Handling Baseline Variables in the Design and Analysis of Randomized Controlled Trials:

An Illustration of the Gap between Statistical Theory and Practice

Jody D. Ciolino, PhD
Northwestern University
Department of Preventive Medicine - Biostatistics

Northwestern Medicine

Context…

- Collaborative and applied: Biostatistics Collaboration Center (BCC)
- Part of the CTSA at Northwestern (NUCATS: NU Clinical and Translational Sciences Institute)
- My role:
  - Clinical trialist/study design specialist
  - Bridging the gap between theory and study design implementation
    - Education
    - Compromise between ideal and real

Northwestern University NUCATS Clinical and Translational Sciences Institute

Outline

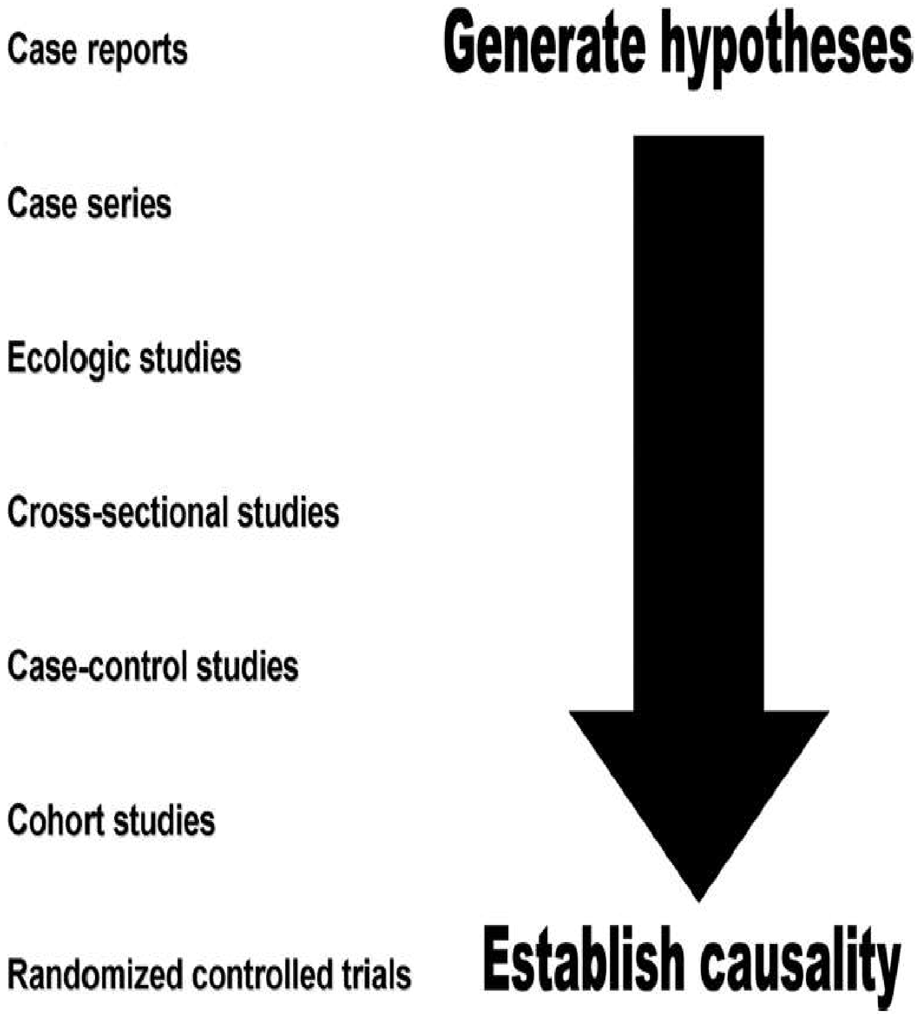
Handling Baseline Variables in Clinical Trials

- Motivation – the 'ideal'
  - Theory
  - Guidelines
- A snapshot of current practice (the 'real')
  - Systematic review of published randomized controlled trials (RCTs)
  - Findings and inferences
- Takeaway messages

Introduction

Why are confounders still a problem?

- We prefer a 'randomized' trial to observational studies
- Randomness: ***on average***, our study arms are 'similar'
  - Measured and unmeasured variables
  - Allows for what we hope is unbiased assessment of intervention effects
- BUT we can only state that expected level of imbalance on all baseline variables = zero
  - *i.e., on average we have 'similar enough' groups where confounding is most likely not an issue*
  - **This means that under purely random assignment, there is a possibility that nontrivial imbalances occur**



Northwestern Medicine®

Some theory…

Why are confounders still a problem?

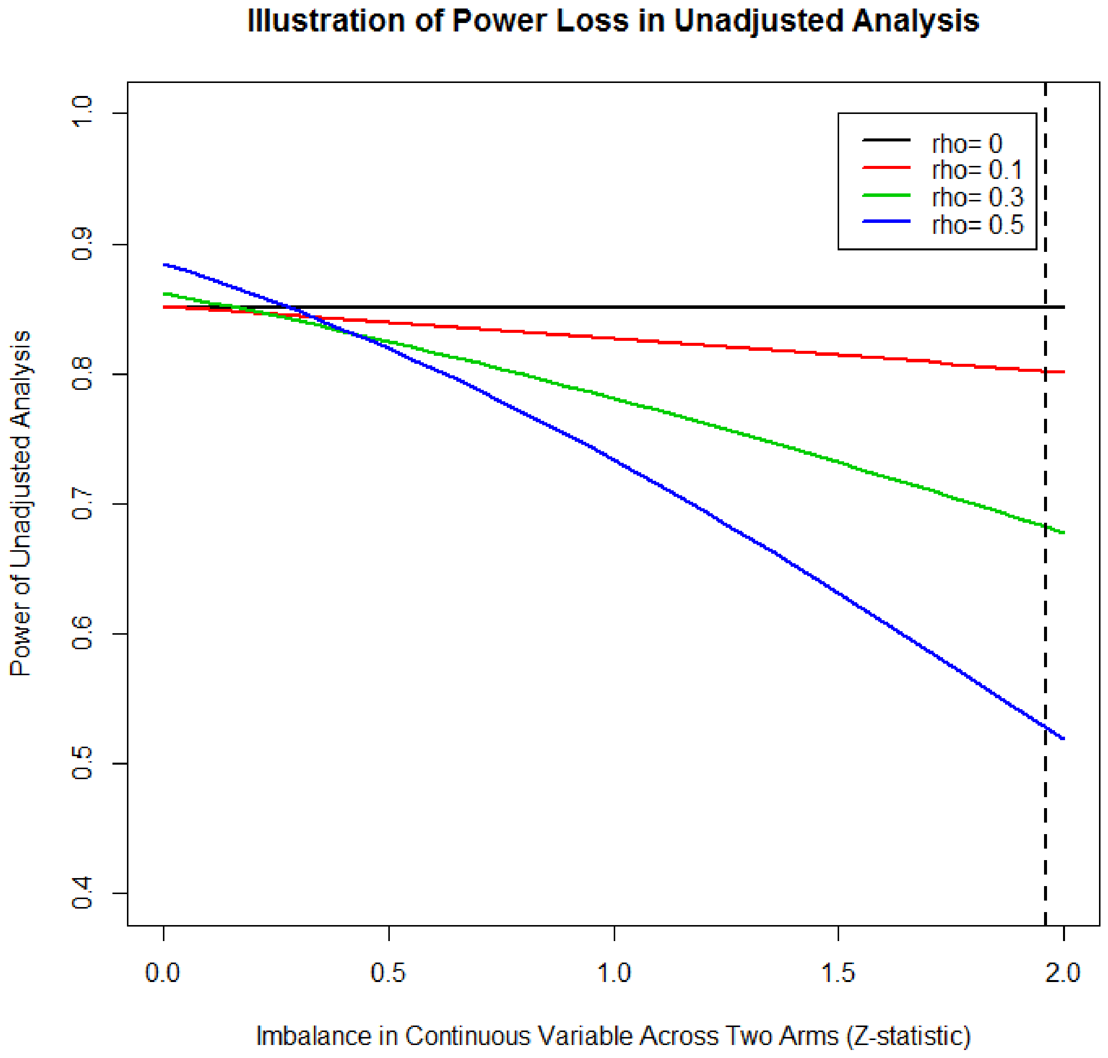
Chance imbalances can affect:

