Statistically Speaking Lecture Series

Sponsored by the Biostatistics Collaboration Center (BCC)

Protocol Development and Review from a Biostatistical Perspective

Jody D. Ciolino, PhD

Associate Professor

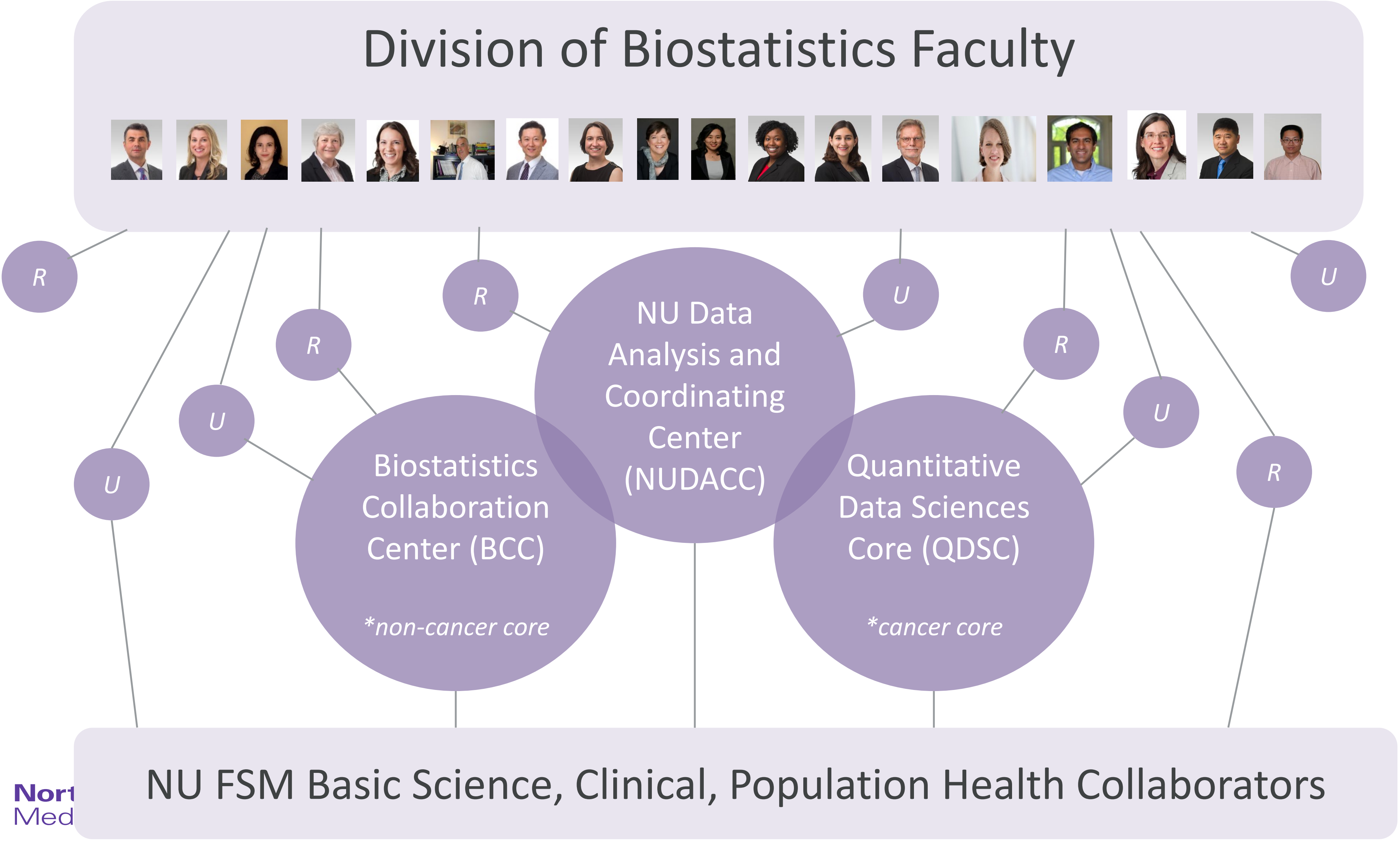
Department of Preventive Medicine-Biostatistics

Biostatistics Collaboration Center (BCC)

Northwestern University Data Analysis and Coordinating Center (NUDACC)

jody.ciolino@northwestern.edu

NU FSM Department of Preventive Medicine - Division of Biostatistics



Division of Biostatistics Faculty

NU Data Analysis and Coordinating Center (NUDACC)

Biostatistics Collaboration Center (BCC)

*non-cancer core

Quantitative Data Sciences Core (QDSC)

*cancer core

NU FSM Basic Science, Clinical, Population Health Collaborators

Slide credit: NUDACC

Conflicts of interest

No conflicts to disclose.

Disclaimer:

The views to follow do not represent all statisticians' perspectives. The opinions and advice to follow reflect those of the presenter only and should not be construed as a representation of the statistical community at large nor Northwestern University / Feinberg School of Medicine.

Northwestern Medicine®

Today's goals:

1. Present basic statistical concepts to keep in mind for any research study.
2. Illustrate the key elements of any protocol or project proposal that require biostatistical thought / input.
3. Promote sound, rigorous, reproducible research for researchers at FSM and beyond.

Outline

- Introduction
- Key deliverables/components
  - **Objectives and hypotheses**
  - **Outcomes**
  - **Sample size**
  - *Data management*
  - *Analysis plan*
- Final message

Introduction

- Good design → good science
- Protocol development is a critical piece in translational research
- ***Bad analysis can be redone, bad design and conduct cannot be redone***
- When it comes to translational research, no matter what the study type, there are some recurring themes and ideas

Introduction

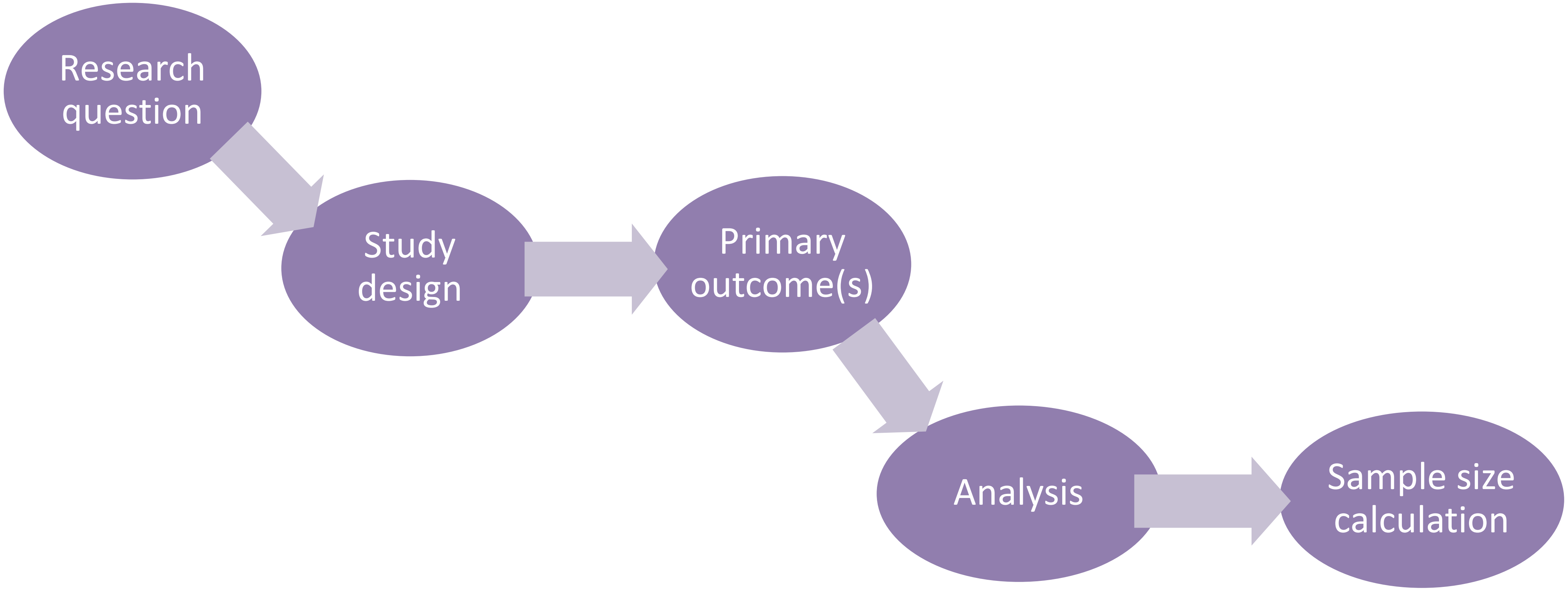
- We’ll focus on design, but protocol should include data integrity and management and high-level analysis approach
- We’ll focus on clinical trials, but a lot of the same concepts apply for basic science and translational studies

