

Statistically Speaking Lecture Series

Sponsored by the Biostatistics Collaboration Center

Capturing Racial and Ethnic Data and Considerations for Analysis

Jody D. Ciolino, PhD

Associate Professor

Director, Master of Science in Biostatistics

Department of Preventive Medicine

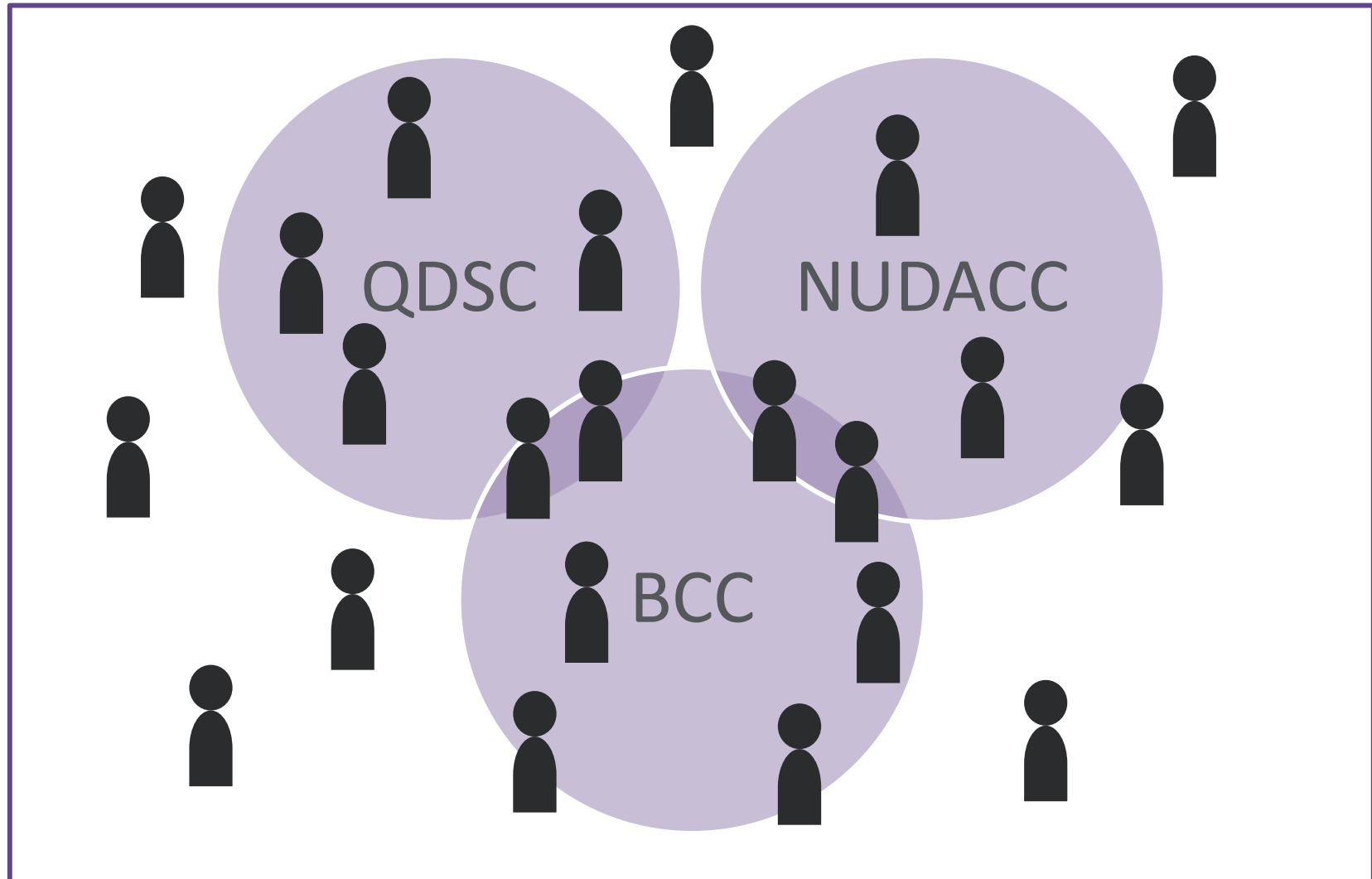
February 3, 2022

Before we begin...

Biostatistics at NU

Overview

Division of Biostatistics (Chief: Denise Scholtens),
Department of Preventive Medicine (Chair: Donald Lloyd-Jones)



Biostatistics Centers and Cores

Overview



Biostatistics Collaboration Center (BCC)

- Supports **non-cancer** research at NU
- Initial 1-2 hour consultation subsidized by FSM Research Office
- Grant, Hourly
- <https://www.feinberg.northwestern.edu/sites/bcc/>

Quantitative Data Sciences Core (QDSC)

- Supports **cancer-related** research at NU
- Free to Lurie Cancer Center (LCC) members
- Grant
- <https://www.cancer.northwestern.edu/research/shared-resources/quantitative-data-sciences.html>

Northwestern University Data Analysis and Coordinating Center (NUDACC)

- Prospective, large **multicenter research**
- Comprehensive support (e.g., clinical monitoring, data analysis, project management)
- Grant
- <https://www.feinberg.northwestern.edu/sites/nudacc/>

Disclaimer:

The views presented today are my own.
They do not represent those of the
statistical community nor those of
Northwestern University at large.

I have no relevant conflicts of interest.

Goals for today

- Provide guidelines and recommendations for you/researchers to consider when conducting studies involving racial or ethnic data
- Data capture → Data analysis → Reporting

Background

- This work initiated from the **Department of Preventive Medicine's (DPM) Working Group** focusing on **Response to Structural Racism in Research**.
- Group formed in the Fall of 2020 – several activities and recommendations since its' inception.
- **Members:** Kiarri Kershaw (Chair); Mercedes Carnethon; Jody Ciolino; Frank Granata; Elizabeth Gray; Mark Huffman; Molly Jones; Monica Rodriguez; Leah Neubauer; Denise Scholtens
- Goals of the group...



