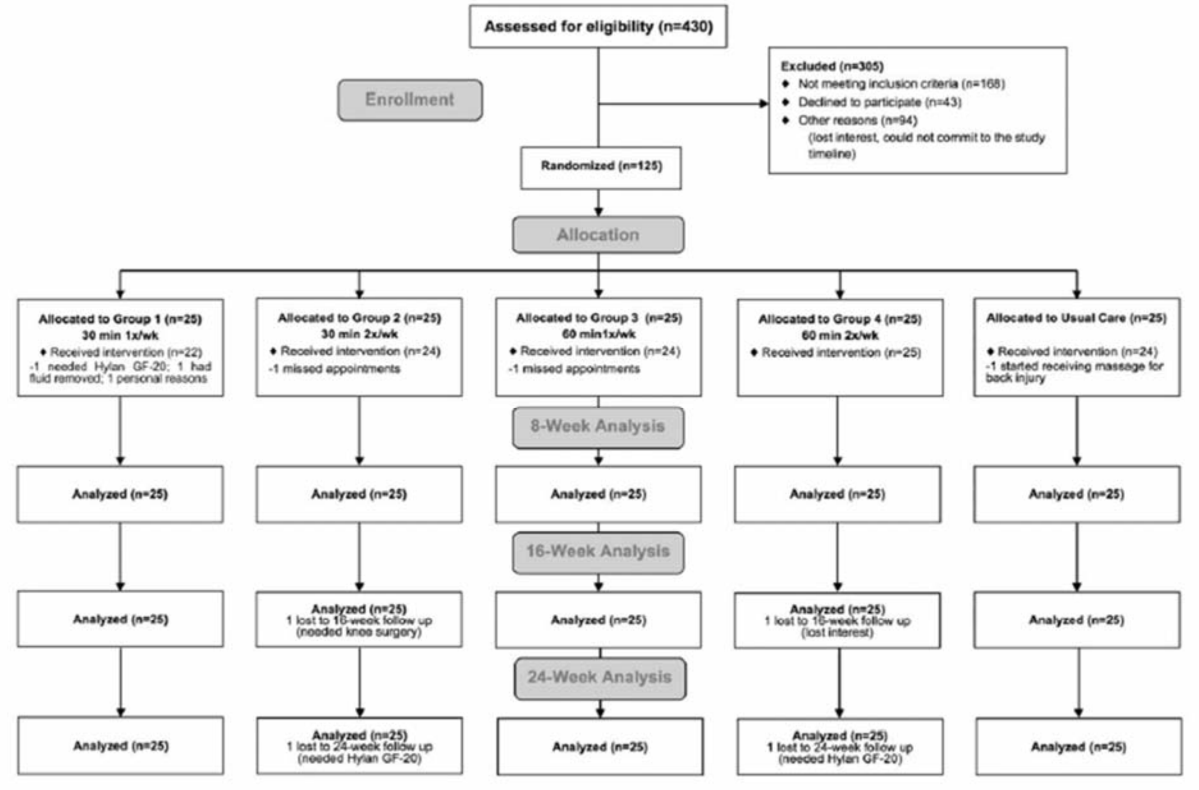

Assessed for eligibility (n=430)
Enrollment
Excluded (n=305)
◆ Not meeting inclusion criteria (n=168)
◆ Declined to participate (n=43)
◆ Other reasons (n=94)
(lost interest, could not commit to the study timeline)
Randomized (n=125)
Allocation
Allocated to Group 1 (n=25) 30 min 1x/wk
◆ Received intervention (n=22)
-1 needed Hylan GF-20; 1 had fluid removed; 1 personal reasons
Allocated to Group 2 (n=25) 30 min 2x/wk
◆ Received intervention (n=24)
-1 missed appointments
Allocated to Group 3 (n=25) 60 min1x/wk
◆ Received intervention (n=24)
-1 missed appointments
Allocated to Group 4 (n=25) 60 min 2x/wk
◆ Received intervention (n=25)
Allocated to Usual Care (n=25)
◆ Received intervention (n=24)
-1 started receiving massage for back injury
8-Week Analysis
Analyzed (n=25)
Analyzed (n=25)
Analyzed (n=25)
Analyzed (n=25)
Analyzed (n=25)
16-Week Analysis
Analyzed (n=25)
Analyzed (n=25) 1 lost to 16-week follow up (needed knee surgery)
Analyzed (n=25)
Analyzed (n=25) 1 lost to 16-week follow up (lost interest)
Analyzed (n=25)
24-Week Analysis
Analyzed (n=25)
Analyzed (n=25) 1 lost to 24-week follow up (needed Hylan GF-20)
Analyzed (n=25)
Analyzed (n=25) 1 lost to 24-week follow up (needed Hylan GF-20)
Analyzed (n=25)

Figure 1. Participant Flow Diagram.