- **Power**
- Type I error rate
- Bias in treatment effect estimates (over/underestimation is possible)

$$\gamma(d_x) = prob\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{d_x^*\rho}{\sqrt{1-\rho^2}} - \frac{\Delta}{\sqrt{(1-\rho^2)\sigma_y^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right]$$

$$= prob\left[Z \geq \frac{Z_\alpha - d_x^*\rho - d_y^*}{\sqrt{1-\rho^2}}\right]$$

Senn, 1989; Ciolino et al., 2011



Imbalance across two arms favoring control arm→

[rho = cor(baseline variable, outcome)]

Northwestern Medicine

Why are confounders still a problem?

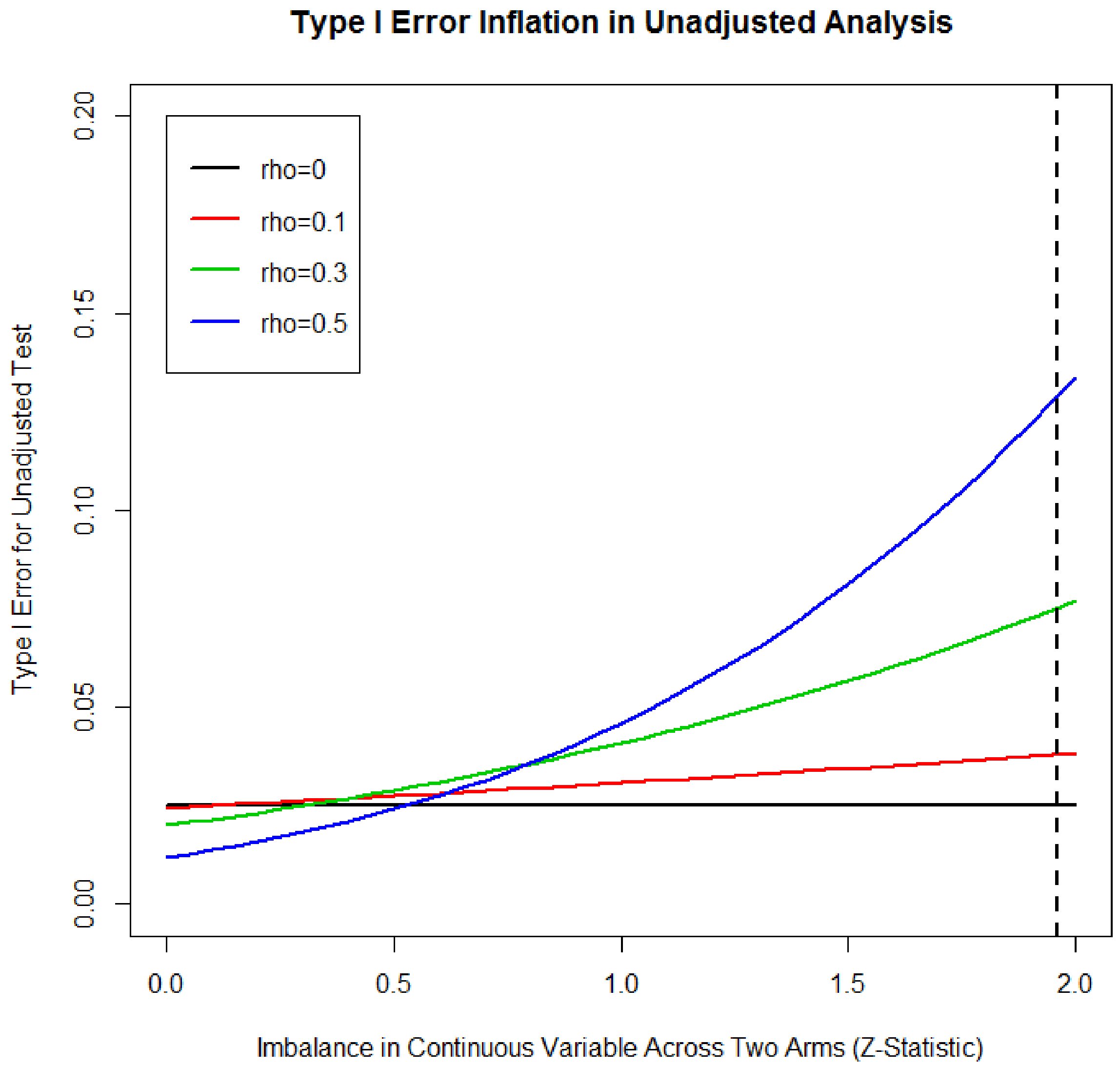
Chance imbalances can affect:

- Power
- **Type I error rate**
- Bias in treatment effect estimates (over/underestimation is possible)

$$\alpha(d_x) = prob\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho d_x}{\sqrt{(1-\rho^2)\sigma_x^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right]$$

$$= prob\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho d_x^*}{\sqrt{1-\rho^2}}\right]$$

Senn, 1989; Ciolino et al., 2011



Imbalance across two arms favoring active arm →

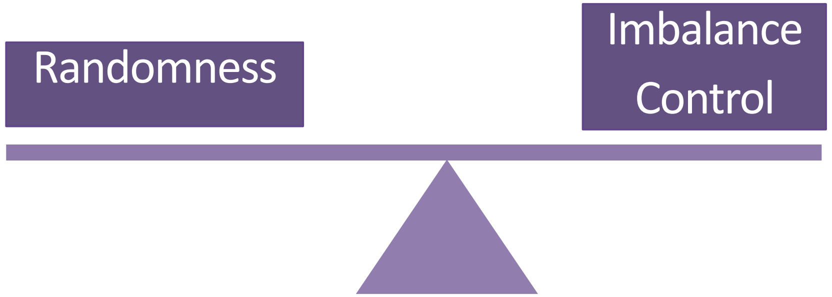
[rho = cor(baseline variable, outcome)]

Northwestern Medicine®

The Ideal…

What *should* we be doing about these variables?