Goal 1:

To create a set of **protocols for embedding a racial and ethnic equity perspective in research**, from inception to dissemination.

Goal 2:

To **identify and strengthen existing and new resources** – both internal and external - to promote scholarship in the areas of health equity, racial and ethnic equity, and structural discrimination.

Goal 3:

To contribute to the **education and training** of future health equity scholars.

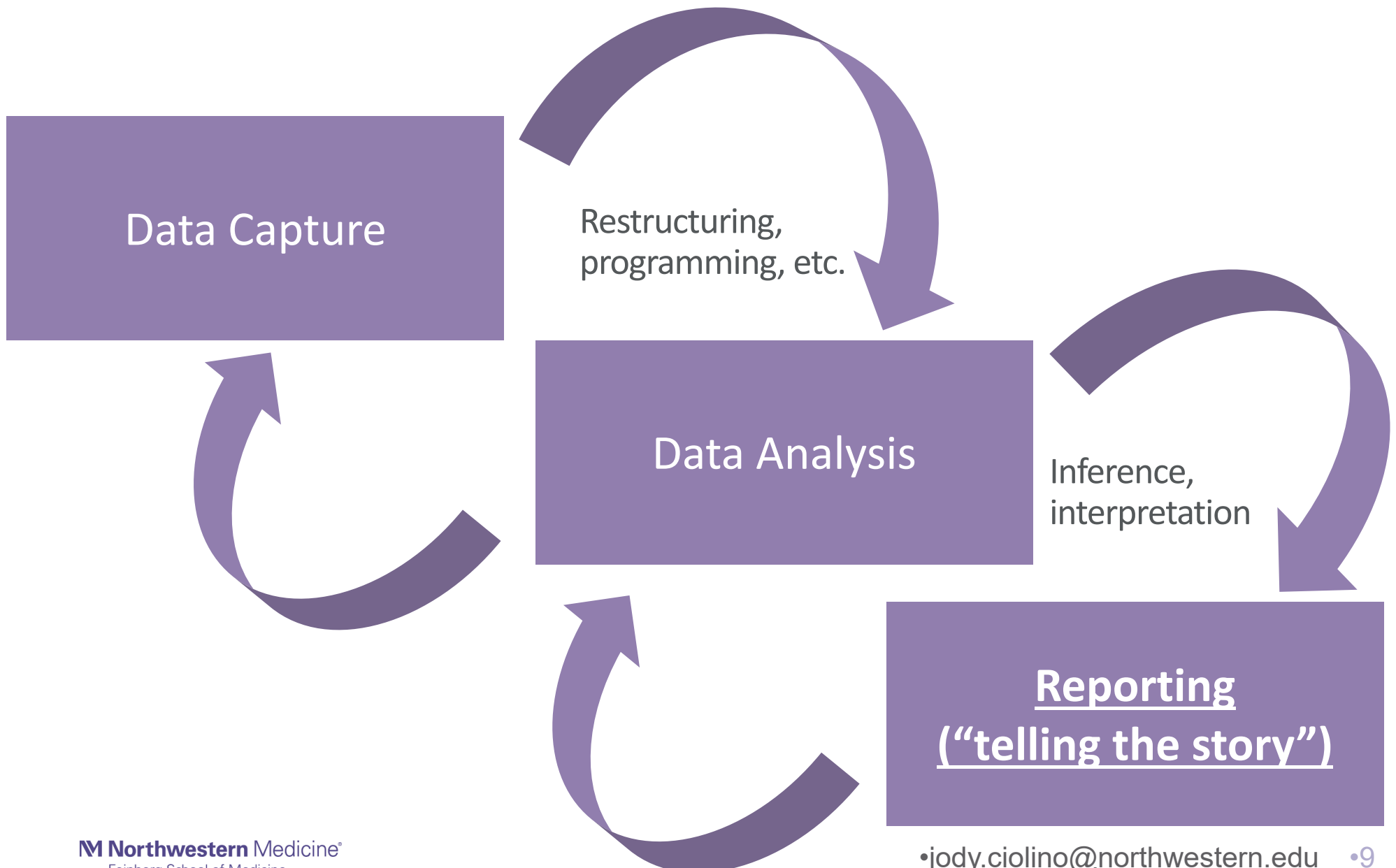
Goal 1:

To create a set of protocols for embedding a racial and ethnic equity perspective in research, from inception to dissemination.

While there is a lot of overlap between goals and the ideas we discuss today certainly pertain to the others, this is our focus today.

Starting with analysis (and reporting)