Perlman et al. 2012

Check List of Key Elements in Published Reports on Clinical Trials

- CONSORT flow diagram
- Baseline demographic and clinical characteristics of each treatment group

Based on Altman et al. 2001 and Friedman et al. 2010

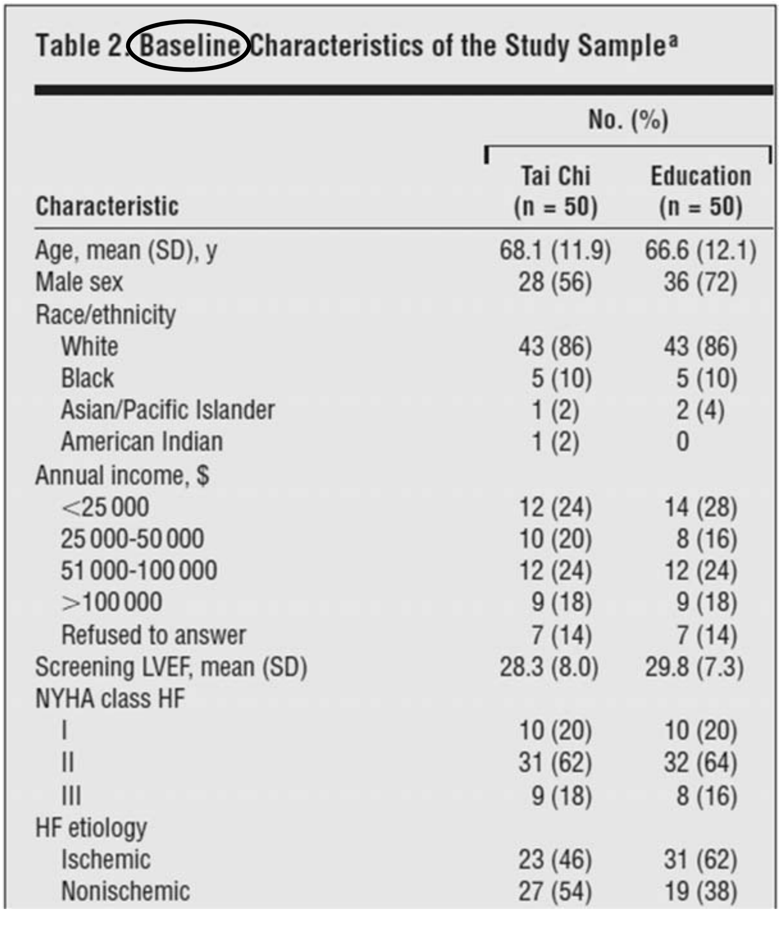


Table 2. Baseline Characteristics of the Study Sample[a]

Characteristic	No. (%) Tai Chi (n = 50)	No. (%) Education (n = 50)
Age, mean (SD), y	68.1 (11.9)	66.6 (12.1)
Male sex	28 (56)	36 (72)
Race/ethnicity		
White	43 (86)	43 (86)
Black	5 (10)	5 (10)
Asian/Pacific Islander	1 (2)	2 (4)
American Indian	1 (2)	0
Annual income, $		
<25 000	12 (24)	14 (28)
25 000-50 000	10 (20)	8 (16)
51 000-100 000	12 (24)	12 (24)
>100 000	9 (18)	9 (18)
Refused to answer	7 (14)	7 (14)
Screening LVEF, mean (SD)	28.3 (8.0)	29.8 (7.3)
NYHA class HF		
I	10 (20)	10 (20)
II	31 (62)	32 (64)
III	9 (18)	8 (16)
HF etiology		
Ischemic	23 (46)	31 (62)
Nonischemic	27 (54)	19 (38)

Yeh et al. 2011

Check List of Key Elements in Published Reports on Clinical Trials

- CONSORT flow diagram
- Baseline demographic and clinical characteristics of each treatment group
- Whether the trial worked as planned
- A clear description of the statistical method(s) used, and whether they differ from the planned analysis
- A clear description of the result of the primary analysis (including multiplicity adjustments)

Based on Altman et al. 2001 and Friedman et al. 2010

Check List of Key Elements in Published Reports on Clinical Trials

- Treatment (medication or therapy) adherence
- Extent of missing data
- Safety information (adverse events and side effects)
- Clinical implications of the findings
- Comparison of the findings with those from other studies
- Results of subgroup analyses (if any)
- Limitations

Based on Altman et al. 2001 and Friedman et al. 2010

Issues and Recommendations

1) May publish trial design before the trial is completed

2) May publish baseline characteristics (not by treatment condition) before data lock, but after recruitment is completed

3) Do not present, report or publish any “preliminary” post-randomization results before data lock

Issues and Recommendations

4) Avoid statistical tests and reports of p-values for differences in baseline characteristics between the treatment conditions (although some would disagree)