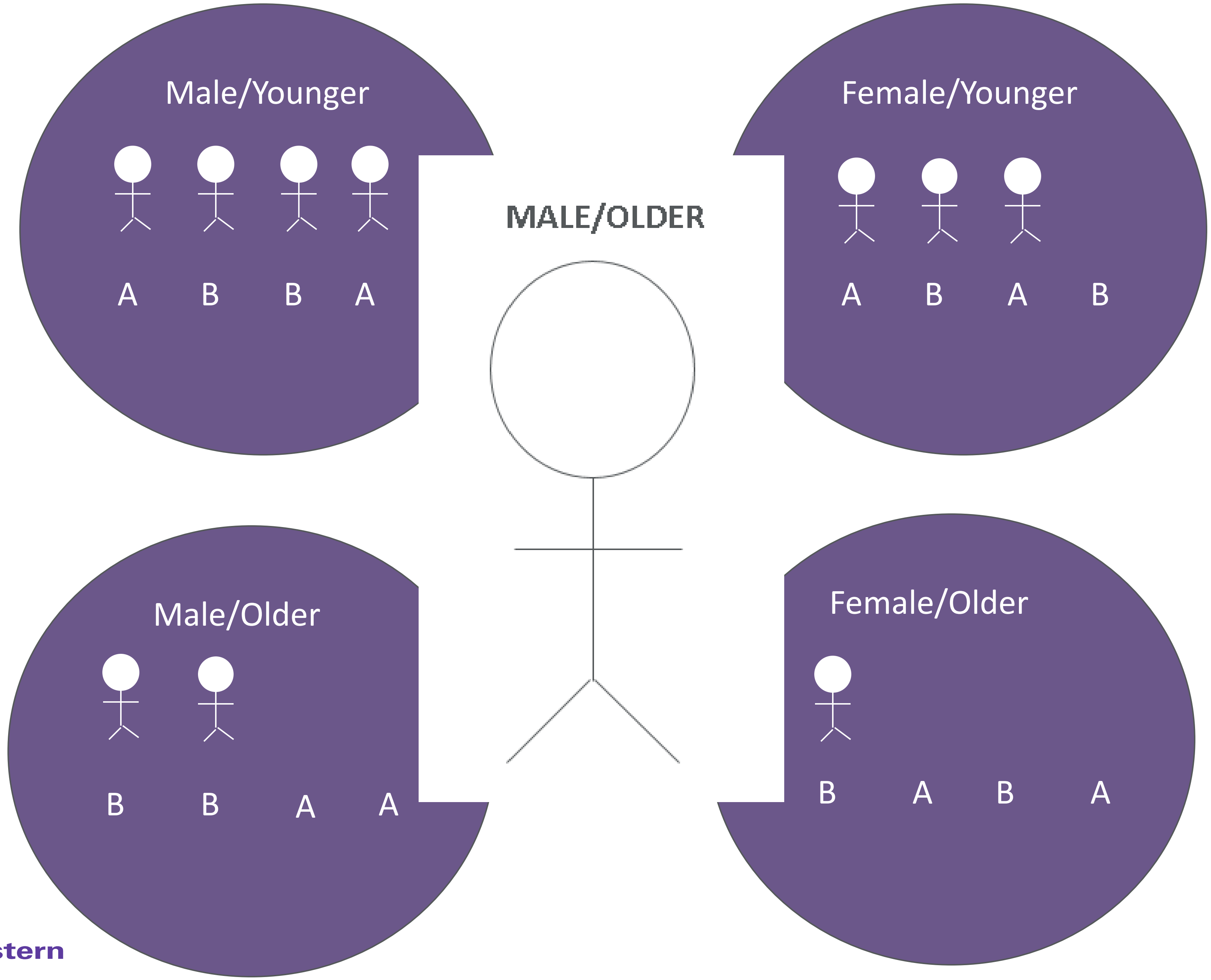
At the Beginning (design)



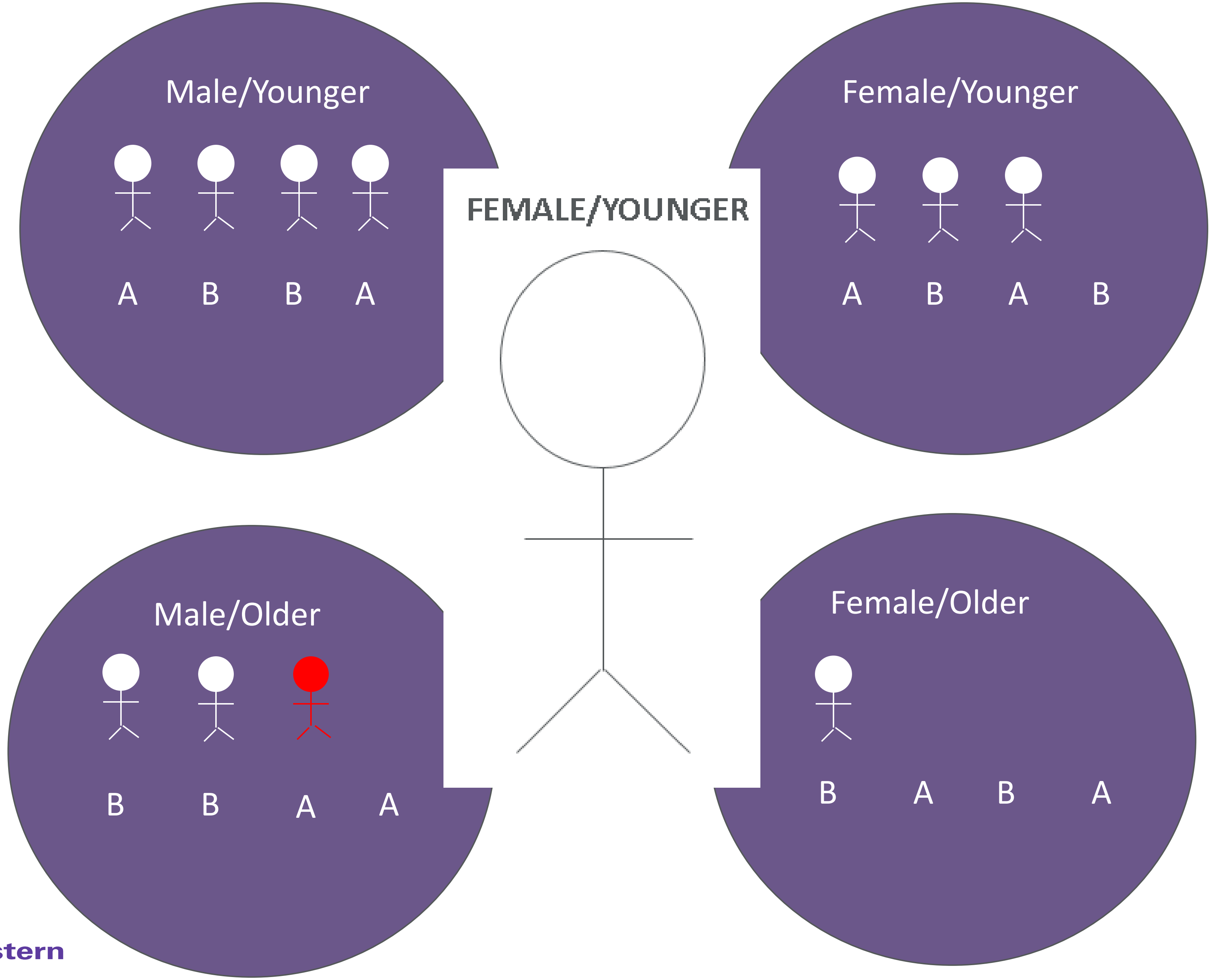
- Many options with regard to randomization or treatment allocation scheme
  - Simple random allocation, random or permuted block, urn designs, etc.
  - **Stratified or stratified block**
  - **Adaptive techniques (e.g., minimization, minimal sufficient balance, etc.)**
- Which one is ‘best’ depends on scenario of the trial…
  - In general, the most flexible designs tend to be the adaptive designs
  - A brief review of designs follows…