Embedding...from inception to dissemination

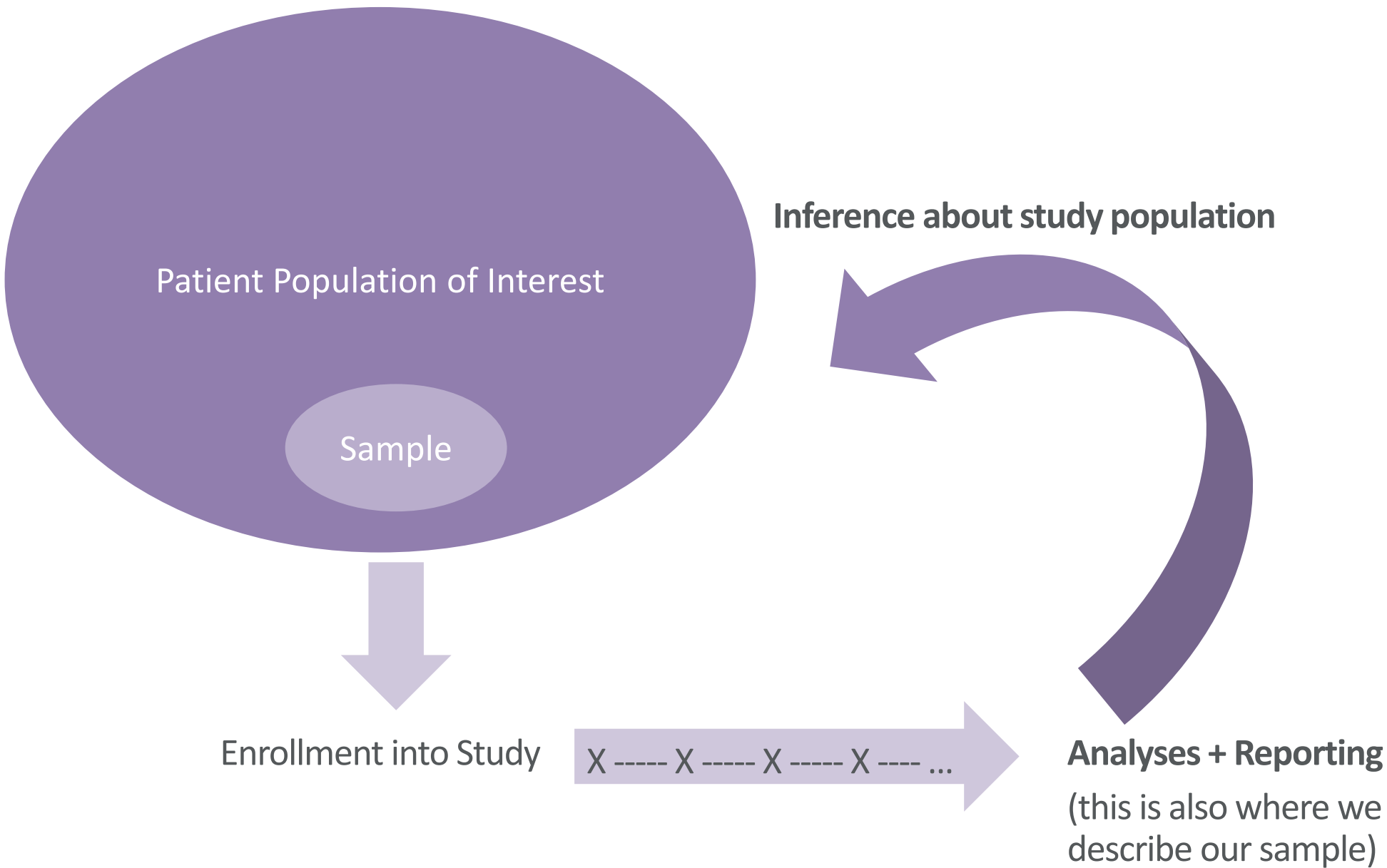


On Analysis and Reporting Regarding Racial and Ethnic Data

Points to consider in advance of the study initiation

- **Audience** – who is going to hear the story?
- Audience may include...
 - People of the community
 - Other researchers
 - Clinicians
 - Funders
 - Etc.
- What is the ultimate **message** you will want to convey with respect to racial/ethnic data?
 - **Describing** the sample
 - “**Adjusting**” or “controlling” for (potential) confounding effects
 - **Heterogeneity** in effects

Describing the Sample – What do we mean and why do we do it?



Describing the Sample

For a research, reviewer, clinical, lay audience

- “Table 1” – typically includes a description of demographic characteristics of your study participants
 - Age
 - Sex
 - Gender identity
 - Race
 - Ethnicity
 - Etc.
- For example...

Table 1. Baseline Characteristics of Study Participants by Group and Overall^a

Characteristic	Physical Therapy Group (n = 43)	Usual Care Group (n = 58)	Total (N = 101)
Age, median (IQR) ^b	45.0 (35.0–57.0)	38.0 (31.0–53.0)	40.5 (31.5–54.0)
Sex, no. (%) of women ^c	27 (62.8)	32 (56.1)	59 (59.0)
Race ^c			
White	20 (46.5)	17 (29.8)	37 (37.0)
Black	13 (30.2)	23 (40.4)	36 (36.0)
Hispanic	06 (14.0)	13 (22.8)	19 (19.0)
Other	04 (9.3)	04 (7.0)	08 (8.0)
Highest education level ^c			
None	01 (2.3)	01 (1.8)	02 (2.0)
High school/GED	11 (25.6)	17 (29.8)	28 (28.0)
College	15 (34.9)	26 (45.6)	41 (41.0)
Graduate or professional school	16 (37.2)	13 (22.8)	29 (29.0)
Pain level, median (IQR)	8.0 (7.0–9.0)	7.0 (6.0–8.0)	7.0 (6.0–8.0)

Kim HS, Ciolino JD, Lancki N, Strickland KJ, Pinto D, Stankiewicz C, Courtney DM, Lambert BL, McCarthy DM. A Prospective Observational Study of Emergency Department–Initiated Physical Therapy for Acute Low Back Pain. *Physical therapy*. 2021 Mar;101(3):pzaa219.

Describing the Sample

For a research, reviewer, clinical, lay audience

- “Table 1” – typically includes a description of demographic characteristics of your study participants
 - Age
 - Sex
 - Gender identity
 - Race
 - Ethnicity
 - Etc.
- **The reason for “Table 1” in Human Subjects’ Research**
 - Describing the “sample” from the population
 - Allowing the audience to understand the generalizability (or lack thereof) of findings
 - Identifying reach and gaps of research

Describing the Sample

Funder or regulatory

When reporting to NIH (as one example)...