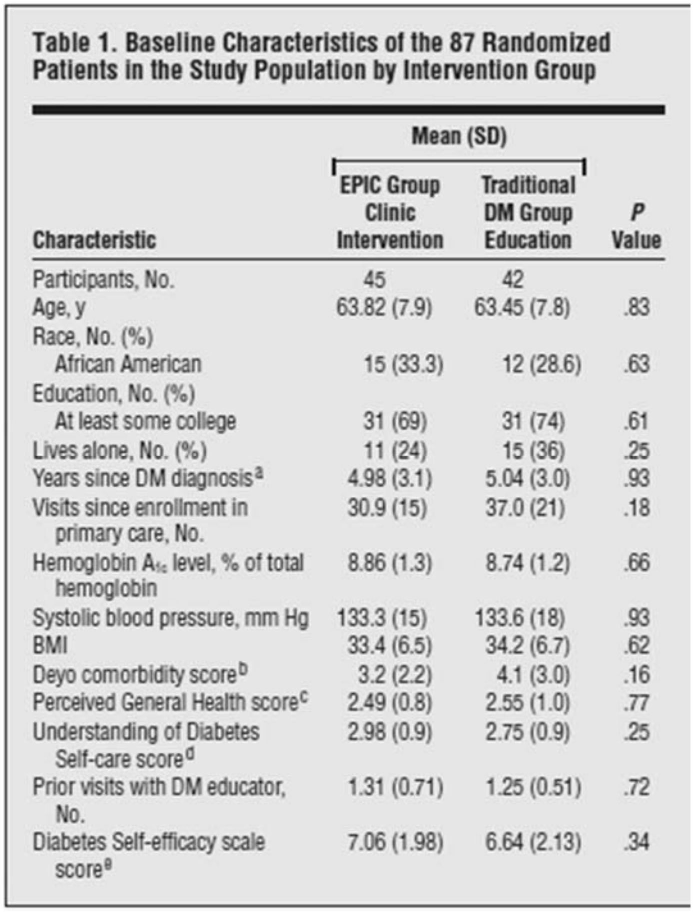


Table 1. Baseline Characteristics of the 87 Randomized Patients in the Study Population by Intervention Group

Characteristic	EPIC Group Clinic Intervention	Traditional DM Group Education	*P* Value
	Mean (SD)		
Participants, No.	45	42	
Age, y	63.82 (7.9)	63.45 (7.8)	.83
Race, No. (%)			
African American	15 (33.3)	12 (28.6)	.63
Education, No. (%)			
At least some college	31 (69)	31 (74)	.61
Lives alone, No. (%)	11 (24)	15 (36)	.25
Years since DM diagnosis[a]	4.98 (3.1)	5.04 (3.0)	.93
Visits since enrollment in primary care, No.	30.9 (15)	37.0 (21)	.18
Hemoglobin A_{1c} level, % of total hemoglobin	8.86 (1.3)	8.74 (1.2)	.66
Systolic blood pressure, mm Hg	133.3 (15)	133.6 (18)	.93
BMI	33.4 (6.5)	34.2 (6.7)	.62
Deyo comorbidity score[b]	3.2 (2.2)	4.1 (3.0)	.16
Perceived General Health score[c]	2.49 (0.8)	2.55 (1.0)	.77
Understanding of Diabetes Self-care score[d]	2.98 (0.9)	2.75 (0.9)	.25
Prior visits with DM educator, No.	1.31 (0.71)	1.25 (0.51)	.72
Diabetes Self-efficacy scale score[e]	7.06 (1.98)	6.64 (2.13)	.34

