Statistically Speaking Lecture Series

Sponsored by the Biostatistics Collaboration Center

Unseen Worlds: How Missing Data Impact Statistical Analyses

Lucia Petito, PhD

Assistant Professor of Biostatistics

Tuesday, May 16, 2023, 12-1pm

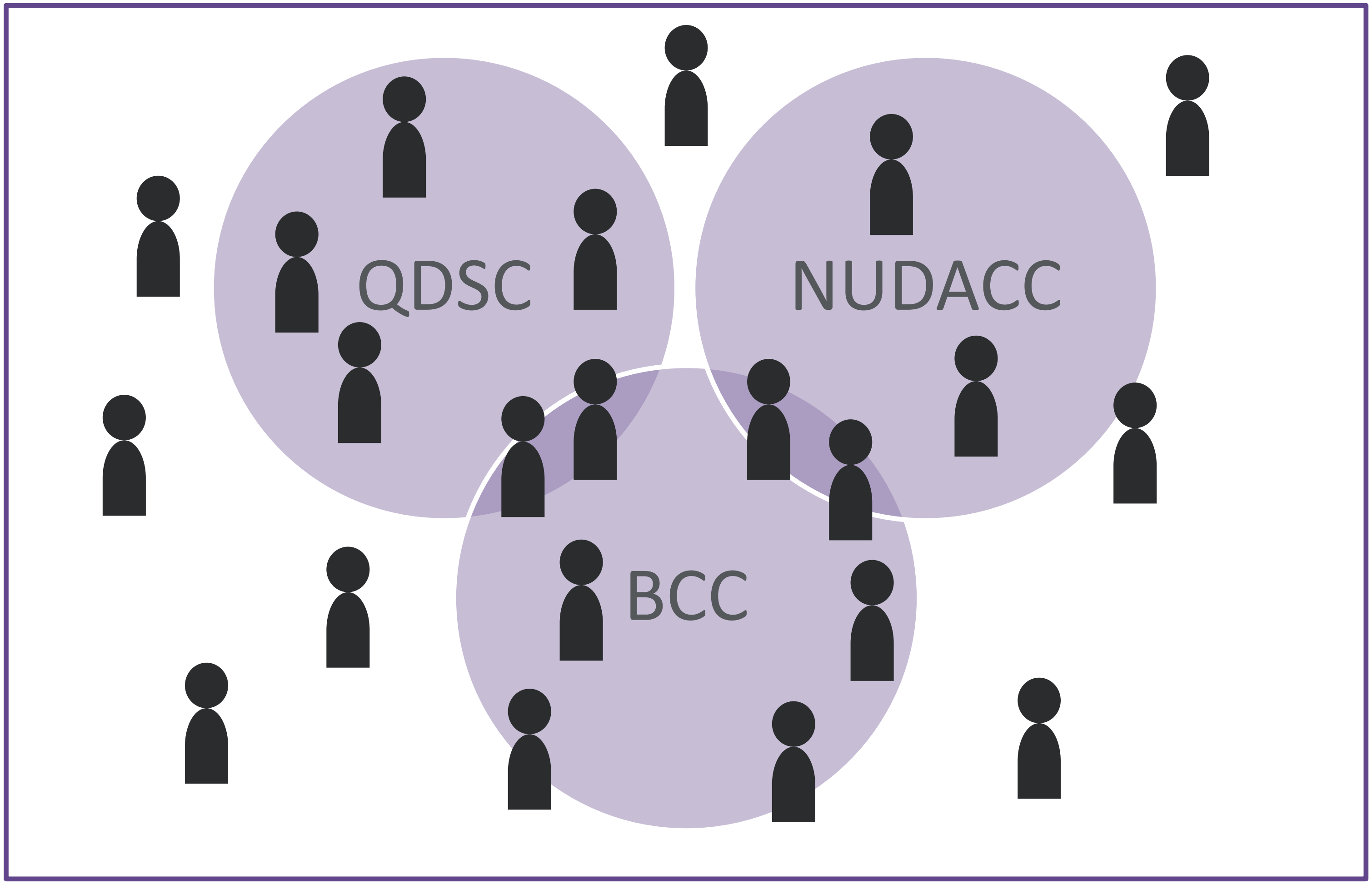
Hybrid: McGaw (Daniel Hale Williams Auditorium) and Zoom