Cumulative (Actual)

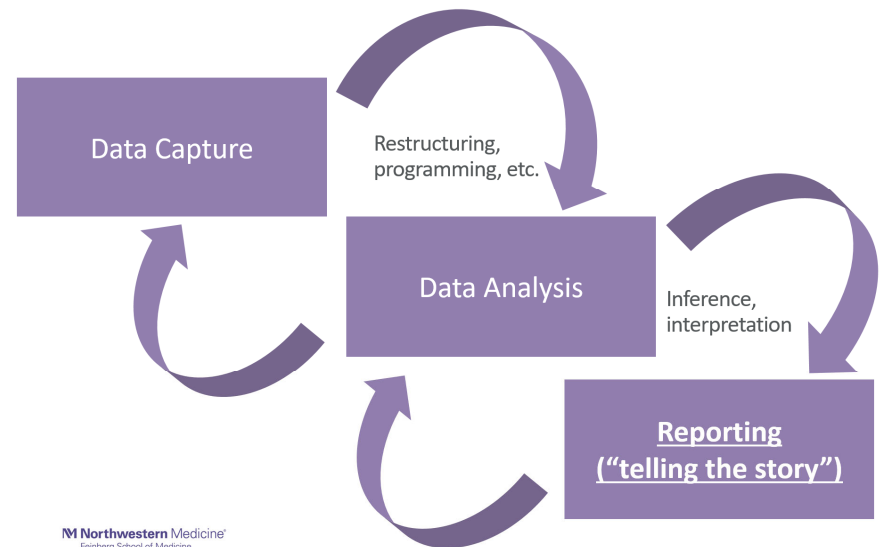
Racial Categories	Ethnic Categories									Total
	Not Hispanic or Latino			Hispanic or Latino			Unknown/Not Reported Ethnicity			
	Female	Male	Unknown/Not Reported	Female	Male	Unknown/Not Reported	Female	Male	Unknown/Not Reported	
American Indian/Alaska Native	0	0	0	0	0	0	0	0	0	0
Asian	0	0	0	0	0	0	0	0	0	0
Native Hawaiian or Other Pacific Islander	0	0	0	0	0	0	0	0	0	0
Black or African American	0	0	0	0	0	0	0	0	0	0
White	0	0	0	0	0	0	0	0	0	0
More than One Race	0	0	0	0	0	0	0	0	0	0
Unknown or Not Reported	0	0	0	0	0	0	0	0	0	0
Total	0	0	0	0	0	0	0	0	0	0

We know ahead of time that we need to report on these categories at minimum
 → (seems obvious) that we must collect these data as well, **at minimum**

Now let's go back to capturing the data



Assume for a moment the sole purpose = to describe our sample

- Knowing the purpose allows us a starting point for data capture
- If NIH-funded study, we can *start* here
- **Race:**
 - American Indian / Alaska Native
 - Asian
 - Native Hawaiian / Pacific Islander,
 - Black or African American
 - White
 - More than one Race
- **Ethnicity:**
 - Hispanic or Latino
 - Not Hispanic or Latino



If we start there, this may be what a data collection tool looks like (for a participant)

Example Case Report Form (CRF)

Visit Date * must provide value	<input type="text"/>  Today M-D-Y
Date of Birth * must provide value	<input type="text"/>  Today M-D-Y
Age * must provide value	<input type="text"/> View equation
Sex as defined at birth * must provide value	<input type="radio"/> Male <input type="radio"/> Female reset
Race: * must provide value	<input type="radio"/> American Indian / Alaska Native <input type="radio"/> Asian <input type="radio"/> Native Hawaiian / Pacific Islander <input type="radio"/> Black or African American <input type="radio"/> White <input type="radio"/> More than one Race reset
Ethnicity: * must provide value	<input type="radio"/> Hispanic or Latino <input type="radio"/> Not Hispanic or Latino reset

While this may serve the needs of the study for purposes of reporting and describing the sample...

From the participant perspective...

This setup may be confusing and rather annoying

Possible participant perspective

- “Why do you need this information?”
- “I would prefer not to give you this information.”

Recommendation

(from Working Group Document)

- Explain the purpose – in the consent or in instructions for survey.
- *“Collecting information on race and ethnicity helps us understand our study participants’ background. It will help people reading the study results understand whether the study may or may not apply to them or their patients.”*
- Allow participants
 - to select “**choose not to answer**” for these questions (preferred) **or**
 - to **skip the question(s)** altogether.

From the participant perspective...

This setup may be confusing and rather annoying

Possible participant perspective

- “I identify with many of these options; it seems silly to make me choose just one.”
- “I don’t know how to answer this. I am not sure what this means.”
- “None of these options really apply to me.”

Recommendation

(from Working Group Document)

- Allow participants to “select all that apply.”
- Allow for an “**unsure**” option.
- Allow an “**other**” option or “some other race or ethnicity”.
 - Consider: If “other”, specify.
 - Note – this may provide inconsistencies with data entry and open up opportunity for error, but may be worth the effort here.

From the participant perspective...

This setup may be confusing and rather annoying

Possible participant perspective

- “What is the difference between race and ethnicity? They seem to be asking the same thing from my perspective.”
- “These categories feel very general. How can you group people of Indian, Chinese, Japanese, Korean, Thai, Indonesian [and many other] descent into just one broad category?”

Recommendation

(from Working Group Document)

- Consider whether is it necessary to break the two constructs apart – you could consider capturing together in one field/set of fields.
- Ask a question that does not specifically state “ethnicity” in the label... “*Do you consider yourself Hispanic or Latino/a/e?*”
- Consider **capturing data on a more granular level** – you can always collapse into larger, “required” categories after the fact. This is **preferred over risking participant difficulty** with answering these questions.

Other options for data collection...

Example CRFs / Data Collection Tools

With which **race** do you **most** identify?

* must provide value

- Black or African American
- White or Caucasian
- Hawaiian or Pacific Islander
- East Asian (e.g., China, Japan, N. or S. Korea)
- Native American or Alaska Native
- South Asian (e.g., India, Pakistan, Sri Lanka, Nepal)
- Multiple races
- Some other race that is not listed above
- Prefer not to disclose

Do you consider yourself **Hispanic** or **Latino/a/x**?

* must provide value

- Yes
- No
- Prefer not to disclose

Other options for data collection...