Northwestern Medicine®

Key Components of Protocol Development

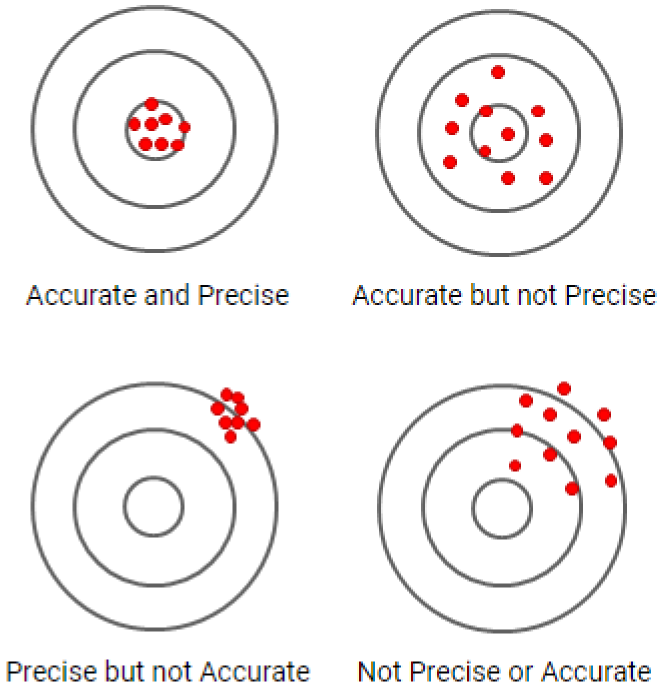
Key Components of Protocol Development

Biostatistical perspective is not simply meant to provide an ‘N’ or a ‘p-value’
Statistical thinking must occur throughout the entire study



Statistical perspectives…

- In general, there are two recurring themes in statistics:
  - Bias
  - Variability
- Both are a “problem” – they make it difficult to estimate underlying parameters with ***accuracy*** and ***precision***



Accurate and Precise

Accurate but not Precise

Precise but not Accurate

Not Precise or Accurate

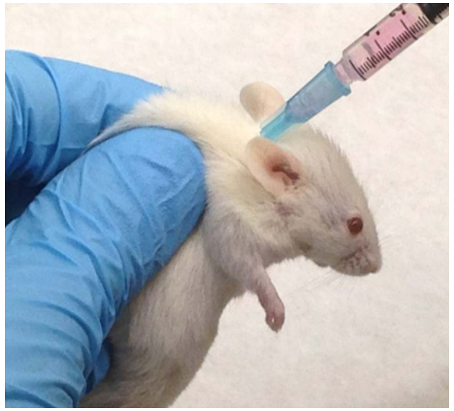
Bias and Variability

Bias

- Results in inaccuracy
- Systematic error
- Examples:
  - Unrepresentative sample
  - Uncalibrated instrument
  - Unfair “advantage” in one randomized arm
  - Unfair (dis)advantage at a clinical site

Variability

- Results in imprecision (more noise)
- Heterogeneity within a sample
- Examples:
  - Moving from “bench” to “bedside”
  - Phase I → Phase II → Phase III trials
  - Adding clinical sites
  - Relaxing inclusion/ exclusion criteria



Key components of any protocol include

- Objectives and hypothesis
- Measurements and outcomes
- Sample size
- Data management
- Analysis plan

Keep in mind … Research question → study design → primary outcome(s) → analysis → sample size calculation

Northwestern Medicine®

Example: A complex, cluster-randomized, non-inferiority study

Patient-Centered Outcomes Research Institute (PCORI) Award (AD-1507-31473). The views, statements, and opinions in this presentation are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

My collaborator…

"I want to conduct a **non-inferiority study** to show that my intervention **delivered by paraprofessional** home visitors (HV) is similar in preventing postpartum depression **when compared to mental health (MH) professionals**."

Darius Tandon, PhD

Associate Professor, Medical Social Sciences

Associate Director, Center for Community Health

Northwestern Medicine®

Me…

(skeptical)

My collaborator…

"The study needs to be **cluster-randomized** because we need intervention at the site level."

"We also need to have a **control arm** (we need to test superiority of the intervention too)."

Darius Tandon, PhD
Associate Professor, Medical Social Sciences
Associate Director, Center for Community Health

Me…

(skeptical)

"A little more information, please…**What is the research question?**"

The MB Study

Some background…

- Mothers and Babies (MB) intervention at home visiting (HV) sites in the Midwest Region
- Goal = promote perinatal mental health and well-being through MB
- MB = group intervention, six sessions of the MB course, perinatal + postpartum

http://www.mothersandbabiesprogram.org/

The MB Study

Some background…

- Previously, MB delivered by mental health (MH) professionals
- Previous studies suggest efficacy of this intervention
- BUT, would be more cost effective/efficient to have paraprofessionals deliver MB

http://www.mothersandbabiesprogram.org/

The question…

There are many, but…

1. Is MB delivered by **home visiting paraprofessionals (HVP) effective** in reducing **depressive symptoms at six months postpartum** when compared to **usual home visiting services** among low-income women?

2. Is MB delivered by **HVP "just as good as" (not inferior to) MB delivered by Mental Health Professionals (MHP)** in reducing **depressive symptoms** at six months postpartum among low-income women?