Stratified Block Design

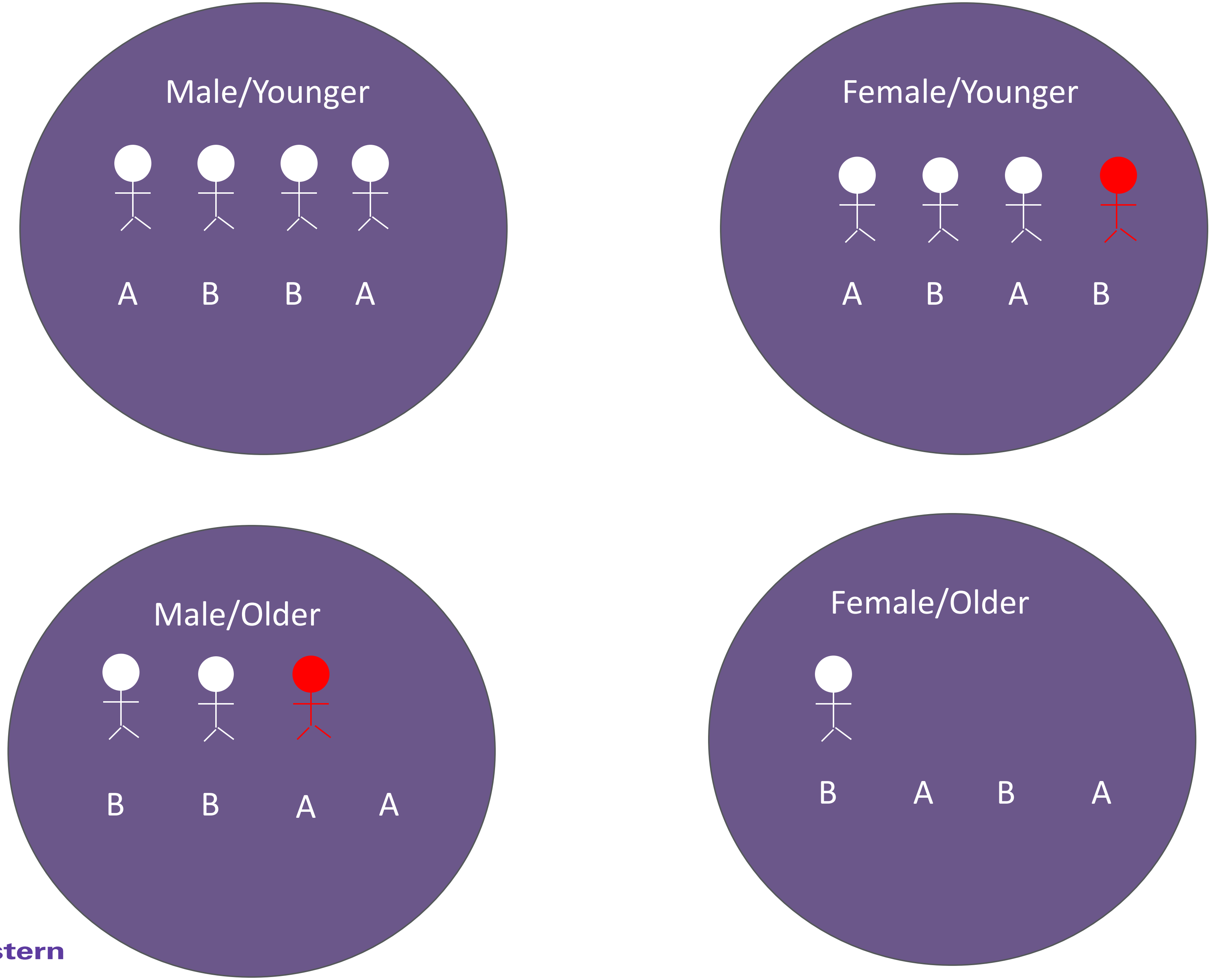
- Most commonly used method for attempting to balance covariates
- Uses blocking within strata of influential covariates
- Example: **Gender** (M/F) and **Age** (older/younger) = important predictors
- We have four strata:
  - Older males
  - Older females
  - Younger males
  - Younger females
- Within each stratum, apply the blocked design



Northwestern Medicine®



Northwestern Medicine


Male/Younger
A B B A
Female/Younger
A B A B
Male/Older
B B A A
Female/Older
B A B A

Drawbacks of Stratified Block Design

- What if we stop the trial now?
  - Unfilled blocks: Male/Older and Female/Older have unfilled blocks
  - How do we really know that we are balancing age? Must categorize continuous variables

- As number of strata increases, performance = similar to simple randomization
- Example: Clinical center (assume 5), Gender (2 categories), age (4 categories: 21-30,30-35,36-40,>40 years), baseline disease status (mild, moderate, severe)
  - → **Each center** has 2 x 4 x 3 = 24 strata that need to be balanced!
  - Thus, 5 x 24 = 120 strata total!
  - Requires pre-generated lists: may be electronic, sealed envelopes, pharmacy houses list, etc. → opportunity for error
- Issues re: unfilled blocks and categorization are magnified