Example CRFs / Data Collection Tools

Race & Ethnicity

Raza y etnia

With which race/ethnic group do you primarily identify (choose one)?

¿Con qué raza o grupo étnico se identifica principalmente (elija uno)?

* must provide value

- White / Caucasian
De raza blanca/caucásica
- Black / Sub Saharan African descent
De raza negra/descendencia africana subsahariana
- Hispanic - Mexican, Mexican American, Chicano/a
Hispana: mexicana, mexicanoamericana, chicano/a
- Hispanic - Puerto Rican
Hispana: portorriqueña
- Hispanic - Cuban
Hispana: cubana
- Hispanic - Other
Hispana: otra
- Native American / Alaskan
Nativa americana/de Alaska
- East Asian - Chinese, Japanese, Korean
Asiática (del Este): china, japonesa, coreana
- South Asian - India, Pakistan, Sri Lanka, Bangladesh
Asiática (del Sur): de India, Pakistán, Sri Lanka, Bangladesh
- Central Asian - Iran, Afghanistan
Asiática (central): Irán, Afganistán
- Middle Eastern
Oriente Medio
- Multi-racial
Multirracial
- Other
Otro

reset

If 'Other', specify:

Si la respuesta es 'Otro', especifique:

* must provide value

Other options for data collection...

Example CRFs / Data Collection Tools

What is your race?		
	No	Yes
American Indian or Alaska Native * must provide value	<input type="radio"/>	<input type="radio"/>
Asian * must provide value	<input type="radio"/>	<input type="radio"/>
Black or African American * must provide value	<input type="radio"/>	<input type="radio"/>
Native Hawaiian or Other Pacific Islander * must provide value	<input type="radio"/>	<input type="radio"/>
White or Caucasian * must provide value	<input type="radio"/>	<input type="radio"/>
Other * must provide value	<input type="radio"/>	<input type="radio"/>
Please specify other race * must provide value	<input type="text"/>	
Are you Hispanic or Latino? * must provide value	<input type="radio"/> Yes <input type="radio"/> No	

Other options for data collection...

Example CRFs / Data Collection Tools

Racial / Ethnic Data

Do you identify with any of the racial / ethnic categories below?

Please check here if you prefer not to disclose racial / ethnic data to the Society: I choose not to disclose

Select all that apply

- Black or of Sub Saharan African Descent
- Central Asian Descent (e.g., Iran, Afghanistan)
- East Asian Descent (e.g., China, Japan, Korea)
- Middle Eastern
- Native American or Alaska Native
- Native Hawaiian / Pacific Islander
- South Asian Descent (e.g., India, Pakistan, Sri Lanka, Bangladesh)
- White
- Hispanic or Latinx
- Some other race or ethnicity (not listed above)

Specify 'Other' race or ethnicity:

Summary on Data Collection

From the DPM Working Group Documents

- 1. Consider the audience and target study population.** Depending upon country, culture, or age group of interest, these data may not apply or the study participants may not understand the social construct or categories of race or ethnicity.
- 2. Explain the purpose** of this data collection to study participants (either within the consent or in the instructions of the relevant survey).
- 3. Consider the goal** of collecting these data (reporting only vs. analyses?).
- 4.** If the study is short-term (e.g., less than one year in length), these data will likely only require collection at “baseline.” If the study is longer-term, investigators should **recognize that racial/ethnic identity from a participant’s perspective may change over time.** Therefore, the investigators could consider allowing participants to provide updates on these variables as their time in the study progresses.

Summary on Data Collection

From the DPM Working Group Documents

5. We **always** suggest obtaining these data points **via participant self-report**, according to the categories in which they identify personally.
 - a. Participants **must be allowed the option to select more than one** racial/ethnic category.
 - b. Participants **must be allowed** to either **skip** questions related to race and ethnicity **OR select a “Prefer not to answer” option**.
 - c. Consider “other” and “unsure” options.
6. Although **certain countries** and **reporting entities recognize a difference between race and ethnicity**, study participants may not. **Consider whether it is necessary to break these two constructs apart into separate segments of data capture of a demographics form.**
7. Regarding the number of racial and ethnic categories, **more options are better** to a degree; however, **participant burden and overwhelm should be considered.**