Naik et al. 2011

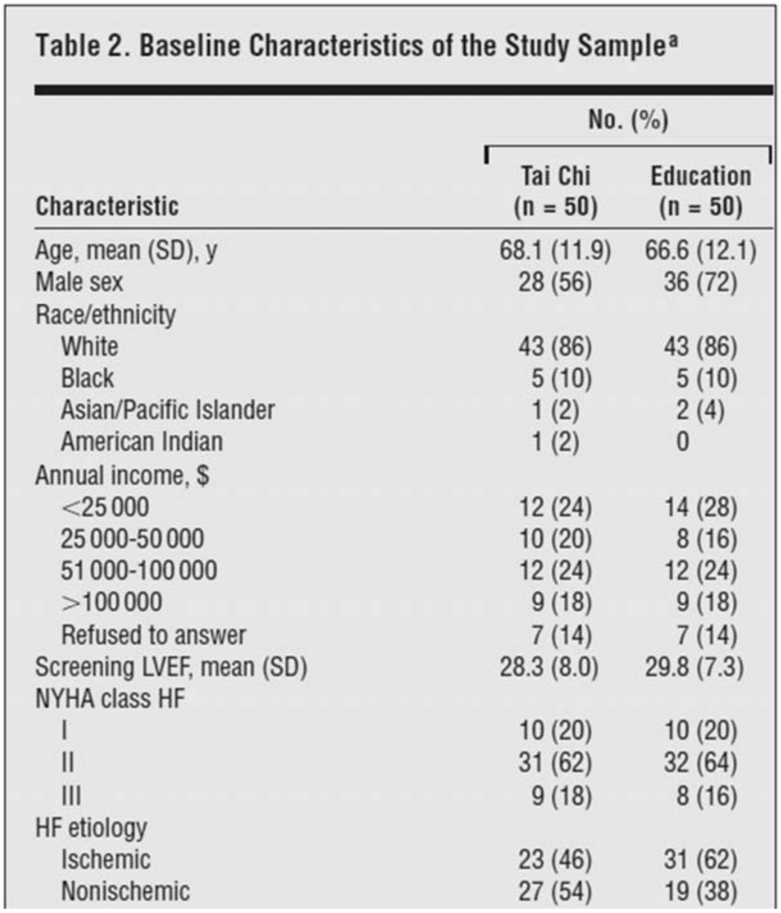
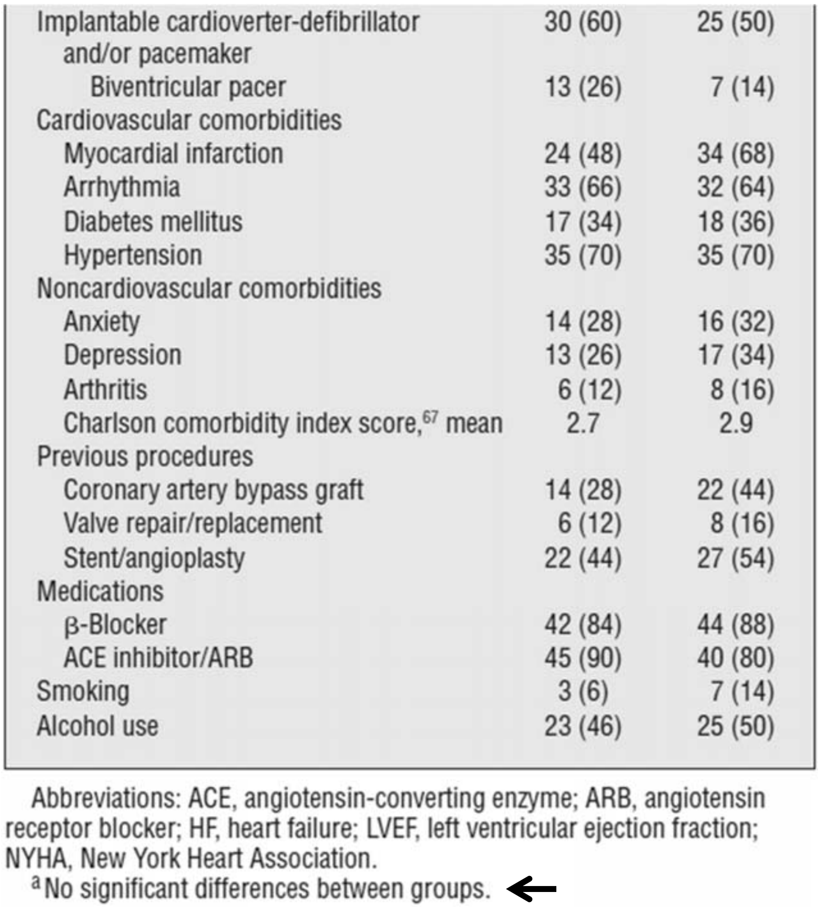


Table 2. Baseline Characteristics of the Study Sample[a]

Characteristic	Tai Chi (n = 50)	Education (n = 50)
	No. (%)	
Age, mean (SD), y	68.1 (11.9)	66.6 (12.1)
Male sex	28 (56)	36 (72)
Race/ethnicity		
White	43 (86)	43 (86)
Black	5 (10)	5 (10)
Asian/Pacific Islander	1 (2)	2 (4)
American Indian	1 (2)	0
Annual income, $		
<25 000	12 (24)	14 (28)
25 000-50 000	10 (20)	8 (16)
51 000-100 000	12 (24)	12 (24)
>100 000	9 (18)	9 (18)
Refused to answer	7 (14)	7 (14)
Screening LVEF, mean (SD)	28.3 (8.0)	29.8 (7.3)
NYHA class HF		
I	10 (20)	10 (20)
II	31 (62)	32 (64)
III	9 (18)	8 (16)
HF etiology		
Ischemic	23 (46)	31 (62)
Nonischemic	27 (54)	19 (38)

Yeh et al. 2011



Implantable cardioverter-defibrillator and/or pacemaker	30 (60)	25 (50)
Biventricular pacer	13 (26)	7 (14)
Cardiovascular comorbidities		
Myocardial infarction	24 (48)	34 (68)
Arrhythmia	33 (66)	32 (64)
Diabetes mellitus	17 (34)	18 (36)
Hypertension	35 (70)	35 (70)
Noncardiovascular comorbidities		
Anxiety	14 (28)	16 (32)
Depression	13 (26)	17 (34)
Arthritis	6 (12)	8 (16)
Charlson comorbidity index score,[67] mean	2.7	2.9
Previous procedures		
Coronary artery bypass graft	14 (28)	22 (44)
Valve repair/replacement	6 (12)	8 (16)
Stent/angioplasty	22 (44)	27 (54)
Medications		
β-Blocker	42 (84)	44 (88)
ACE inhibitor/ARB	45 (90)	40 (80)
Smoking	3 (6)	7 (14)
Alcohol use	23 (46)	25 (50)

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin receptor blocker; HF, heart failure; LVEF, left ventricular ejection fraction; NYHA, New York Heart Association.

[a] No significant differences between groups. ←

Yeh et al. 2011

Issues and Recommendations

5) Clearly indicate the pre-specified *primary* hypothesis and the corresponding result in the body of the paper, as well as in the "Summary" or "Abstract"

Example
Primary Aim, Primary & Secondary Endpoints

From the Abstract:

Method. *This study aimed to directly compare the effects of CBT versus IPT for the treatment of panic disorder with agoraphobia…*
Primary outcomes were panic attack frequency and an idiosyncratic behavioral test.
Secondary outcomes were panic and agoraphobia severity, panic-related cognitions, interpersonal functioning and general psychopathology.