Northwestern Medicine®
Feinberg School of Medicine

Biostatistics at NU

Overview

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



Biostatistics Centers and Cores

Overview

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute

Robert H. Lurie Comprehensive Cancer Center of Northwestern University

Stanley Manne Children's Research Institute

Ann & Robert H. Lurie Children's Hospital of Chicago

Shirley Ryan AbilityLab

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/

Outline

Today we will cover:

- **What is Missing Data?**
- **Pitfalls of Complete Case Analyses**
- **Single Imputation Methods**
- **Multiple imputation**

What is Missing Data?

What is missing data?

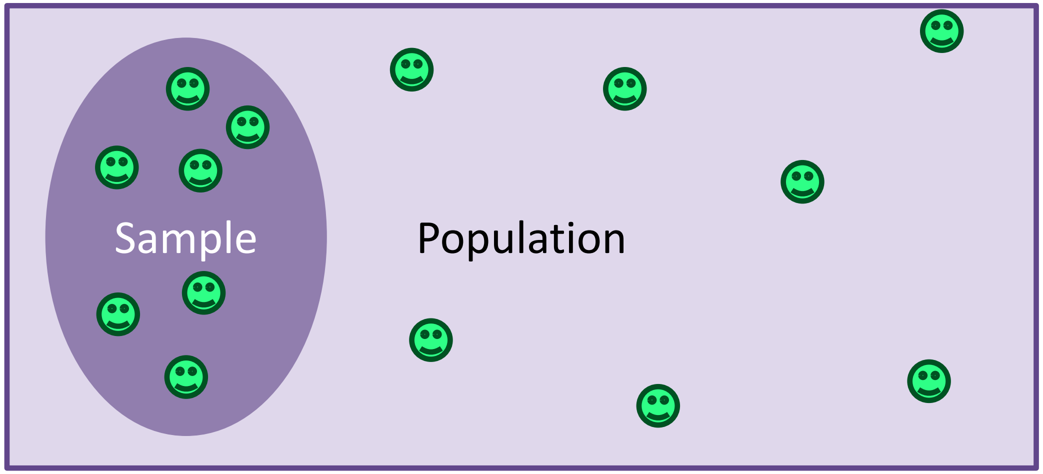
Broadly speaking, data that you wish you had, but you do not.

- Whether data are missing or not can often depend on the research question
- In some cases, the same data can be missing for some questions, but not missing for others
- In other cases, it is clear which data are missing

Examples of missing data

Scenario: Sampling to make inferences on a population

- The population units that were not included in your sample can be considered missing data.
- If you could have it, you would rather have data on the entire population rather than a sample
- Solution: Weight the survey data by the inverse probability of selection



Examples of missing data

Scenario: Estimating the prevalence of a disease

- Question on an NHANES survey:

"Has a doctor ever told you that you have diabetes?"

- Possible responses:
  - Yes
  - No
  - Don’t know

- Here the "Don’t knows" are strictly missing data - either a doctor has informed you, or s/he hasn’t

- What about EHR data collected at, say, an urgent care center?

Image source: https://www.nm.org/about-us/careers

Examples of missing data

Scenario: Loss to follow-up in longitudinal studies

- In any study where participants are asked to come in repeatedly, there will be some participants who miss assessments
- Of particular concern are drop-outs: participants who miss an assessment and never return
  - May be systematically different from those participants who remain in a study

- Example: Randomized placebo-controlled trial for efficacy of a pharmacotherapy to treat depression.
  - Control participants who become more depressed may drop-out
  - Implication: Analyses based on the observed (not-missing) data might make the treatment look less effective than had the very depressed controls not dropped out