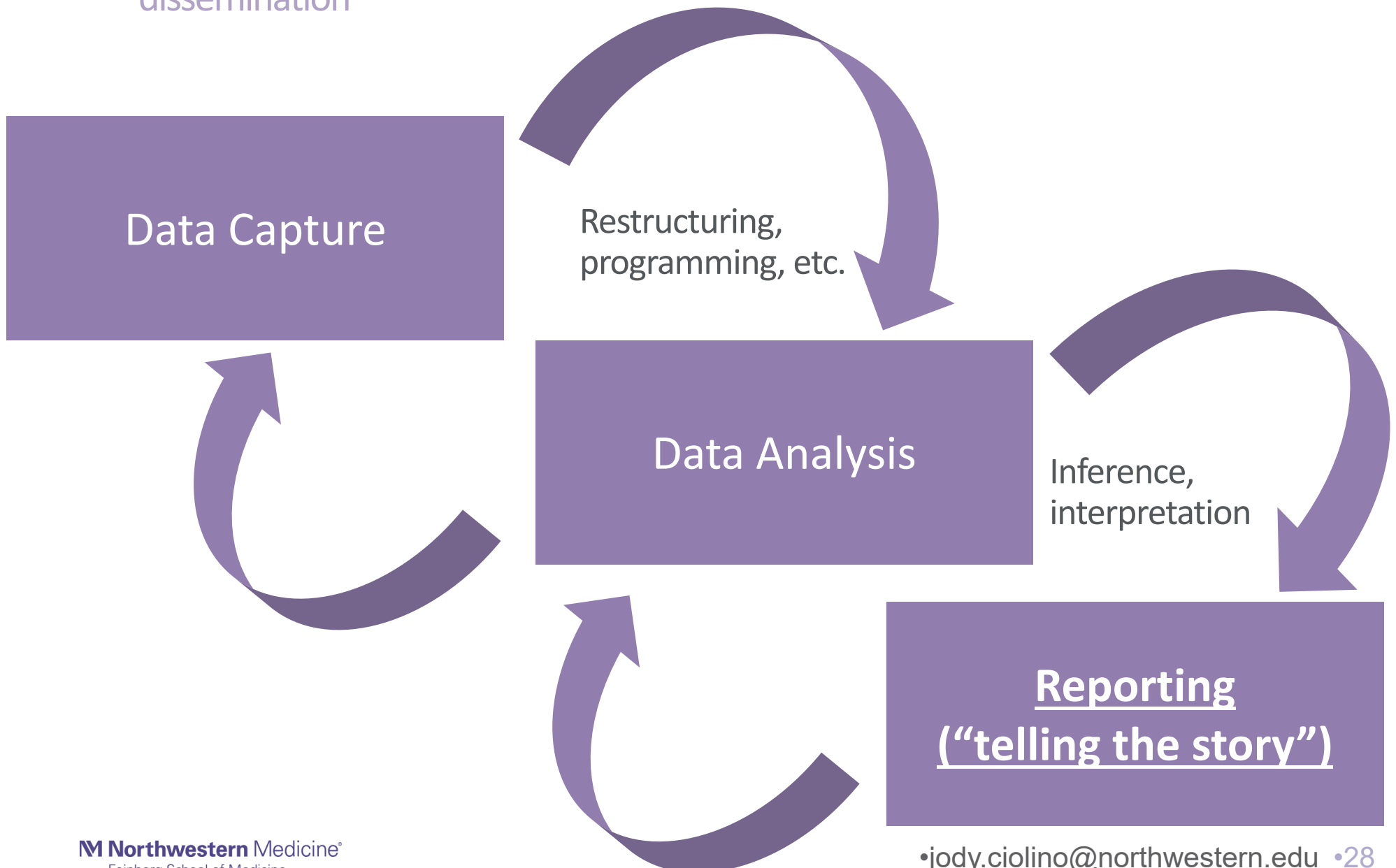
Summary on Data Collection

From the DPM Working Group Documents

8. **Avoid forcing participants to identify with broad categories** to the extent possible. It will be easier to collapse categories for reporting and analytic reasons after data collection, rather than risk participant difficulty in response.
9. **Some recommendations** of phrasing toward study participants:
 - a. “Which of these best describe your race / ethnicity? (Select all that apply)”
 - b. “Do you identify with any of the following racial/ethnic categories?”
 - c. “Indicate the racial / ethnic categories with which you (most) identify below. (Select all that apply)”
 - d. In each case, followed by a series of checkboxes or a matrix of yes / no fields as outlined above.

Circling back to analysis (and reporting)

Recall from Working Group goals: Embedding...from inception to dissemination



On Analysis and Reporting Regarding Racial and Ethnic Data

Points to consider in advance of the study initiation

- **Audience** – who is going to hear the story?
- Audience may include...
 - People of the community
 - Other researchers
 - Clinicians
 - Funders
 - Etc., etc.
- What is the ultimate **message** you will want to convey with respect to racial/ethnic data?
 - **Describing** the sample
 - **“Adjusting”** or “controlling” for (potential) confounding effects
 - **Heterogeneity** in effects

Using Race / Ethnicity Variables in Analysis

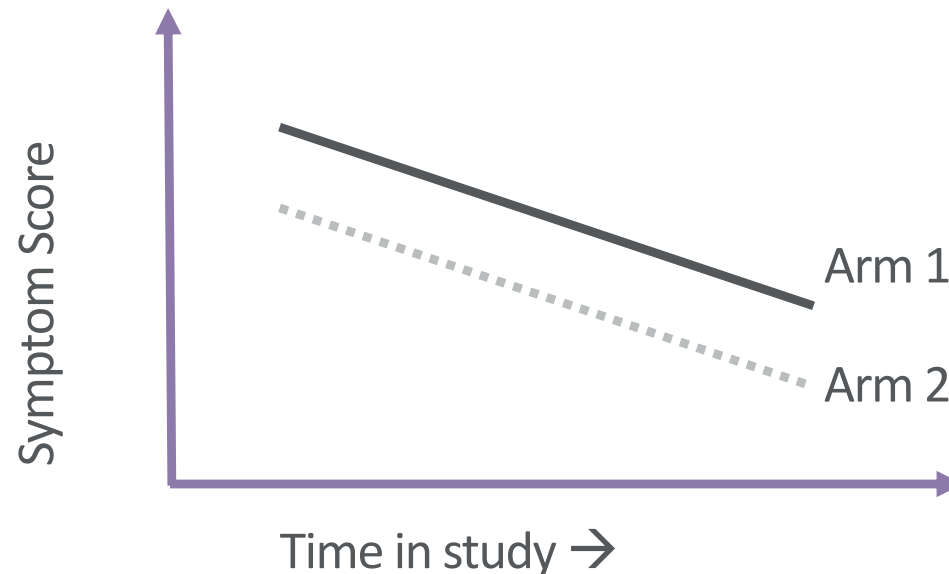
What is the meaning?

- Consider an example...
- Study participants are enrolled into a randomized interventional study (a clinical trial).
 - **Arm 1** = cognitive behavioral therapy
 - **Arm 2** = treatment with antidepressant
- Participants are **followed for 2 years**, and **have study visits every 6 months**.
- Primary outcome of interest is **depressive symptom severity score**.
- **Primary aim**: to evaluate intervention effects on depressive symptom scores over time.

Race / Ethnicity Variables in Analysis

Hypothetical Example

- Analysis plan:
 - H0: mean depression score for patients treated with CBT = mean depression score for patients treated with antidepressants.
 - H1: mean scores in the two patient groups are not equal.
 - Longitudinal modelling techniques to compare depressive symptom scores over time:
 - $Y = \text{intercept} + \text{time} + \text{study arm} + \text{error} + \text{error}$



What typically comes next?