Vos et al. 2012

Example – Primary and Secondary Results

From the Abstract:

Results. *Intention-to-treat (ITT) analyses on the primary outcomes indicated superior effects for CBT in treating panic disorder with agoraphobia.*
Per-protocol analyses emphasized the differences between treatments and yielded larger effect sizes.
Reductions in the secondary outcomes were equal for both treatments, except for agoraphobic complaints and behavior and the credibility ratings of negative interpretations of bodily sensations, all of which decreased more in CBT.

Vos et al. 2012

Issues and Recommendations

6) Clearly identify all secondary analyses (including subgroup analyses) as such, and as exploratory findings that need to be confirmed. Indicate whether they were pre-specified.

Reporting Secondary Analyses

A well-reported secondary analysis must make clear to the reader the uncertainty of the result – so clear, in fact, that it should be an obvious part of the conclusions that implementation should await confirmation as the primary outcome in an adequately powered trial.

Marler 2012

Issues and Recommendations

7) Do not highlight (in “Summary” or “Abstract”) *only* what turned out to be statistically significant

Example

… an important medical conference had just featured a study claiming that the new arthritis drug Celebrex was safer on the stomach than more established drugs…

The truth was that Celebrex was no better at protecting the stomach from serious complications than other drugs. It appeared that way only because Pfizer and its partner, Pharmacia, presented the results from the first six months of a yearlong study rather than the whole thing.

In Documents on Pain Drug, Signs of Doubt and Deception
The New York Times (Health), June 24, 2012

Example (cont'd)

Then and now, Pfizer has defended its decision to release partial results from the 2000 study and denies any intent to deceive. Company officials have said the drug has demonstrated its worth and safety…

Pfizer has argued that presenting the limited data was legitimate because so many people taking a comparison drug, diclofenac, dropped out, biasing the later results.

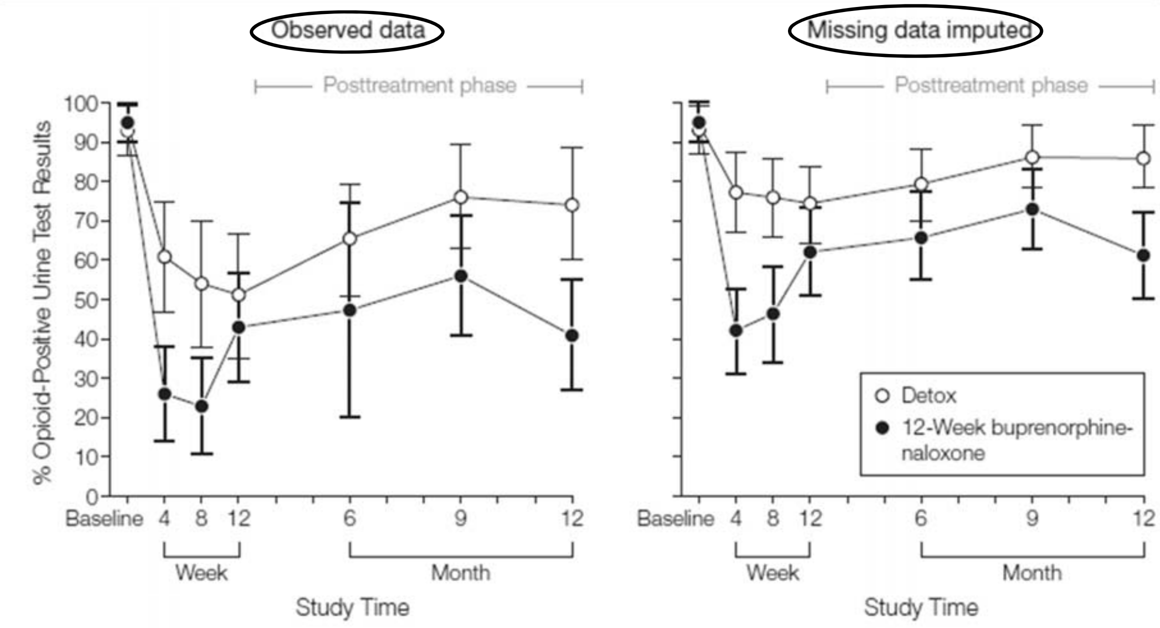
In Documents on Pain Drug, Signs of Doubt and Deception
The New York Times (Health), June 24, 2012

Issues and Recommendations

8) Show the full picture (e.g. graphs over time) with confidence intervals

9) Include in the primary manuscript results of any sensitivity analysis

Figure 2. Percentage of Opioid-Positive Urine Test Results at Baseline and Weeks 4, 8, and 12 and Follow-up Months 6, 9, and 12



No. of patients							
Detox	78	59	53	53	46	45	42
12-Week[a]	74	58	52	49	47	45	49

Woody et al. JAMA 2008

Detox indicates detoxification group. Error bars indicate 95% confidence intervals.
[a] 12-Week buprenorphine-naloxone group.