Covariate-Adaptive Methods

- AKA ‘minimization’ (Taves, Simon, Pocock [1970s])
- Choose imbalance function to minimize (range, variance) for each variable

(D_i, *i=1,…,# variables*)

- Weight each variable wish to balance (w_i)
- Let overall imbalance = $D = \sum w_i D_i$
- For incoming subject, calculate *D* under assignment to each possible arm
- Assign subject to arm with smallest *D* with higher probability (0.5,1]
- Well known*, less commonly implemented than stratified block
- More recent methods can handle both categorical and continuous variables (e.g., Minimal sufficient balance [Zhao et al., 2012])

Minimization: Example

- Incoming subject = Male, BMI <30 kg/m^2, Cholesterol >6.0 mmol/l
- Use **'range'** as measurement of imbalance
- Use **equal weight** for each of these variables
- Assign to treatment A:

  Imbalance=|5-5|+|5-3|+|4-2|=4
- Assign to treatment B:

  Imbalance=|4-6|+|4-4|+|3-3|=2
- Minimize imbalance by assigning to treatment B
- Use probability of assignment to B = (0.50, 1]

Features of 17 Subjects Entered Into a Trial of Obesity

Variable	Category	A	B
Sex	Male[1]	4	5
	Female	4	4
BMI (kg/m^2)	<30[1]	4	3
	≥30	4	6
Fasting cholesterol (mmol/l)	≤6.0	5	7
	>6.0[1]	3	2
Total number of subjects already allocated		8	9

1. Values for next subject to be allocated.

McEntegart 2005; Drug Information Journal

Minimization/Covariate-Adaptive Methods

- More flexible: adaptive, weighting, more covariates, differing variable types (categorical, continuous, etc.)
- More difficult to guess treatment assignment when balancing several covariates
- Does not handle imbalance as well as stratified block in presence of interactions
- **Complex: requires algorithmic feedback on ongoing basis**
  - **Interactive voice response**
  - **Web-based**
  - **Need to consider: back-up, speed of process, 24-hour availability**
- Taves (2010) reports <2% of published randomized clinical trials use minimization

Back to the question: what *should* we do at design?

- Think about potential confounders at the beginning of the trial
- Attempt to control imbalance to avoid impact on (unadjusted) analyses
- Consider **covariate adaptive techniques**

> "With modern technologies such as IVR and IWR,
> generation of a randomization sequence takes little
> time and effort but affords big rewards in scientific accuracy and credibility."
> (Lin et al. 2015; *Contemporary Clinical Trials*)

- Instances where variables are unknown or few…stratified block randomization (or simple) may be acceptable; just keep limitations in mind

What about at the end of the study?

- Good news!
- ***Appropriate*** adjustment ***often*** solves many of the statistically-related problems (Ciolino et al. 2011, 2014; Raab and Day 2000; Ford and Norrie 2002)
  - Increases precision on treatment effect estimate
  - Decreases bias in treatment effect estimate
  - → tends to preserve type I error rate and power
- Bad news?
  - We can't adjust for everything
  - **Sometimes the benefit of adjusted analyses depends heavily on nature of outcome and magnitude/direction of imbalance** (*Gail et al. 1984; Greenland 1999; Hauck et al. 1998; Ciolino et al. 2013)*
    - Binary outcome/nonlinear relationships
    - Precision may decrease and unadjusted estimates ≠ 'adjusted' estimates (See Steingrimsson et al. 2017)

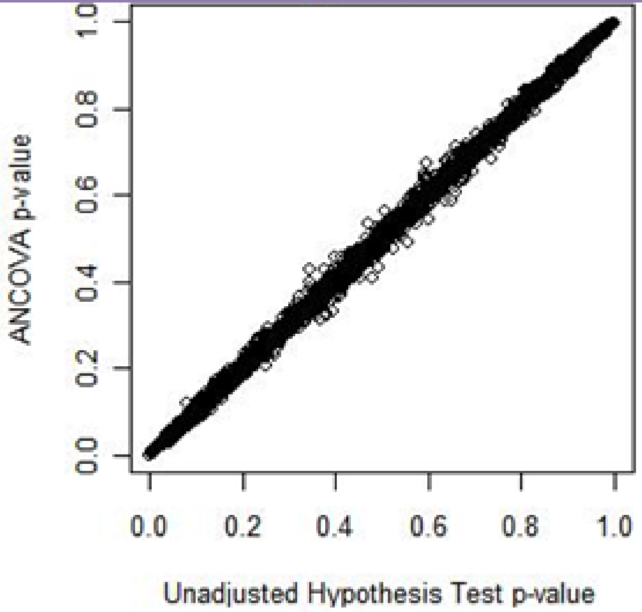
At the End (Analysis)

- When in doubt, adjust
- CONSORT (2009):
  - Adjustment may be 'sensible, especially if one or more variables is thought to be prognostic' (Journal of Clinical Epidemiology, 2010)
  - Ideally…pre-specified in the protocol or analysis plan

Continuous Outcome:

Adjusted vs. Unadjusted p-value

ZERO correlation w/baseline variable and outcome



Binary Outcome:

(Simulated data; Ciolino 2013)

β_x	Unadjusted power	Adjusted power	Benefit	Unadjusted bias
$-0.6\tilde{\beta}_{tx}$	76.78%	79.98%	3.20%	–2.5%
$-1.0\tilde{\beta}_{tx}$	66.12%	75.52%	9.40%	–5.3%
$-1.5\tilde{\beta}_{tx}$	48.92%	66.46%	17.54%	–9.2%

At the End (Analysis)

- What we *should not* be doing:
- Allow baseline test for significant differences to dictate adjustment (Senn, Ciolino et al., CONSORT)
- Failing to pre-specify or to transparently explain post hoc decisions to adjust
- CONSORT (J Clinical Epidemiology, 2010) →

Table 1.

Charact	P Value
Age —	0.84
Sex —	0.77
Mal	
Fem	
Race or	0.60
Asia	
Blac	
His	
Whi	
Oth	
Not	

- "Unfortunately significance tests of baseline differences are still common…"
- "[these tests]…assess the probability the observed baseline differences could have occurred by chance; however, we already know that any differences are caused by chance."
- "illogical", "superfluous", and misleading
- "…comparisons at baseline should be based on consideration of prognostic strength *and* the size of any chance imbalances."