From these questions, we can start to formulate our study design…

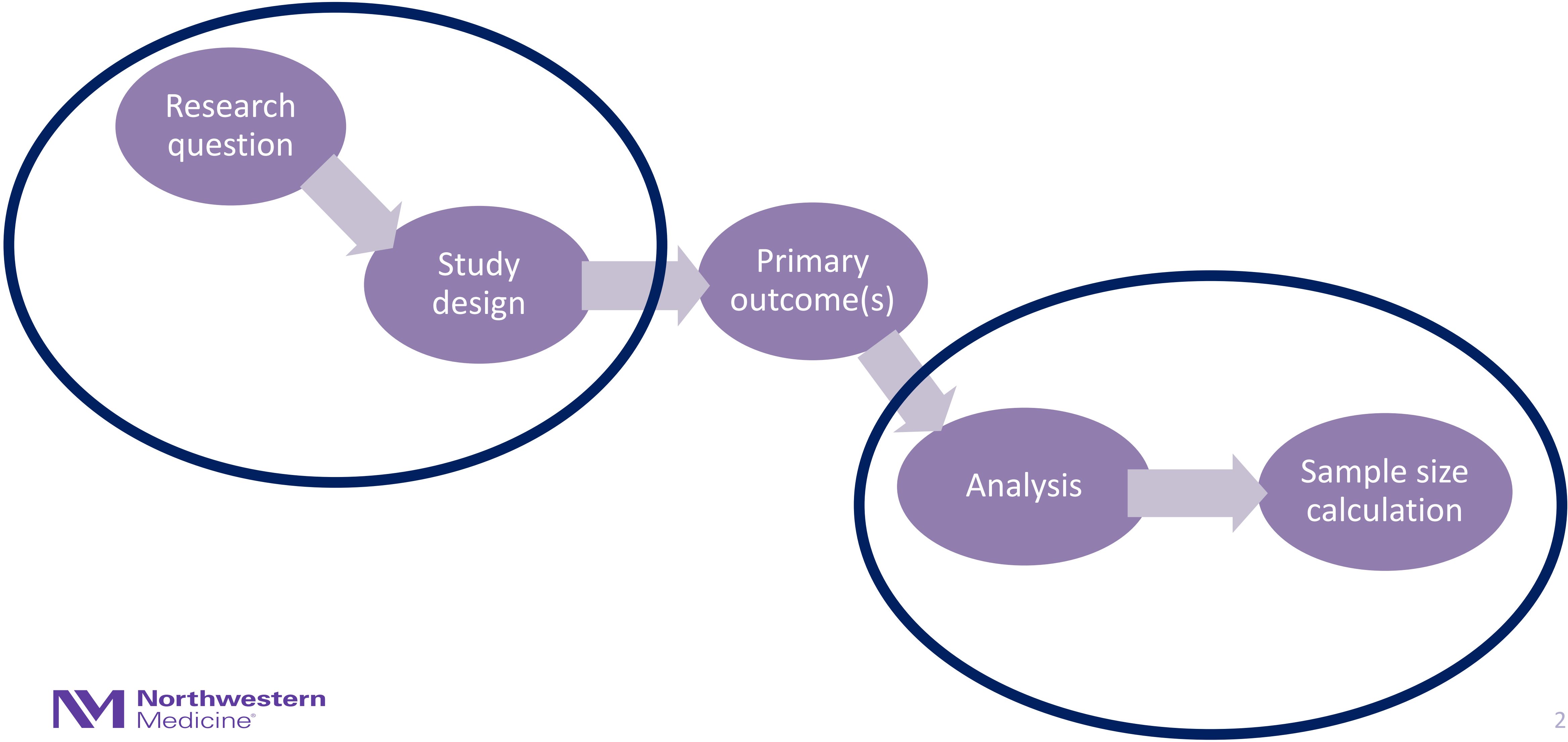
The design…

The options are endless

- **Three** arms
  - MB delivered by MHP
  - MB delivered by HVP
  - Control arm (usual HV activities)
- **Added complexities**…
  - We cannot randomize individuals to this group-based therapy + each site will be randomized to just one of these arms → **Cluster randomization**
  - We want sites to have a “good chance” of being randomized to *an* intervention (MHP or HVP) → **1:3:3 allocation**

Statistically Speaking…

The question and study design already start to complicate analyses and power/sample size considerations



Cluster-Randomized Design

- Must consider intra-cluster or intra-class correlation
  - Are individuals within a within a site or group likely to be more similar for some reason?
  - If so, this creates a non-zero intra-cluster correlation
- Small clusters relative to total N → similar to individual randomization
- In general, if ICC is large → problems for sample size calculations
- The larger the ICC, the larger the required sample size inflation

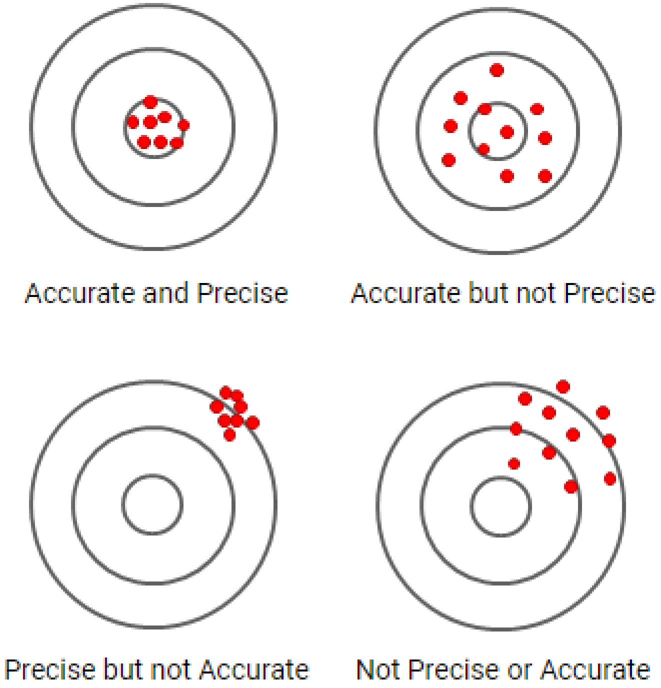
$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Between-cluster variability

Total (between + within) variability

Statistical Issues in the Design of the MB Study

- Cluster-randomized studies are more prone to biases
  - When analyzing participant-level data, we need to think about potentially similarities in participants within a site
  - The measure of similarity of participants within a site is known as “intra-cluster correlation” (**ICC**)
- What if all the **rural** sites were allocated to one arm?
- What if the **largest** sites were allocated to one arm?
- What if a **whole site** drops out of the study after randomization?



The Design of MB

- Allocation ratio (C:MHP:HVP) = 1:3:3
- Initial plan: 42 sites total (6:18:18)
- We would like to ensure **imbalance control** on **key baseline factors** at the site level as well
- **Issues:**
  - We could not implement in all 42 sites at once
  - We had an adaptive randomization method (it got complicated quickly)
  - Site dropout

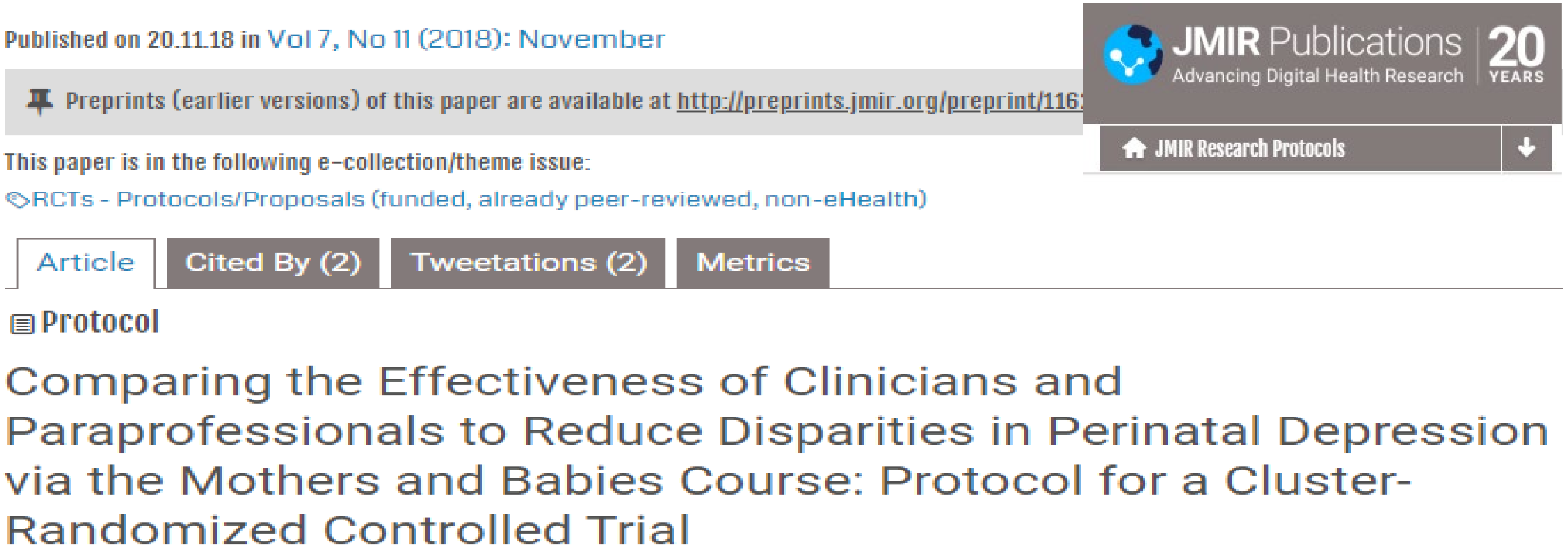
The Design of MB

- ‘Waves’ of randomization

Wave 1 (N=14)	2:6:6 Allocation	(1 C + 1 MHP drop out in meantime)
Wave 2 (N=19)	4:7:8 Allocation	(2 HVP + 1 MHP drop out)
Wave 3/3.5 (N=12)	1:6:5 Allocation	(1 MHP + 1 HVP drop out)

- Notes:
  - Account for dropouts + current assignments in each ‘wave’
  - In Wave #3, we reached a point in which we enrolled one-at-time → employed adaptive methods for the last few sites
  - Randomized = 45, dropout = 8 → 37 active sites (6 C:16 MHP:15 HVP)

The Design of MB



Published on 20.11.18 in Vol 7, No 11 (2018): November

Preprints (earlier versions) of this paper are available at http://preprints.jmir.org/preprint/116

This paper is in the following e-collection/theme issue:

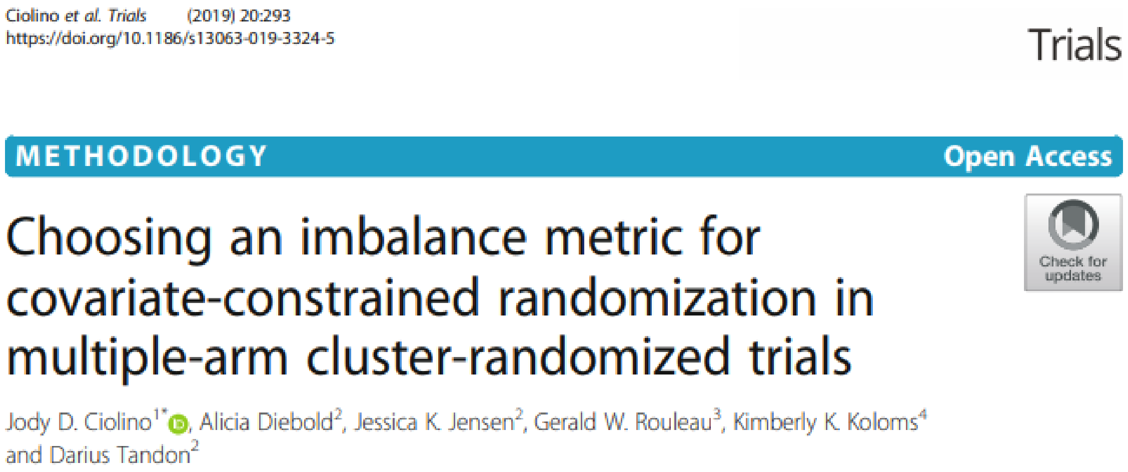
RCTs - Protocols/Proposals (funded, already peer-reviewed, non-eHealth)

Protocol

Comparing the Effectiveness of Clinicians and Paraprofessionals to Reduce Disparities in Perinatal Depression via the Mothers and Babies Course: Protocol for a Cluster-Randomized Controlled Trial

Ciolino et al. Trials. 2019;20:293.

Jensen JK, **Ciolino JD**, Diebold A, Segovia M, Degillio A, Solano-Martinez J, Tandon SD. JMIR research protocols. 2018;7(11):e11624.



Ciolino *et al. Trials* (2019) 20:293
https://doi.org/10.1186/s13063-019-3324-5

Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials

Jody D. Ciolino[1*], Alicia Diebold[2], Jessica K. Jensen[2], Gerald W. Rouleau[3], Kimberly K. Koloms[4] and Darius Tandon[2]

Key components include

Objectives and hypothesis

Measurements and outcomes

Sample size

Data management

Analysis plan

Power – what is it?

- **Probability** that we 'find something' significant in our data when we should
- **Probability** that our data shows us what is really going on in the population
- **Example: MB study**
  - If MB delivered by HVP *is* more efficacious than usual care, then power = probability that we conclude (based on our statistical test) that HVP arm has lower depressive symptoms scores at six months postpartum when compared to the Control arm
  - If we assume MB delivered by HVP *is not inferior* to MB delivered by MHP, then power = probability that we conclude non-inferiority (based on our statistical test)

Things that affect power

- Assumptions, **Assumptions**, ***Assumptions***
- Variability
  - In outcome
  - For cluster-randomized studies, the interplay in variability between sites vs. within sites (ICC)
- Effect size
  - Superiority study: minimal clinically important difference
  - Non-inferiority study: margin of non-inferiority
- Type I error (false positive rate): can be one or two-sided
- **Sample size**

MB Study Study Sample Size Considerations

- Outcome: Quick Inventory of Depressive Symptoms Self Report Score (QIDS-SR16)
- Information needed from investigators:
  - Variability: standard deviation in outcome + ICC estimate
  - Clinically meaningful difference (across arms)
  - Number of sites / Number of participants per site

MB Study Study Sample Size Considerations

Superiority Aim: HVP vs. Control

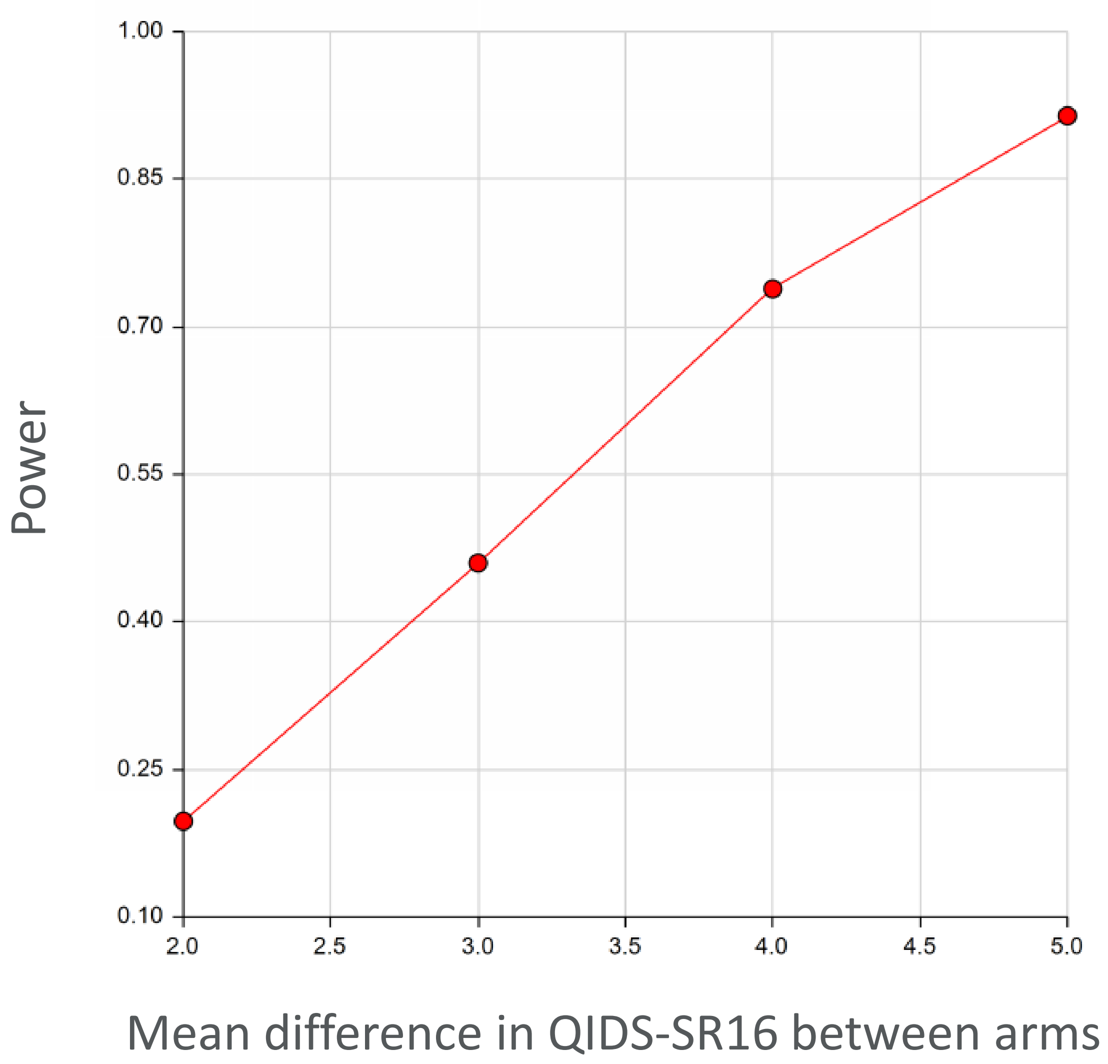
- Assume:
  - QIDS-SR16 standard deviation = 6 points
  - ICC estimate = 0.02
  - **Clinically meaningful difference** (across arms) = 5 points
  - Number of control sites = 5
  - Overall type I error = 0.05

- →we need **n=16** participants per site to allow **for 90% power to detect this difference**
- What if we want to be powered to detect a smaller difference?

Why does effect size matter?