Issues and Recommendations

10) Report the values of p-values, not intervals, i.e. avoid $p<0.05$, $p<0.01$, etc.

👎

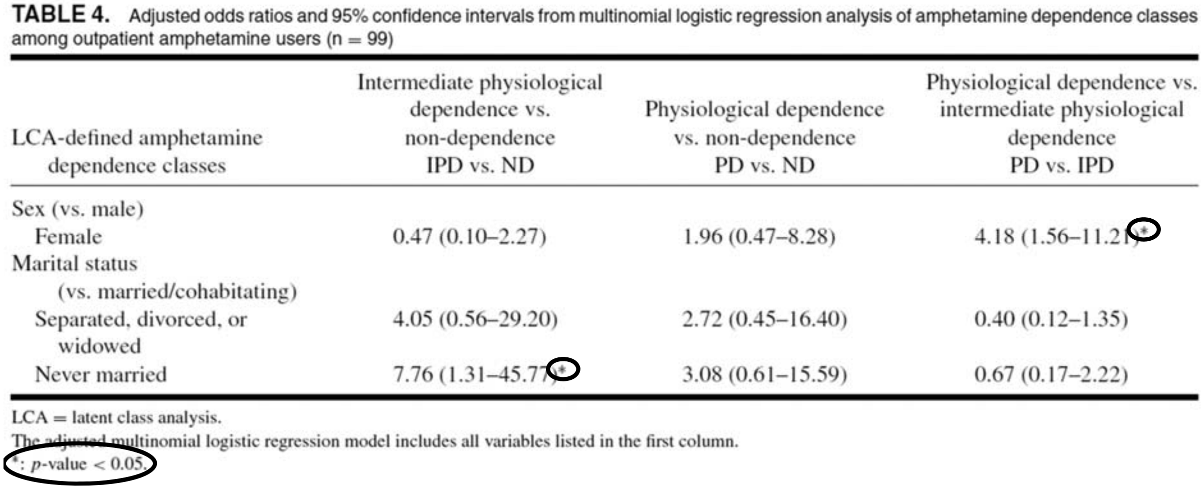


TABLE 4. Adjusted odds ratios and 95% confidence intervals from multinomial logistic regression analysis of amphetamine dependence classes among outpatient amphetamine users (n = 99)

LCA-defined amphetamine dependence classes	Intermediate physiological dependence vs. non-dependence IPD vs. ND	Physiological dependence vs. non-dependence PD vs. ND	Physiological dependence vs. intermediate physiological dependence PD vs. IPD
Sex (vs. male)			
Female	0.47 (0.10–2.27)	1.96 (0.47–8.28)	4.18 (1.56–11.21)*
Marital status (vs. married/cohabitating)			
Separated, divorced, or widowed	4.05 (0.56–29.20)	2.72 (0.45–16.40)	0.40 (0.12–1.35)
Never married	7.76 (1.31–45.77)*	3.08 (0.61–15.59)	0.67 (0.17–2.22)

LCA = latent class analysis.
The adjusted multinomial logistic regression model includes all variables listed in the first column.
*: *p*-value < 0.05.