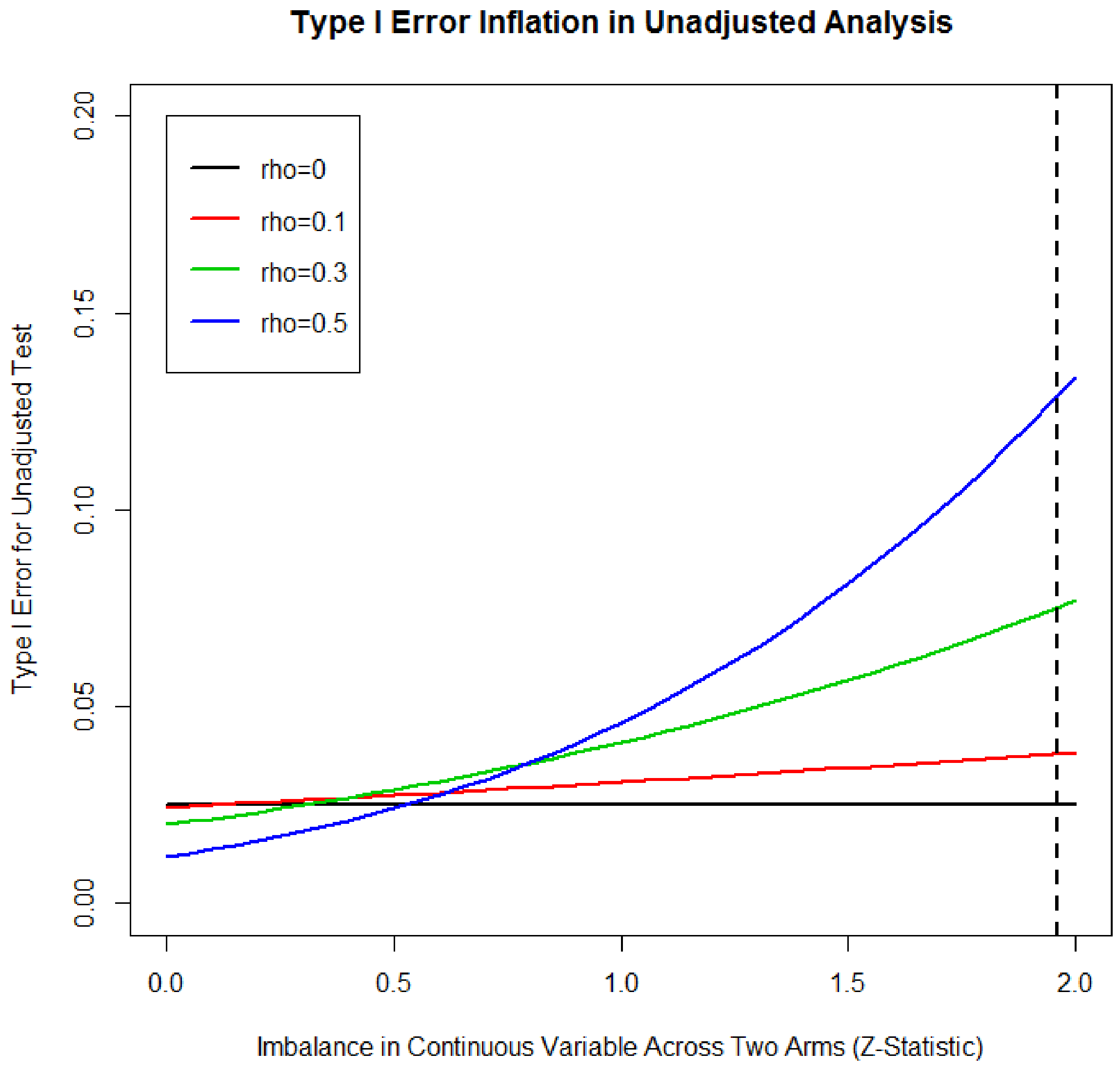
Recall…

Chance imbalances can affect:

- Power
- **Type I error rate**
- Bias in treatment effect estimates (over/underestimation is possible)

$$\alpha(d_x) = prob\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho d_x}{\sqrt{(1-\rho^2)\sigma_x^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right]$$
$$= prob\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho d_x^*}{\sqrt{1-\rho^2}}\right]$$

Senn, 1989; Ciolino et al., 2011



Imbalance across two arms favoring active arm →

[rho = cor(baseline variable, outcome)]

Northwestern Medicine®

A Snapshot of Current Practice in RCTs

Systematic Review of Reported Methods of Handling Baseline Variables in Published RCTs

Objectives:

1. Explore the frequency of use for each allocation scheme type in published RCTs.
2. Explore the handling of covariates in the analysis phase in published RCTs.

Methods

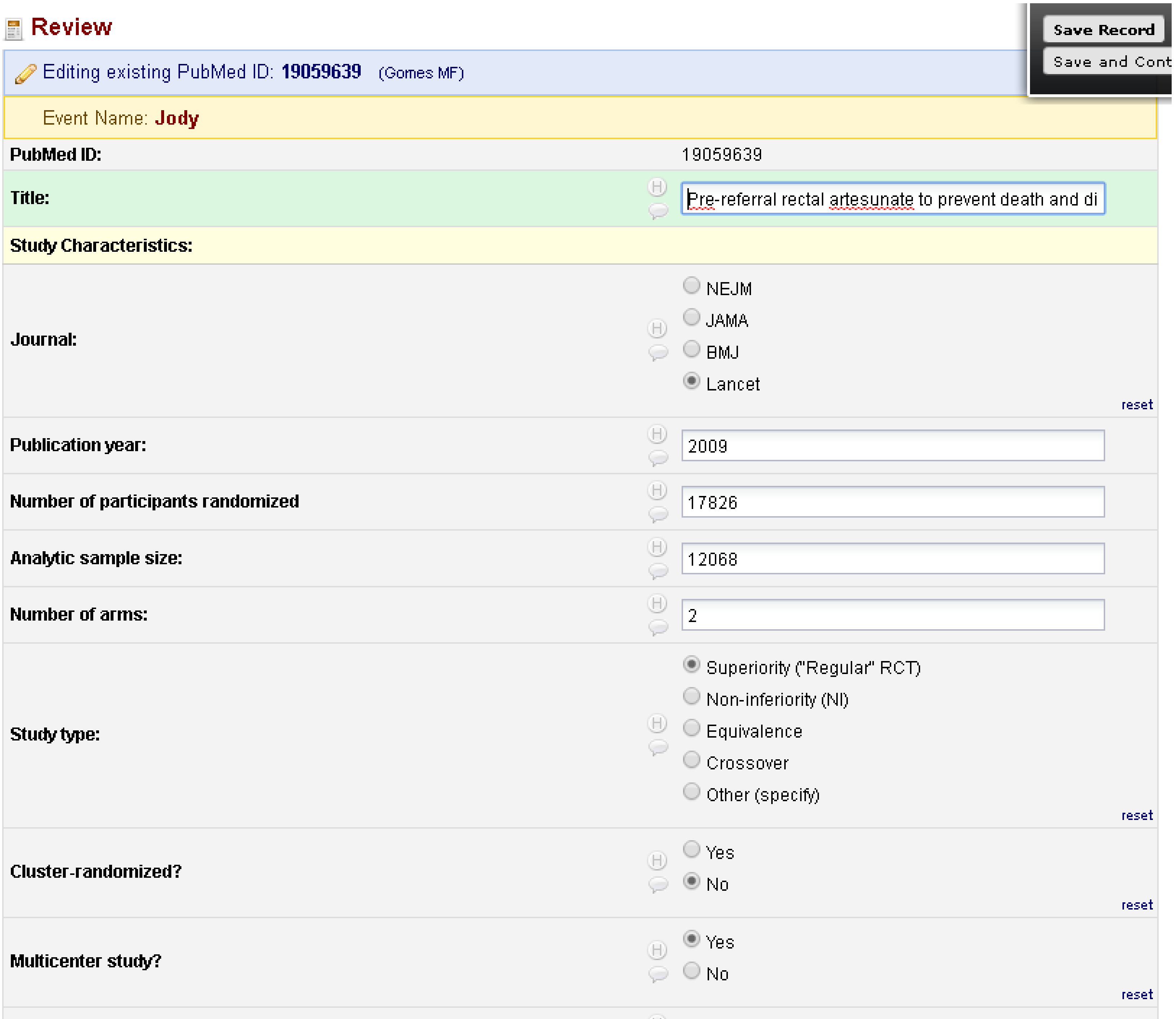
Methods:

- Search PubMed for articles indexed as "RCT" in *NEJM*, *JAMA*, *BMJ*, *Lancet*
- Two time periods: 2009 (before updated CONSORT); 2014 (five years later)
- Extracted trial characteristic variables and

1. **Covariate involvement** in randomization (binary variable: yes vs. no/unable to determine)
2. Use of **adjustment** vs. no adjustment in analyses (binary)
3. Use of **covariate-adaptive** techniques (binary) for allocation (within a subset of trials)
4. Whether adjusted analyses were **pre-specified** (within a subset of trials)

Data Capture: REDCap

(Research Electronic Data Capture)


Review
Save Record
Save and Cont
Editing existing PubMed ID: 19059639 (Gomes MF)
Event Name: Jody

PubMed ID:	19059639
Title:	Pre-referral rectal artesunate to prevent death and di
Study Characteristics:	
Journal:	○ NEJM ○ JAMA ○ BMJ ◉ Lancet reset
Publication year:	2009
Number of participants randomized	17826
Analytic sample size:	12068
Number of arms:	2
Study type:	◉ Superiority ("Regular" RCT) ○ Non-inferiority (NI) ○ Equivalence ○ Crossover ○ Other (specify) reset
Cluster-randomized?	○ Yes ◉ No reset
Multicenter study?	◉ Yes ○ No reset

Northwestern Medicine®

343 Articles Identified through PubMed Search:

(randomized controlled trial[Publication Type] AND ("N Engl J Med"[Journal] OR "JAMA"[Journal] OR "BMJ"[Journal] OR "Lancet"[Journal]) AND (("2009/01/01"[PDAT] : "2009/06/30"[PDAT]) OR ("2014/01/01"[PDAT] : "2014/06/30"[PDAT])))

↓

45 Articles Excluded:

7 = Not an RCT
5 = Research letter/comment/editorial
19 = Secondary analysis
13 = Reporting on multiple trials
1 = Other (country policy change RCT not fitting mold)

↓

298 Articles Included in Full Review

102 (34%) from *NEJM*
59 (20%) from *JAMA*
38 (13%) from *BMJ*
99 (33%) from *Lancet*

131 (44%) from 2009
167 (56%) from 2014

Northwestern Medicine®

Summary of Findings – Typical trial

- Two-armed (79%), multicenter (92%), superiority (86%)
- Lasting for a median of three years with median 12 months of follow-up
- Stratified block method of allocation (69%) with accompanying analysis that tended to adjust (84%) for baseline variables

Snapshot of Practice - Design

Allocation Method	Overall N (%)	2009 N (%)	2014 N (%)
Purely random	4 (1)	3 (2)	1 (1)
Permuted/Random Block	24 (8)	9 (7)	15 (9)
Stratification/ Stratified Block	**205 (69)**	**82 (63)**	**123 (74)**
Covariate Adaptive	**32 (11)**	**18 (14)**	**14 (8)**
Other	4 (1)	2 (2)	2 (1)
Unable to determine	29 (10)	17 (13)	12 (7)

Overall, 81% of studies included baseline variables in allocation scheme →

Potentially influential study characteristics:

- Longer studies (p=0.016)
- ***Fewer arms*** (p=0.025)
- Multicenter (p=0.021)
- Time-to-event outcome (p=0.005)

Northwestern Medicine®

Snapshot of Practice – Design

Number of Baseline Variables Included in Randomization	Overall N (%)	2009 N (%)	2014 N (%)
1	95 (39)	42 (41)	53 (38)
2	86 (36)	32 (31)	54 (39)
3	40 (17)	17 (17)	23 (17)
4	11 (5)	5 (5)	6 (4)
5 or more	9 (4)	6 (6)	3 (2)

Overall, 81% of studies included baseline variables in intervention allocation

Snapshot of Practice - Analysis

Primary Analyses	Overall N (%)	2009 N (%)	2014 N (%)
Unadjusted Only	49 (16)	27 (21)	22 (13)
Adjusted Only	87 (29)	27 (21)	60 (39)
Both	162 (54)	77 (59)	85 (51)

- 91% (226) pre-specified (or gave benefit of the doubt)
- **43%** (126) report statistical test for significant differences in baseline variables