With n=16 participants per site with at least 5 sites per arm, we have 90% power to detect a mean 5-point difference QIDS-SR16 across arms

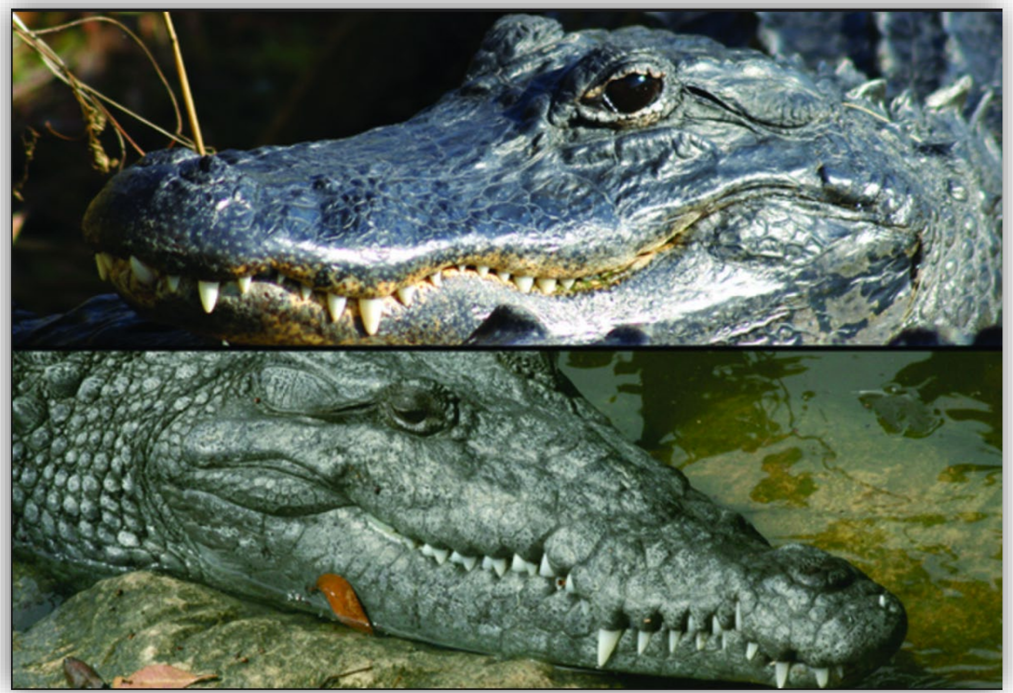
If we hold sample size constant, how likely are we to detect smaller differences across arms?



Northwestern Medicine®

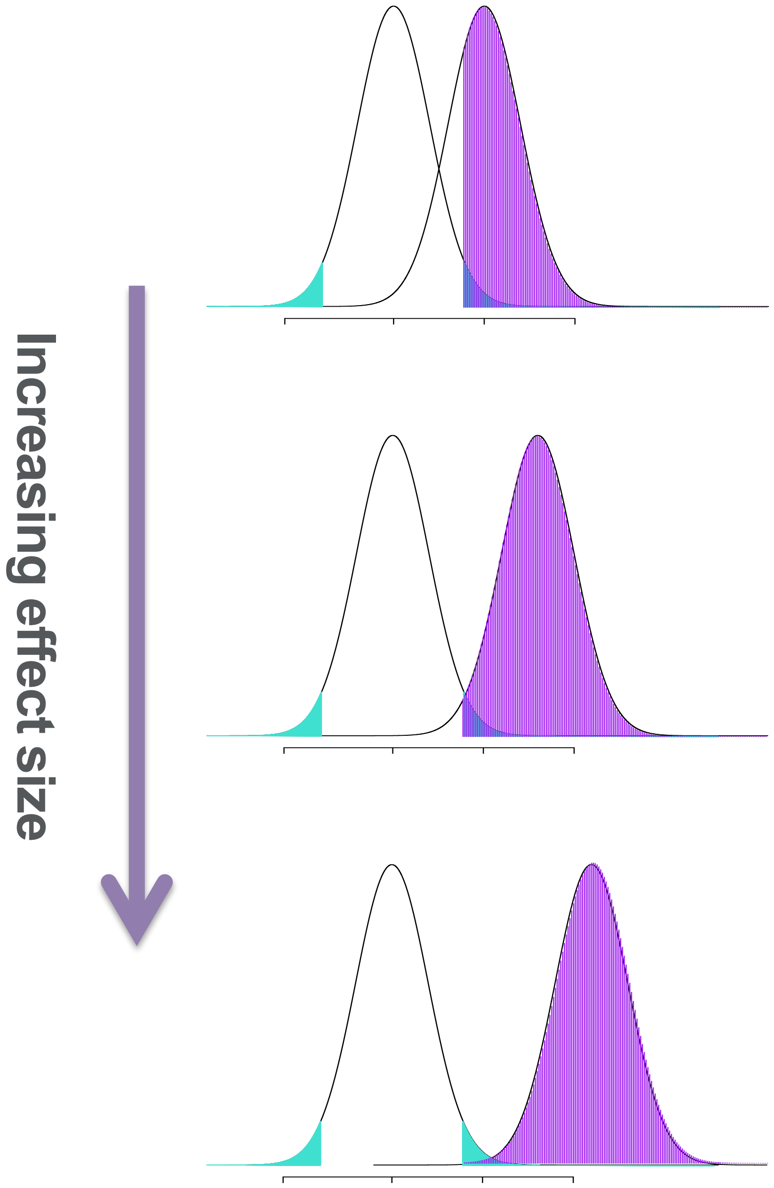
Why does this happen?

- Intuitively, the more similar two things are, the more difficult it is to tell them apart from one another
- It’s easier to tell the difference between two things that are not similar



vs.

Statistically…



Increasing effect size

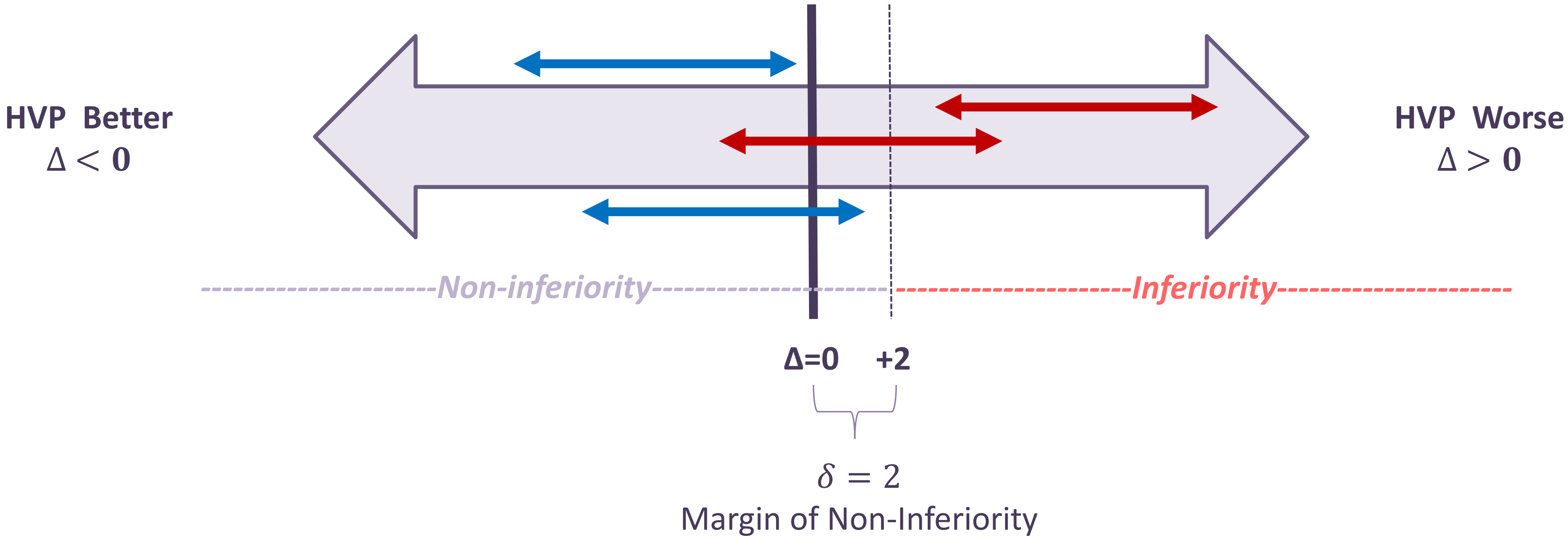
What about Non-inferiority?