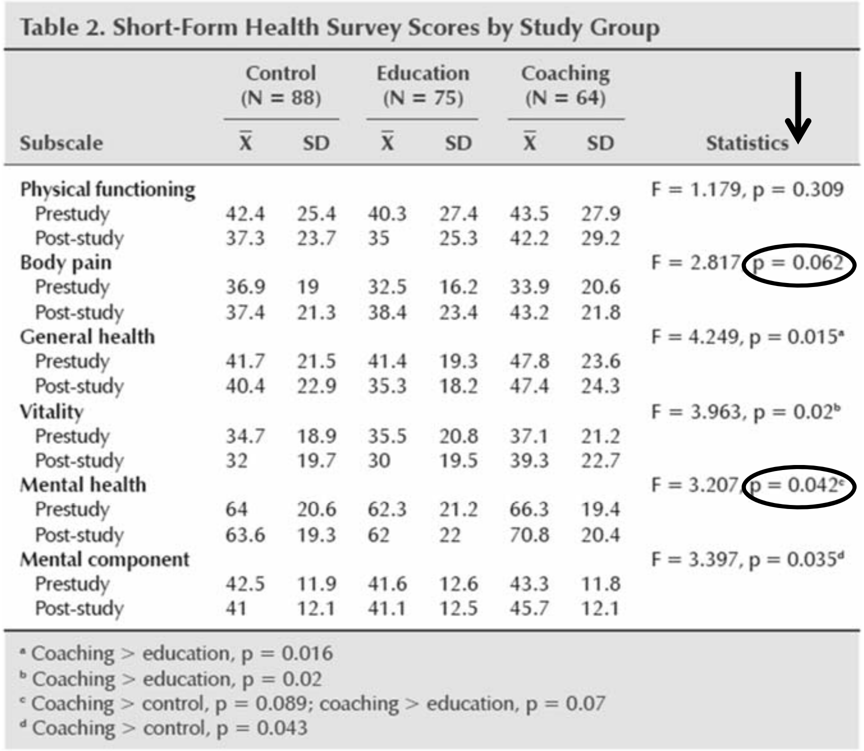


Table 2. Short-Form Health Survey Scores by Study Group

	Control (N = 88)		Education (N = 75)		Coaching (N = 64)		
Subscale	X̄	SD	X̄	SD	X̄	SD	**Statistics** ↓
Physical functioning							F = 1.179, p = 0.309
Prestudy	42.4	25.4	40.3	27.4	43.5	27.9	
Post-study	37.3	23.7	35	25.3	42.2	29.2	
Body pain							F = 2.817, p = 0.062
Prestudy	36.9	19	32.5	16.2	33.9	20.6	
Post-study	37.4	21.3	38.4	23.4	43.2	21.8	
General health							F = 4.249, p = 0.015[a]
Prestudy	41.7	21.5	41.4	19.3	47.8	23.6	
Post-study	40.4	22.9	35.3	18.2	47.4	24.3	
Vitality							F = 3.963, p = 0.02[b]
Prestudy	34.7	18.9	35.5	20.8	37.1	21.2	
Post-study	32	19.7	30	19.5	39.3	22.7	
Mental health							F = 3.207, p = 0.042[c]
Prestudy	64	20.6	62.3	21.2	66.3	19.4	
Post-study	63.6	19.3	62	22	70.8	20.4	
Mental component							F = 3.397, p = 0.035[d]
Prestudy	42.5	11.9	41.6	12.6	43.3	11.8	
Post-study	41	12.1	41.1	12.5	45.7	12.1	

[a] Coaching > education, p = 0.016
[b] Coaching > education, p = 0.02
[c] Coaching > control, p = 0.089; coaching > education, p = 0.07
[d] Coaching > control, p = 0.043

👍

Thomas et al. 2012

Issues and Recommendations

11) Report the value of the treatment effect and the corresponding confidence interval. Similarly for all other important outcomes.

12) Translate the statistical results into simple-English clinical terms in “Results”, and explain the impact of these results on clinical practice in “Discussion”

Example

Time to suicidal ideation was significantly longer in patients allocated to SSRI compared to those allocated to IPT (HR=2.21, 95% CI 1.04–4.66, P=.038), even after controlling for treatment augmentation, benzodiazepine use, and comorbidity with anxiety disorders.

Rucci et al. 2011

Issues and Recommendations

13) When multiple tests are conducted for *primary* analyses, adjust for multiplicity, and state that it was done

14) When multiple tests are conducted for *secondary* and *exploratory* analyses, indicate the total number of tests that were conducted

Questions or Comments

References (1 of 7)

Ad Hoc Working Group for Critical Appraisal of the Medical Literature, *A proposal for more informative abstracts of clinical articles*, Annals of Internal Medicine, 1987, 106:598-604.

Altman DG & Doré CJ, *Randomization and baseline comparisons in clinical trials*, Lancet, 1990, 335:149-153.

Altman DG et al., for the CONSORT Group, *The revised CONSORT statement for reporting randomized trials: explanation and elaboration,* Annals of Internal Medicine, 2001, 134:663-694.

Bassler D, Briel M, Montori VM, Lane M et al., *Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects: Systematic Review and Meta-regression Analysis*, JAMA, 2010, 303(12):1180-1187.

Berry DA, *Subgroup Analyses,* Biometrics, 1990; 46:1227-30.

Briel M, Lane M, Montori VM et al., *Stopping randomized trials early for benefit: a protocol of the Study Of Trial Policy Of Interim Truncation-2 (STOPIT-2)*, Trials, 2009, 10:49-58.

Committee for Proprietary Medicinal Products (CPMP), *Points to Consider on Adjustment for Baseline Covariates*, Statistics in Medicine, 2004, 23:701-709.

Dmitrienko A, Molenberghs G, Chuang-Stein C & Offen W, *Analysis of Clinical Trials Using SAS: A Practical Guide*, 2005, SAS Institute Inc.

References (2 of 7)

Dmitrienko A et al. (editors), *Multiple Testing Problems in Pharmaceutical Statistics*, Chapman and Hall/CRC Biostatistics Series, 2009.

Dmitrienko A, *Key Multiplicity Problems in Clinical Trials*, presentation at the 2011 FDA/Industry Statistics Workshop, Washington, DC.

European Agency for the Evaluation of Medicinal Products (EMEA), Committee for Proprietary Medicinal Products (CPMP), *Points to Consider on Multiplicity Issues in Clinical Trials*, 19 September 2002.

FDA/ICH, *Guidance for Industry: E09 Statistical Principles for Clinical Trials*, September 1998.

Friedman LM, Furberg CD & DeMets DL, *Fundamentals of Clinical Trials*, 4th Edition, Springer, 2010

Graham JW, *Missing Data Analysis: Making It Work in the Real World*, Annual Review of Psychology, 2009, 60: 549-576.

Grouin J-M, Coste M & Lewis J, *Subgroup analyses in randomized clinical trials: statistical and regulatory issues*, J Biopharm Stat, 2005; 15:869-82.

Hulley SB et al., *Designing Clinical Research*, Lippincott Williams & Wilkins, Philadelpha, PA, 2007.

References (3 of 7)

International Committee of Medical Journal Editors, *Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication*, updated April 2010, http://www.icmje.org/.