After primary analyses

- **Sometimes controlling for multiple potential covariates** of interest
 - Example – “controlling for race”, “controlling for ethnicity”
 - How often do we hear: “Did you control for race?” from reviewers?
- **Recommendation** – if this is not of interest, not part of the study goals, really consider whether it makes sense to perform additional analyses controlling for race.

What typically comes next?

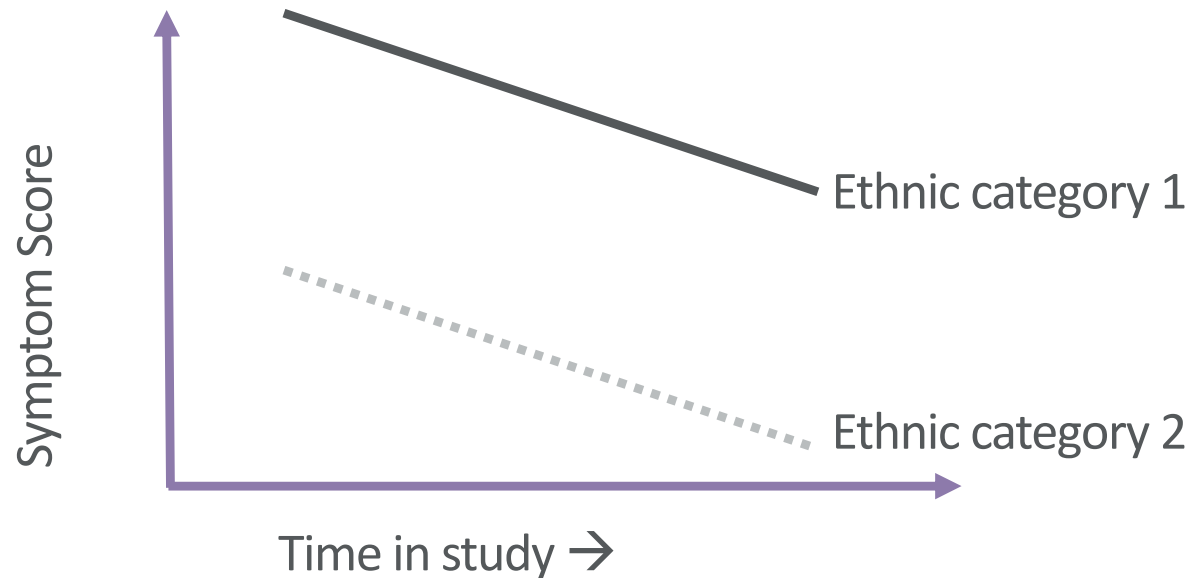
After primary analyses

- **If we have reason to believe that race or ethnicity does have an effect** on outcome (in this case depression scores), and it is **scientifically justified**, consider adjusting for race or ethnicity.
- Ideally, this would be pre-specified adjustment.
- **Example:** perhaps there is reason to believe that those self-identifying as Hispanic have differing depressive symptoms scores, in general, than those not self-identifying as Hispanic.
- To control for this (assumed) difference, your new model becomes something like:

$$Y = \text{intercept} + \text{time} + \textit{ethnicity} + \textbf{study arm} + \text{error} + \textit{error}$$

What does this mean?

$Y = \text{intercept} + \text{time} + \textit{ethnicity} + \text{study arm} + \text{error} + \text{error}$



Assumption: there is an inherent difference in symptom scores between these two groups, regardless of treatment. \rightarrow *Is this a valid / justified assumption?*

- It is impossible, in a randomized study to obtain perfectly “balanced” distributions of all variables across study arms.
- Adding in variables for which there is an assumption like this helps increase precision on the estimate of interest (i.e., that study arm effect).

What about heterogeneity in treatment effects?

Usually more exploratory in nature, unless these are primary goals of the research study

- **What does this mean in context of our example?**
 - Do we have evidence of a differing study arm effect within groups of participants self-identifying as specific ethnic categories?



This may be one hypothesized view of this heterogeneity – there are countless.

Statistically...

$$Y = \text{intercept} + \text{time} + (\text{ethnicity}) * \text{study arm} + \text{error} + \text{error}$$

In analyses, things get complicated very quickly

Some general recommendations

- **Pay attention to your study population**
 - Are adjustments or evaluations for heterogeneity even possible?
 - Did you collect the data in such a way to allow for this?
 - Even if these analyses are a part of the study goals, they may not be possible as group / cell counts may be too low.

In analyses, things get complicated very quickly

Some general recommendations

- **Always go back to the original goal** of collecting racial / ethnic data in relation to overarching study goals.
 - If simply to describe – use journal, funder, etc. as starting point.
 - Things get more complicated when the goals include evaluating confounding or heterogeneity of effects within subgroups...
 - There is a “push” and “pull” between precise and granular representation of identity and model degrees of freedom.

An example on considerations for analyses

Suppose...

- We have a study similar to the one on depressive symptoms previously described.
- We would like to control for race in analyses.
- We **pre-specify this plan to adjust for self-identified race in analysis.**
- However, here is our distribution of our participants' self-identified racial categories:

Self-Identified Race	N	%
Asian	4	4%
Black	1	1%
Multiple Races	5	6%
White	78	89%
Total	88	100%

Example, continued

Statistically, the largest subgroup provides the most “information” in the model. Larger subgroups tend to result in increased precision and more stable model estimates.

Self-Identified Race	N	%
Asian	4	4%
Black	1	1%
Multiple Races	5	6%
White	78	89%
Total	88	100%

If we leave these categories as is and put “race” with four levels into our statistical model...

- We will most likely end up with nonsensical parameter estimates and confidence intervals (model instability issues) due to **really low cell counts**.
- We may end up with a lack of **anonymity** issue in reporting here as well.