- **Non-inferiority** goal: illustrate therapy is “not worse than” standard of care / some other therapy by some (small) amount
  - Generally sacrifice some efficacy to allow of potential benefits of novel therapy (maybe less side effects, less expensive)
  - Thus, we need to determine *a priori* a **margin of non-inferiority**
  - That is, what amount of efficacy are we willing to sacrifice for the benefit of the novel therapy?
- **Margin of non-inferiority** must be substantially **smaller than** a clinically meaningful difference

Non-inferiority Aim of the MB Study

$\Delta = \mu_{HVP} - \mu_{MHP}$

μ = Adjusted 24-week QIDS score


HVP Better
$\Delta < 0$
HVP Worse
$\Delta > 0$
--------------------Non-inferiority--------------------
--------------------Inferiority--------------------
Δ=0
+2
$\delta = 2$
Margin of Non-Inferiority

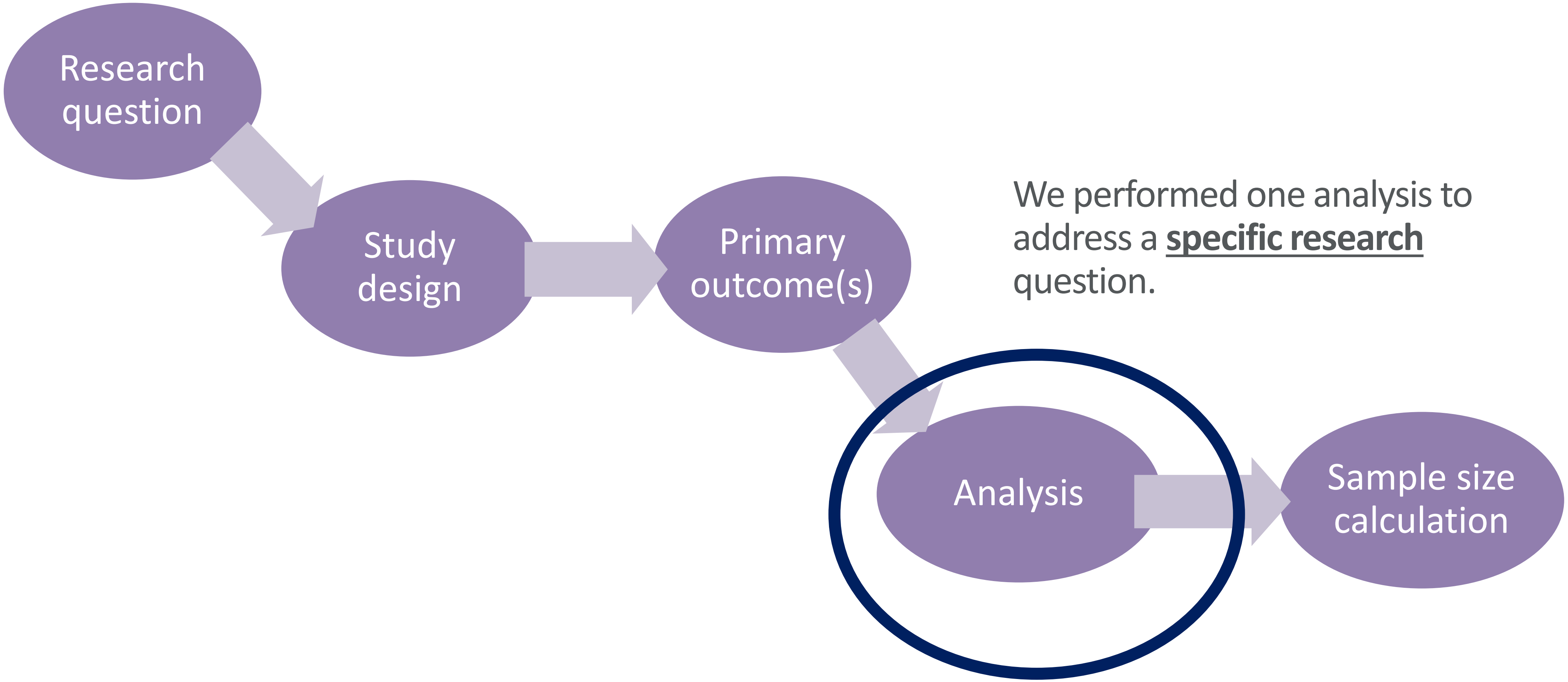
Sample size for Non-inferiority Aim

- Assume:
  - QIDS-SR16 standard deviation = 6 points
  - ICC estimate = 0.02
  - **Margin of NI** = 2 points
  - Number of sites per intervention arm = 15
  - Overall type I error = 0.05
- →we need approximately 22-26 participants per site = 30 sites total x (22-26) participants per site = **660 – 780 total** (*in just the intervention arms*)
- If we want to be able to make the claim of non-inferiority, we need to be powered to do so (requires much larger sample size than superiority)

Northwestern Medicine®

Common Misconception

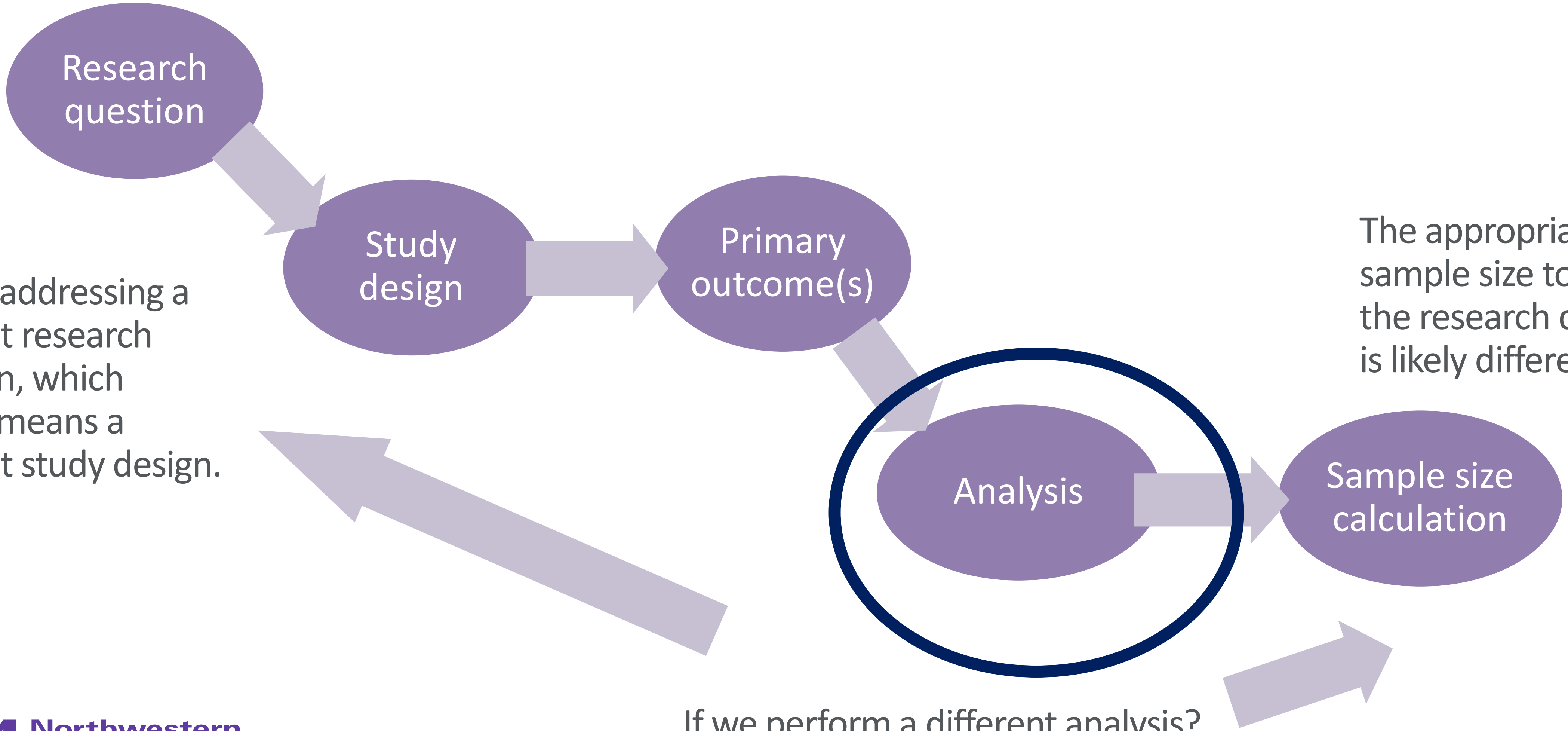
We designed our study as a superiority and the result was insignificant → can we reframe this as a non-inferiority analysis?