Northwestern Medicine®

Snapshot of Practice – Some Interesting Findings

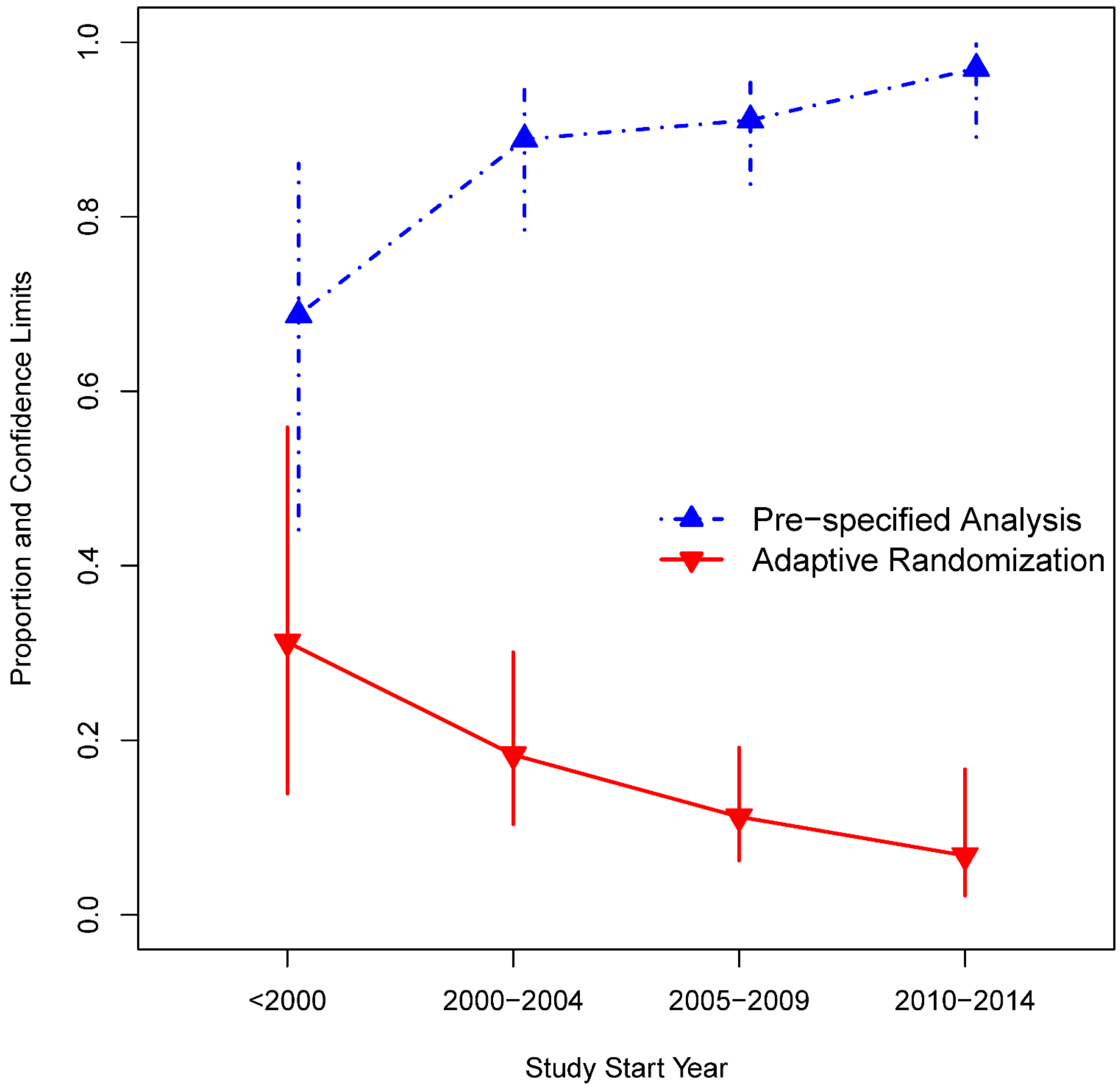
- **Adaptive allocation techniques:**
  - **(-) increasing study start year (p=0.005; OR=0.89 [0.82,0.96])**
    31% initiated before 2000 vs. just 7% 2010 or later
  - (+) increasing number of baseline covariates in randomization
    (p=0.031; OR=4.92 [2.99,8.09])
  - (+) increasing study length (p=0.040; OR=1.11 [1.00,1.24])

- **Pre-specified adjusted analyses:**
  - (+) increasing study start year (p=0.014; OR=1.12 [1.02,1.22])
    69% before 2000 vs. 97% 2010 or later
  - (+) multicenter (p=0.046; OR=3.45 [1.02,11.62])

Are we progressing?

+ direction for pre-specified analyses

- direction for adaptive randomization methods



Northwestern Medicine®

Summary of Findings – The Positives

- *Typical trial*
  - *Two-armed, multicenter, superiority*
  - *Lasting for a median of three years with median 12 months of follow-up*
  - *Stratified block method of allocation with accompanying analysis that tended to adjust for baseline variables (may not be the same that were used in allocation)*
- **Positive progress**
  - Dominant use of baseline variables in design (81%) and analysis (84%)
  - Largely pre-specified adjusted analysis (91%), with increasing prevalence of pre-specification over time
  - Adjusted analyses associated with covariate involvement in randomization (p=0.010) and increasing number of covariates (p=0.031)
  - Increased number of covariates associated with use of adaptive methods (100% with at least five variables, p<0.001)

Summary of Findings – Identifying Gaps

Areas of potential gap between ideal/theory/guidelines and practice/real

- Dominant use (69%) of stratified block despite shortcomings
- 11% employ covariate-adaptive methods, with less prevalence over time
- Less involvement of baseline variables in general as the number of arms increases:
  - Two arms: 83% involved at design vs. five or more arms: 58%
  - None of five- (or more) armed trials used adaptive methods
- "substantial and confusing variation…in handling baseline covariates" (Austin et al. 2010)
  - 10% of the time unable to determine allocation technique
  - 'unclear' as high as 23% of the time (may be related to number of arms/trial complexity)
  - Superfluous test of baseline differences in 43% of trials (similar to 38% in review by Austin et al. in 2010)

Northwestern Medicine®

Why the gap?

Some Anecdotes

Common questions/comments from collaborators when questioned about baseline variable relevance for their outcomes:

- Shouldn't the randomization take care of it?
- On average, yes; there is no guarantee (*every* trial will exhibit some baseline variable imbalance)
- There are no 'significant differences' at baseline, so we don't need to worry (our randomization 'worked')
- Not necessarily (even 'insignificant' imbalances have an impact [if we fail to adjust] on analyses)
- We stratified, so these variables should be balanced
- See above + stratification may not always help the cause

Some Anecdotes, cont'd

Common questions/comments from collaborators when questioned about baseline variable relevance for their outcomes:

- Can't we just adjust for these in analyses?

- Yes, but…
  - What about face validity?
  - What if we have two many variables for which we'd like to adjust?
  - We can't adjust for everything nor do we know all influential variables ahead of time
  - Unadjusted effect ≠ 'adjusted' effect

Why the gap?

- Anecdotal evidence suggests lack of education/understanding
  - Over-simplification of design (‘it’s just a simple/small trial’)
  - Poor planning/time commitment to design and a pre-specified analysis plan
  - Sometimes a ‘black box’ issue
- Programming/software requirements and expense
- Lack of statistician or programmer involvement from beginning to end
- Individual trial logistical complexities overpower design and analysis considerations

Some Takeaway Messages

- We should be thinking about baseline variables in design and analysis phase of RCTs
  - Complex methods of randomization and/or analyses have potential to increase efficiency and reduce bias in intervention effect estimation
  - BUT these methods are often misunderstood or simply not used
- Increased education and collaborative efforts can help mitigate these gaps
- Sometimes practical constraints simply cannot be avoided
  - Something can (and will) always come up
  - We cannot predict everything with 100% certainty when designing a study
  - In these situations: critical thinking (‘trickle down effects’); involvement of a statistician throughout; compromises between ideal and real; transparency in reporting

Thank you!

jody.ciolino@northwestern.edu

(references available upon request)

Northwestern Medicine