Examples of missing data

Scenario: Compliance status in randomized trials

- Trials are often interested in the efficacy of a treatment or whether it worked for those who took it.
- Complier Average Causal Effect: Effect of the treatment among those who complied with the intervention
- Calculate the difference in outcomes between compliers randomized to treatment and compliers randomized to control
- For those randomized to the treatment condition, compliance status is observed (not missing).
  - Usually know who took their pills, came to treatment sessions, etc.
- But what about controls? The did not have a treatment to comply with.
  - Here compliance status among the controls is missing data

Image source: https://curehht.org/resource/prophylactic-antibiotics/prescription-medication-in-orange-pill-bottles/

Examples of missing data

Scenario: Measurement error in dietary intake

- Participants in lifestyle interventions are often asked to self-report their dietary intake
- Self-reported diet is subject to measurement error due to memory limitations, poor quantification of portion sizes, etc.
- Dietary intake measured without error can be viewed as **missing data**

- Validation studies: Studies where both self-report (biased) and urinary measures of dietary intake (unbiased) are obtained on the same set of participants
  - Use information from validation studies to "fill-in" missing true intake in a lifestyle intervention

Examples of missing data

Scenario: Unmeasured confounders in observational data

- With observational data, association between a treatment and an outcome does not necessarily imply causation
- There may be a third factor (confounder) related to both the treatment and outcome that might explain their association
- We can control for observed confounders but there always exists the potential for unobserved (i.e. missing) confounders
- Example: Positive association between breast feeding (the treatment) and child IQ (the outcome) might be explained by a third factor related to parenting practice. Mothers who breastfeed may also do other things that are good for their child.

So, why do we care about missing data?

Same reasons we always do, Pinky…

Bias and precision!

Image source: https://medium.com/@christina.j.hunte/how-to-take-over-2018-5-steps-from-pinky-and-the-brain-cfdb90ec0778

So, why do we care about missing data?

- Broadly speaking, not taking into account missing data will:
  - Reduce **precision** due to a loss of information
  - Result in **biased** inferences especially when missing data are systematically different from observed data

- Ad hoc or unprincipled methods can often make a bad situation worse
  - May attenuate or distort relationships between variables
  - Provide overly precise inferences

Pitfalls of Complete Case Analyses

Complete case analysis

- Easiest solution: complete case analysis
- Only includes “complete cases” – cases where all variables needed to fit a particular model are available
- Also known as “listwise deletion”
- Default method in (all) statistical packages
  - R, SAS, Stata, SPSS, etc
- Software wants a rectangular data frame, so it will **delete** data to make one
  - Unintended consequence: The same dataset may use different subsets of the data to calculate summary statistics!

Complete Case Analysis - Example

Let’s assume we have some data from a randomized experiment

- **Y** is the outcome
- **A** is the study arm assignment
- **W**s are covariates for adjustment

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Cases with some missing

Complete Case Analysis - Example

Estimating the mean of Y:

$$\bar{Y} = \sum_i Y_i = 26$$

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Cases with some missing

Complete Case Analysis - Example

Estimating the mean of Y:

$$\bar{Y} = \sum_{i} Y_i = 26$$

Uses all 8 observations!

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Cases with some missing

Complete Case Analysis - Example

Estimating the mean of Y among the treated:

$$E[\,Y \mid A = 1\,] = \sum_{i_{a=1}} Y_i = 26.25$$

Uses all 4 observations!

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Cases with some missing

Complete Case Analysis - Example

Estimating the mean of Y among the untreated:

$$E[\,Y \mid A = 0\,] = \sum_{i_{a=0}} Y_i = 25.75$$

Uses all 4 observations!

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Cases with some missing

Complete Case Analysis - Example

Estimating the conditional mean of Y on A and all 3 Ws:

$$E[\,Y \mid A, W_1, W_2, W_3\,] = ???$$

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Discarded data ☹

Complete Case Analysis - Example

Estimating the conditional mean of Y on A and all 3 Ws:

$$E[\,Y \mid A, W_1, W_2, W_3\,] = ???$$

Using linear regression will **only use the first 4 observations**

ID	W_1	W_2	W_3	A	Y
1	0	0.28	1	0	24
2	1	0.73	1	0	22
3	0	0.67	2	1	29
4	1	0.92	3	1	27
5	NA	0.15	3	1	23
6	1	0.93	NA	1	26
7	0	NA	3	0	29
8	1	NA	NA	0	28