We performed one analysis to address a **specific research** question.

Common Misconception

We designed our study as a superiority and the result was insignificant → can we reframe this as a non-inferiority analysis?



We are addressing a different research question, which usually means a different study design.

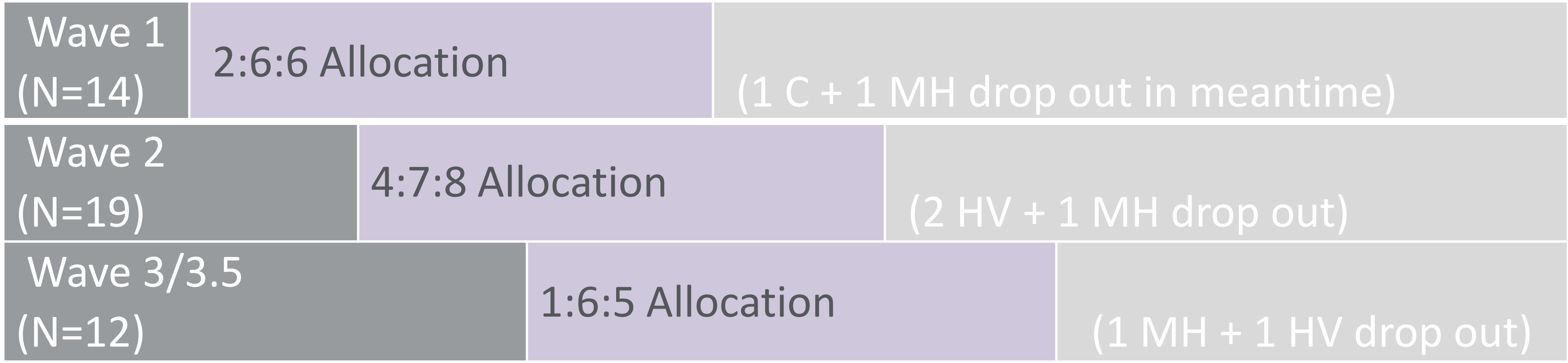
The appropriate sample size to address the research question is likely different.

If we perform a different analysis?

Northwestern Medicine

Recall…The Design of MB

- Allocation ratio (C:MHP:HVP) = 1:3:3
- Initial plan: 42 sites total (6:18:18)



Wave 1 (N=14) | 2:6:6 Allocation | (1 C + 1 MH drop out in meantime)

Wave 2 (N=19) | 4:7:8 Allocation | (2 HV + 1 MH drop out)

Wave 3/3.5 (N=12) | 1:6:5 Allocation | (1 MH + 1 HV drop out)

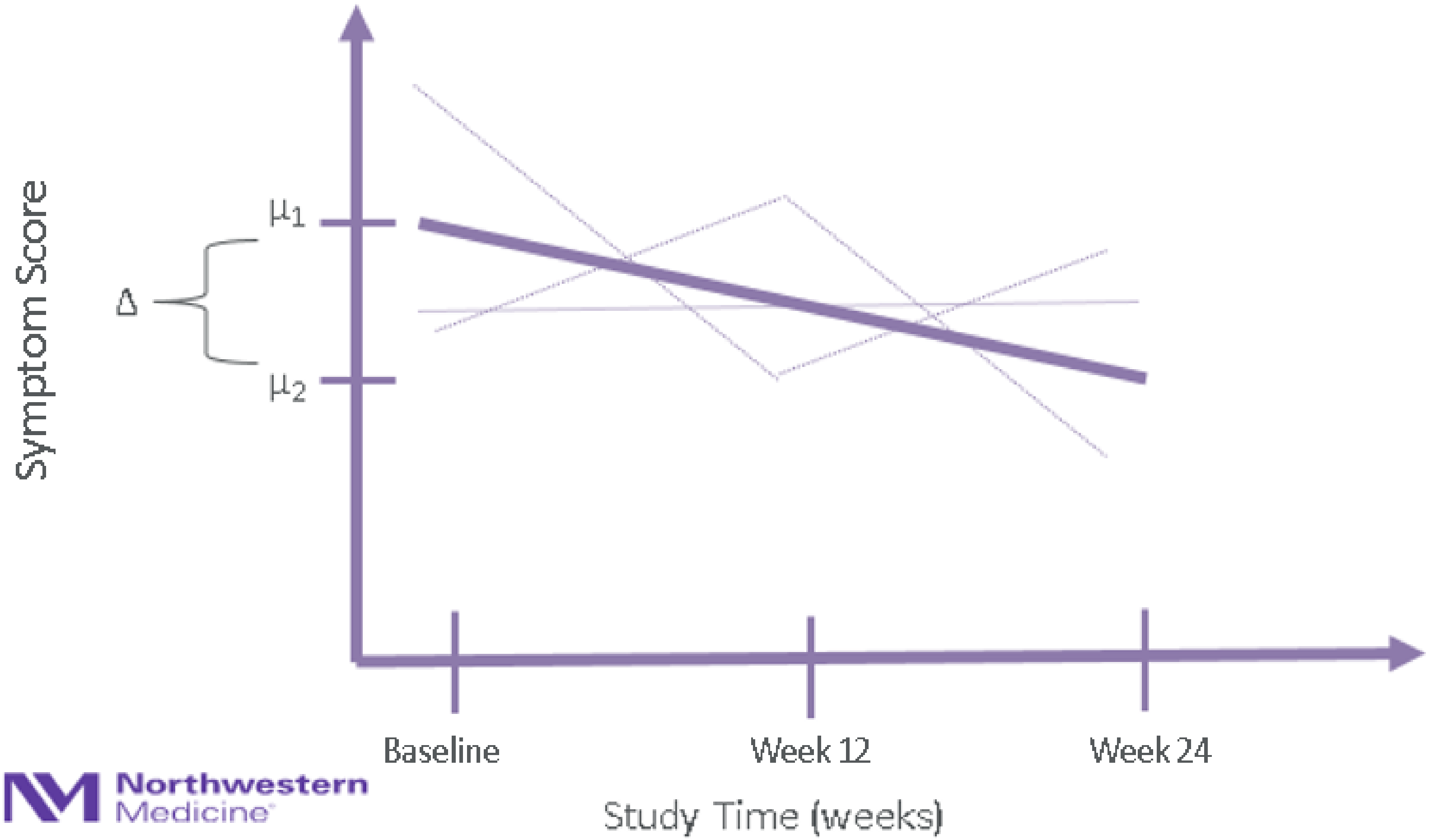
- Randomized = 45, dropout = 8 → 37 active sites (6 C:16 MHP:15 HVP)

Northwestern Medicine®

Research question and sample size

- Recall: Research question → study design → primary outcome(s) → analysis → sample size calculation
- If we calculate a sample size that is simply not feasible, one strategy would be to go back to one of the upstream elements and rethink it
- For example…

What is the question/outcome?



1. Average change in depressive symptom scores (change from baseline) → Δ
2. Meeting "success" definition of crossing below a score threshold → Binary
3. Average score after 24 weeks follow-up → μ_2
4. Score trajectory over 24 weeks →dotted lines

Sample size take-home points

- Sample size calculations are an iterative process in the design of a study
- Sample size and power calculations are based on *assumptions*
- Why are underpowered studies so prevalent?
  - Poor planning / consideration ahead of time: outcomes, analyses, dropout rates, recruitment rates, exaggerated 'meaningful differences' (based on previous, small studies)
  - Bad luck
- Research question → study design → primary outcome(s) → analysis → Sample size calculation

Key components include

Objectives and hypothesis

Measurements and outcomes

Sample size

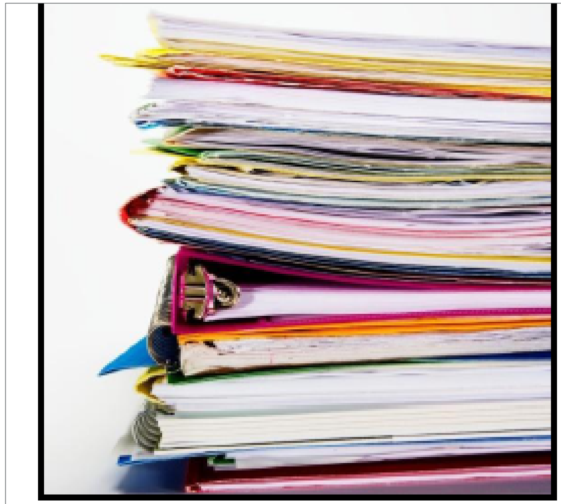
Data integrity and management

Analysis plan

Why should we care about data management?