So what should we do?

Can we collapse these categories?

Self-Identified Race	N	%
Asian	4	4%
Black	1	1%
Multiple Races	5	6%
White	78	89%
Total	88	100%



Self-Identified Race	N	%
Asian / Black / Multiple Races	11	11%
White	78	89%
Total	88	100%

- Does it still make sense to adjust in analyses?
- The team decided “no” and determined it more appropriate to report/summarize only.
- Reason: it did not make sense/there was no scientific justification for assuming that depressive symptoms would be inherently different across such a broad subgroup.

In analyses, things get complicated very quickly

Some general recommendations

- **Pre-specify as much as possible according to study goals in a Statistical Analysis Plan (SAP)**
 - Allow for flexibility and contingency plans in the event of low cell counts or violations of basic assumptions.
- Consider inference - **Will this practice of adjustment or evaluation of interaction terms involving race truly allow for meaningful and actionable conclusions?**
- **Perform both unadjusted and adjusted analyses.**
 - Pay attention to differences in inferences across models.
 - Pre-specify which analyses are “primary”, “secondary”, “exploratory”, “sensitivity,” etc.
- **Take care to ensure preservation of anonymity to the extent possible.**

In analyses, things get complicated very quickly

Some general, more “statsy” recommendations

- **Pay attention to parameterization** and categorization of groups.
- Treat race / ethnicity as a **factor** (i.e., not as an ordinal variable).
- **Consider collapsing** categories, but **pay attention to inference** on coefficients for collapsed categories.
- If participants had the option to select >1 category, consider using a series of indicator variables (1/0) for each category rather than grouping them into mutually exclusive categories.

Tying it all together

- We have presented some **considerations to help serve as guidelines and recommendations for Human Subjects' Research.**
- Please keep in mind that **research studies will all have unique needs**, and there is **no single set of best practices** that would apply to all studies.
- Today we focused on data capture and analysis and reporting for Race and Ethnicity.
- The work product(s) from the DPM Working Group also include(s) brief discussion capturing data on sex and gender identity, and considerations for analyses. Many of the ideas discussed today apply to these concepts as well.

Acknowledgements

- Thank you all for your attention on these important concepts.
- **A special thanks to the DPM Working Group** on Response to Structural Racism:
 - Kiarri Kershaw (Chair)
 - Mercedes Carnethon
 - Jody Ciolino
 - Frank Granata
 - Elizabeth Gray
 - Mark Huffman
 - Molly Jones
 - Monica Rodriguez
 - Leah Neubauer
 - Denise Scholtens

Upcoming *Statistically Speaking* Lectures

Thursday, March 3
1-2pm

Considerations when Leveraging Electronic Health Records for Causal Inference: A “Create-Your-Own-Data” Adventure
Lucia C Petito, PhD, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine

All lectures will be held via Zoom:

<https://northwestern.zoom.us/j/99300239887?pwd=eHJZS1BOM3ZiWW9PN0lpWDYxQXdYQT09>

Thank you!

Biostatistics Collaboration Center

Search web or people

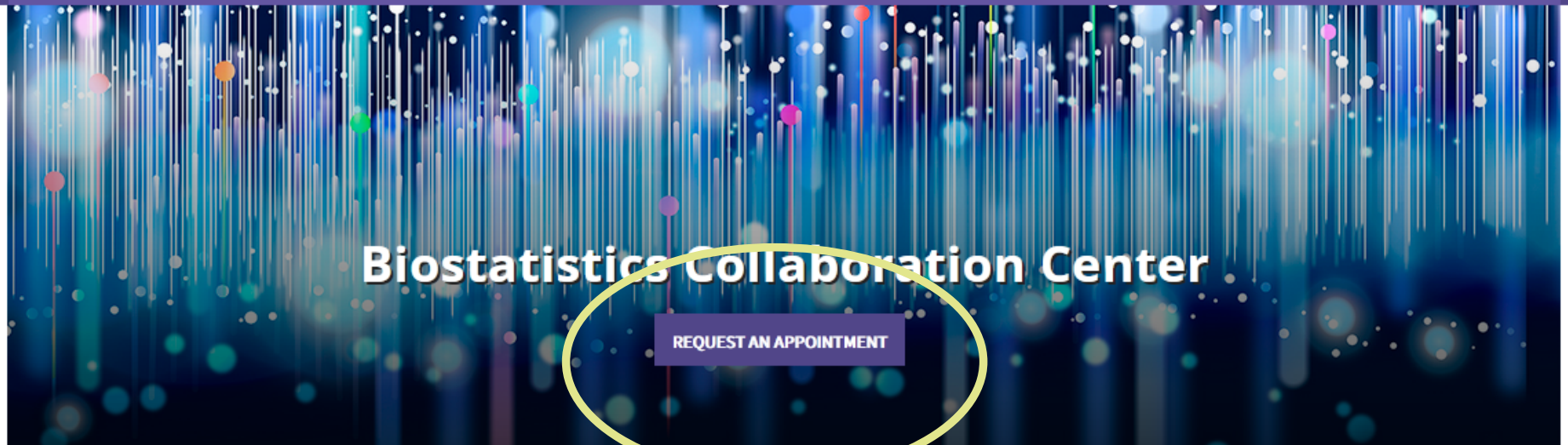


About Us ▾

Research Services and Support

Education

People



Expertise in biostatistics, statistical programming and data management

Since 2004, the Biostatistics Collaboration Center (BCC) has partnered with Northwestern investigators at every level – from residents and postdoctoral fellows to junior faculty and well-established senior investigators.



Meet Our Team

Our faculty and staff provide research and health services in a variety of settings across Chicago and Evanston. We work with clinical partners. Meet our team and their expertise.

OUR PEOPLE

Please note, we have been experiencing a high volume of requests and responses may be delayed. Your patience is appreciated as we work to field requests in as timely a manner as possible!