Ioannidis JPA et al., for the CONSORT Group, *Better reporting of harms in randomized trials: an extension of the CONSORT statement*, Annals of Internal Medicine, 2004, 141:781-788.

Jennison C & Turnbull BW, *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC, 2000.

Lachin JM*, A review of methods for futility stopping based on conditional power*, Statistics in Medicine, 2005, 24:2747-2764.

Lan KKG & Wittes J, *The B-Value: A Tool for Monitoring Data*, Biometrics, 1988, 44:579-585.

Marler JR, *Secondary Analysis of Clinical Trials – A Cautionary Note*, Progress in Cardiovascular Diseases, 2012, 54:335–337.

Moher D et al., The PRISMA Group, *Preferred reporting items for systematic reviews and meta-analyses, Public Library of Science Medicine*, 2009, 6(7):e1000097.

Moher D et al., *CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials*, BMJ, 2010;340:c869

References (4 of 7)

Motulsky H, Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking, Second Edition, Oxford, 2010

Moyé LA, *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators*, Springer, 2006.

Moyé LA, *Rudiments of Subgroup Analyses*, Progress in Cardiovascular Diseases, 2012, 54:338-342.

Naik AD et al., *Comparative Effectiveness of Goal Setting in Diabetes Mellitus Group Clinics Randomized Clinical Trial*, Archives of Internal Medicine, 2011, 171:453-459.

National Research Council, *The Prevention and Treatment of Missing Data in Clinical Trials, National Academies*, July 2010.

The Pediatric Eye Disease Investigator Group (PEDIG) public website contains links to all PEDIG study publications, including studies used as examples in this presentation: http://pedig.jaeb.org

Perlman AI et al., Massage Therapy for Osteoarthritis of the Knee: A Randomized Dose-Finding Trial, PLoS ONE, 2012, 7:e30248.

Petticrew M et al., *Damned if you do, damned if you don't: subgroup analysis and equity*, Journal of Epidemiology and Community Health, 2012, 66:95-98.

References (5 of 7)

Piaggio G et al., *Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement*, JAMA, 2006, 295:1152-1160.

Piantadosi S, *Clinical Trials: A Methodologic Perspective*, John Wiley and Sons, 2005.

Pocock SJ et al., Statistical problems in the reporting of clinical trials: a survey of three medical journals, New England Journal of Medicine, 1987, 317:426-432.

Pocock SJ & Ware JH, *Translating statistical findings into plain English*, Lancet, 2009, 373:1926-1928.

Proschan MA, Lan KKG & Wittes JT, *Statistical Monitoring of Clinical Trials: A Unified Approach*, Springer, 2006.

Proschan MA, *Sample size re-estimation in clinical trials*, Biometrical Journal, 2009, 51(2):348-357.

Rucci P et al., *Treatment-Emergent Suicidal Ideation During 4 Months of Acute Management of Unipolar Major Depression with SSRI Pharmacotherapy or Interpersonal Psychotherapy in a Randomized Clinical Trial*, Depression and Anxiety, 2011, 28:303-309.

Schulz KF, Altman DG, Moher D, CONSORT Group, *CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials*, PLoS Med, 2010, 7(3): e1000251. doi:10.1371/journal.pmed.1000251

References (6 of 7)

Sun X et al., *Is a subgroup claim believable? A user’s guide to subgroup analyses in the surgical literature*, J Bone Joint Surg Am, 2011; 93:e8(1-9).

Sun X et al., *Credibility of claims of subgroup effects in randomised controlled trials: systematic review,* British Medical Journal, 2012, 344:e1553 (Published 15 March 2012).

Thomas ML et al., *A Randomized, Clinical Trial of Education or Motivational-Interviewing-Based Coaching Compared to Usual Care to Improve Cancer Pain Management*, Oncology Nursing Forum, 2012, 39:39-49.

Vos SPF et al., *A randomized clinical trial of cognitive behavioral therapy and interpersonal psychotherapy for panic disorder with agoraphobia*, Psychological Medicine, 2012, April 30, 1-12 (epub).

Wakim P et al., *Relation of study design to recruitment and retention in CTN trials*, American Journal of Drug and Alcohol Abuse, 37:426–433, 2011.

Wang R et al., *Statistics in medicine – reporting of subgroup analyses in clinical trials*, New England Journal of Medicine, 2007, 357:2189-2194.

Wheelan C, *Naked Statistics: Stripping the Dread from the Data*, Norton, 2013.

Wikipedia, *Anscombe’s Quartet*, accessed on 11/30/2011.

References (7 of 7)

Woody GE et al., *Extended vs Short-term Buprenorphine-Naloxone for Treatment of Opioid-Addicted Youth: A Randomized Trial*, Journal of the American Medical Association, 2008, 300(17):2003-2011.

Yeh GY et al., *Tai Chi Exercise in Patients With Chronic Heart Failure: A Randomized Clinical Trial*, Archives of Internal Medicine, 2011, 171:750-757.

Zhu L, Ni L & Yao B, *Group Sequential Methods and Software Applications*, The American Statistician, 2011, Vol. 65, No. 2, 127-135.