- Formal statistical training tends to focus on
- Study design → **study conduct** → analysis methods

- But **study conduct** (including capturing and managing data) can also have large impact on our ability to answer the study question
  - Bias
  - Variability
  - Poor data quality
  - Missing data
  - Etc.



Who should see the data?

Where do I find the data?

How do I know the data have not changed?

The Data Management Plan

- Data management plan (DMP) may be housed in the study protocol or as a separate document
- It explains the process of collecting, storing, reviewing, sharing data,
- Also outlines: responsibilities, timing, security, etc.

Topics to Cover in the DMP:

1. CRF creation – who, how, when, etc.
2. Database design and build
3. Edit check specifications
4. Testing and release
5. Data workflow (paper trails if applicable)
6. Reports/metrics
7. Query management
8. Managing special/non-CRF data
9. Coding special terms (medications, Adverse Events)
10. Handling AEs/SAEs
11. Data transfer/database locking procedures

The investigator is ultimately responsible for ensuring the accuracy, completeness, and timeliness of the data reported

The investigator is responsible for ensuring the accuracy, completeness, and timeliness of the data reported

The protocol should provide details regarding the type(s) of data that will be collected and any relevant data standards or ***common data elements***

Specify whether data will be paper or electronic, distributed or central, batched or ongoing processing, and any related requirements; what data will be collected on CRFs and what data will be collected from other sources

Further details should be provided in the ***MOP or the data management plan****, including detailed descriptions of source documentation, CRFs, instructions for completing forms, data handling procedures, and procedures for data monitoring*

Information should include the role in data collection, review of data, trial materials, and reports, as well as retention of source documents, files, and records

It is not acceptable for the CRF to be the only record of a participant's inclusion in the study. Study participation should be captured in a participant's medical record

Provide a list of planned data standards, formats, terminologies and their versions, used for the collection, tabulation, analysis of study data

NIH-FDA Clinical Trial Protocol Template – v1.0 7 Apr 2017

Statistical Considerations

This section of the protocol should have the following subsections describing the statistical tests and analysis plans:

- ***Statistical Hypotheses:*** *State the formal and testable null and alternative hypotheses for Primary and Secondary Efficacy Endpoint (s); specify the type of comparison (e.g. superiority, equivalence,..) and time period for which each endpoint will be analyzed*
- ***Sample Size Determination:*** *Include number of participants to recruit, screen, ansd enroll to have adequate power to test the key hypotheses for the study. Provide all information needed to validate your calculations.*
  - *Discuss whether the sample size provides sufficient power for addressing secondary endpoints or exploratory analyses (e.g., subgroup analyses or moderator analyses involving an interaction term.*
  - *Method for adjusting calculations for planned interim analyses, if any*
- ***Populations for Analyses:*** Clearly identify and describe the analysis datasets (e.g., which participants will be included in each). As a guide, this may include, but is not limited to, any or all of the following:
  - Intention-to-Treat (ITT) Analysis Dataset (i.e., all randomized participants)
  - Modified Intention-to-Treat Analysis Dataset
  - Safety Analysis Dataset
  - Per-Protocol Analysis Dataset
  - Other Datasets that may be used for sensitivity analyses

NIH-FDA Clinical Trial Protocol Template – v1.0 7 Apr 2017

Statistical Considerations

- Statistical Analyses
  - General approach
  - Analyses of the primary and secondary efficacy endpoint(s)
    - Define the measurement
    - Describe the scale and the statistical procedure(s)
    - Describe how results of statistical procedure(s) will be presented (e.g., adjusted means (Leastsquares means (LSMEANS)) with standard errors, odds ratios with 95% confidence intervals, prevalence rates, number-needed-to-treat)
    - Describe details to check assumptions required for certain types of analyses (e.g., proportional hazards, transformations)
    - Describe how missing data will be handled
    - Describe the statistical adjustment used for controlling for Type I error if more than one endpoint
- Safety Analyses
- Baseline Descriptive Statistics
- Planned Interim Analyses
- Sub-Group Analyses
- Tabulation of Individual Participants Data
- Exploratory Analyses

- **NIH-FDA Clinical Trial Protocol Template – v1.0 7 Apr 2017**

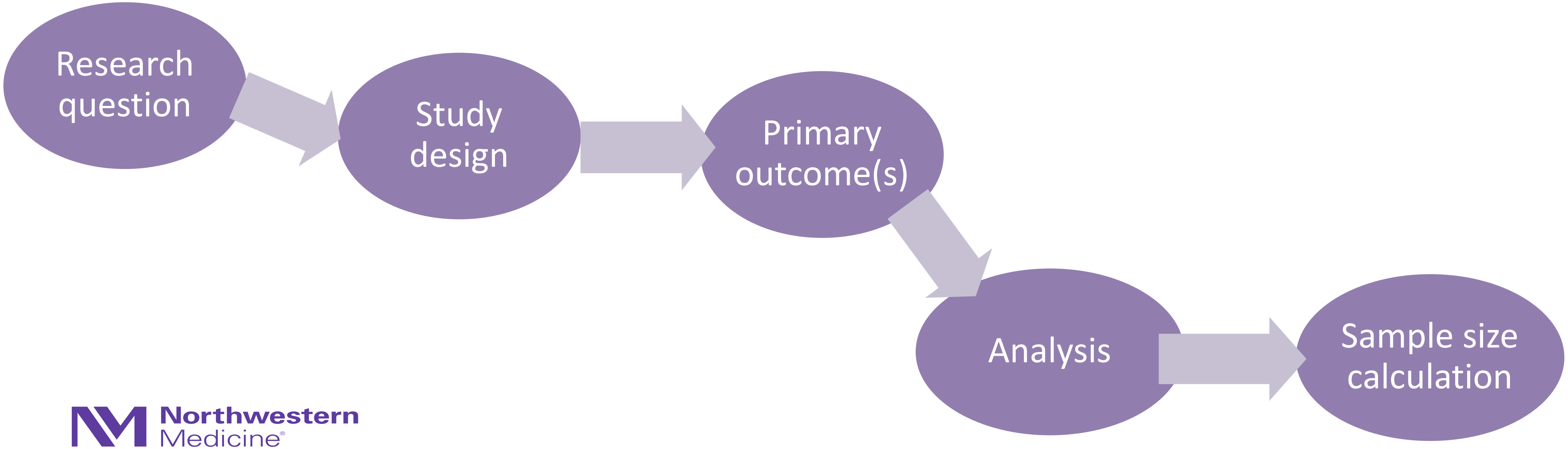
What is a Statistical Analysis Plan (SAP)?

- Note: the SAP may be housed within the protocol, depending on the study type and complexity
- The SAP is a technical document that describes in detail the planned statistical analysis of a clinical study as outlined in the protocol
- Although the SAP is often a standalone document, it should be reviewed in conjunction with the study protocol

Take-home points

jody.ciolino@northwestern.edu

- Good design ➔ good science
- Thoughtful protocol development is a critical piece in any translational research – it is a PROCESS
- Statistical concepts that should always be considered throughout: bias and variability



Research question

Study design

Primary outcome(s)

Analysis

Sample size calculation

Northwestern Medicine®

Your feedback is important to us!
(And helps us plan future lectures)

jody.ciolino@northwestern.edu

Statistically Speaking: Upcoming Lectures

We hope to see you again!

Wednesday, January 15	**To p or not to p: reflections on recent p-value statements** **Mary Kwasny, ScD**, Professor, Division of Biostatistics, Department of Preventive Medicine
Wednesday, March 18	**Biostat Basics: Some Practical Things to Know** **Nina Srdanovic, MS**, Statistical Analyst, Division of Biostatistics, Department of Preventive Medicine
Monday, May 11	**Logistic Regression: Odds & Ends** **Lauren Balmert, PhD**, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine

All lectures will be held from Noon to 1 pm in Baldwin Auditorium,
Robert H. Lurie Medical Research Center, 303 E. Superior St.

http://www.feinberg.northwestern.edu/sites/bcc/education/lecture/2019.html