Complete cases

Discarded data ☹

Complete case analysis

- Pro: Easiest option to implement

- Might do OK, but only with *small* amounts of missing data

- What constitutes a *small* amount of missing data? Must consider:
  - Fraction of incomplete cases
  - Observed information among cases with missing values
  - Parameter being estimated

- In the best case scenario, complete case analyses will provide **unbiased** results, but will be **inefficient**, resulting in increased variance of estimates
  - Reduced sample size → Reduced study power
  - Impossible to know if bias will be towards or away from the null

Complete case analysis

Consider the following scenario:

- 100 individuals
- 10 variables
- 10% of each variable is missing completely at random

What is the chance of observing all 10 variables for an individual?

$$\Pr[Complete] = \Pr[X_1 obs] \times \Pr[X_2 obs] \times \ldots \times \Pr[X_{10} obs] = 0.9^{10} \approx 0.35$$

So, on average, only **35 individuals** would contribute complete data on all 10 variables in a sample of 100 individuals!

Missing Data Mechanisms

- Deciding on how to handle missing data is often dependent on understanding the process that gave rise to the missing data
  - This is referred to as the **missing data mechanism**

- These mechanisms are often divided into three categories:
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR) (aka Missing Not at Random - MNAR)

Missing Data Mechanisms

- Deciding on how to handle missing data is often dependent on understanding the process that gave rise to the missing data
  - This is referred to as the **missing data mechanism**

- These mechanisms are often divided into three categories:
  - **Missing Completely at Random (MCAR)**
    - The probability of being missing is completely unrelated to all observed and unobserved patient characteristics
    - Least plausible mechanism, but only one where complete case analysis yields unbiased results
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR) (aka Missing Not at Random - MNAR)

Missing Data Mechanisms

- Deciding on how to handle missing data is often dependent on understanding the process that gave rise to the missing data
  - This is referred to as the **missing data mechanism**

- These mechanisms are often divided into three categories:
  - Missing Completely at Random (MCAR)
  - **Missing at Random (MAR) (aka "ignorable")**
    - Does not assume patients with missing values are similar to those with complete data
    - Instead, assumes that observed values can be used to "explain" which values are missing and help predict what the missing values would be
    - Assumed by most of the currently used valid techniques for handling missing data
  - Not Missing at Random (NMAR) (aka Missing Not at Random - MNAR)

Missing Data Mechanisms

- Deciding on how to handle missing data is often dependent on understanding the process that gave rise to the missing data
  - This is referred to as the **missing data mechanism**

- These mechanisms are often divided into three categories:
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR) (aka Missing Not at Random - MNAR)
    - Most problematic mechanism
    - Occurs when missing values are dependent on unobserved or unknown factors
    - *If present, statistical adjustment for missing data is not possible*

Missing Data Mechanisms

- Deciding on how to handle missing data is often dependent on understanding the process that gave rise to the missing data
  - This is referred to as the **missing data mechanism**

- These mechanisms are often divided into three categories:
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR) (aka Missing Not at Random - MNAR)

- **These assumptions are not testable** ☹

General Principles for Missing Data Analysis

1. Formulate the estimand of interest (what is the scientific question?)
2. Determine whether missing values are meaningful for analysis
3. Document the reasons why data are missing
4. Decide on a primary set of assumptions about the missing data mechanism
5. Conduct statistical analysis
6. Assess the robustness of inferences to various missing data assumptions

Northwestern Medicine®
Feinberg School of Medicine

Single Imputation Methods

Single Imputation

- **Imputation** is the process of replacing missing values with plausible values
- **Pro:** By filling in holes, can create a rectangular data set which is easy to analyze
- **Con:** Done incorrectly it can create bias or understate uncertainty
- Generally speaking, *single imputation* [replacing each missing value with one imputed value] *does not lead to valid results* because the analyst is making up data

Single Imputation

Some common forms of single imputation include:

- (Unconditional) Mean imputation
- Regression (conditional mean) imputation
- Last observation carried forward (for longitudinal data)
  - Special case: Assume a return to baseline (e.g. in weight loss studies)
- Assume all missing values are informative and identical
  - Current smoker (e.g. smoking cessation studies)
  - Absence of code (e.g. EDW analysis) implies absence of condition

Single Mean Imputation

- Principle: Replace missing values with the average of all observed values

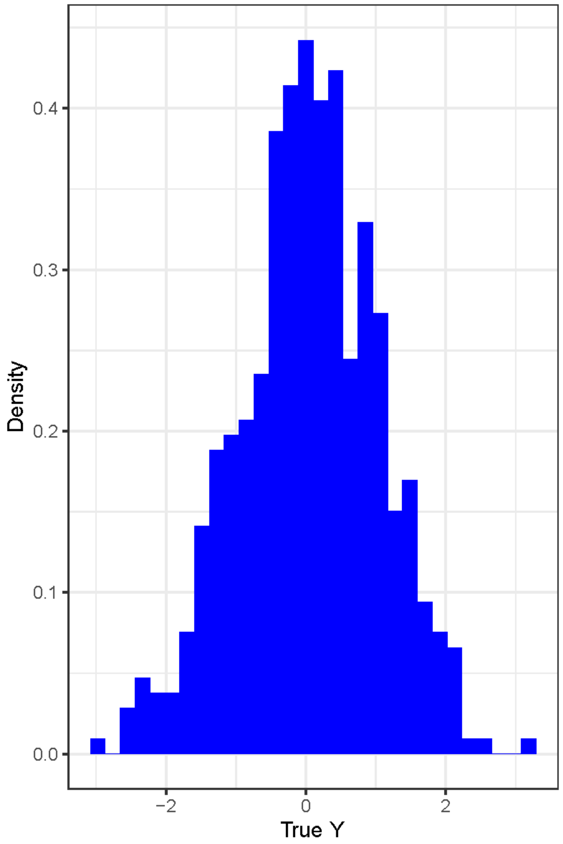
- Pros: Easy to implement

Single Mean Imputation

- Principle: Replace missing values with the average of all observed values

- Pros: Easy to implement

- Cons:
  - **Will create a spike in the data distribution**



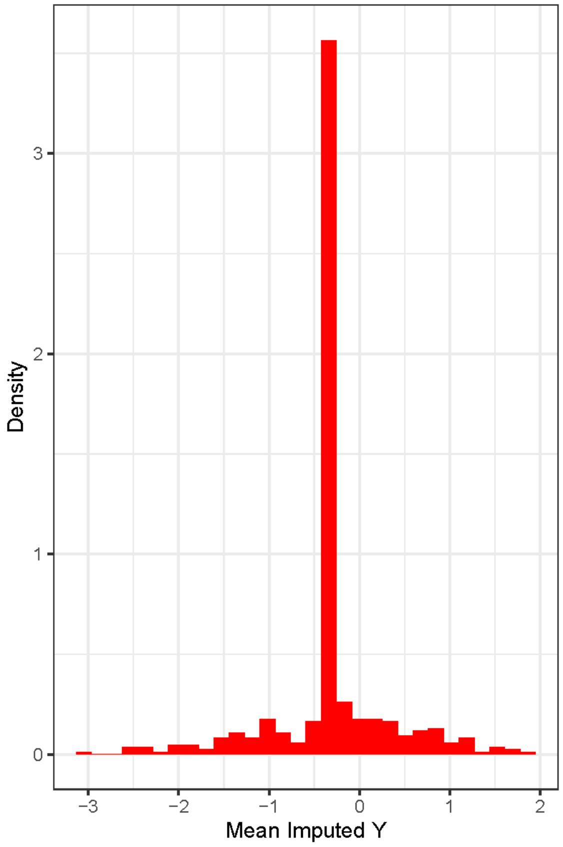
Histogram of observed data

Single Mean Imputation

- Principle: Replace missing values with the average of all observed values

- Pros: Easy to implement

- Cons:
  - **Will create a spike in the data distribution**



Histogram of data after imputation

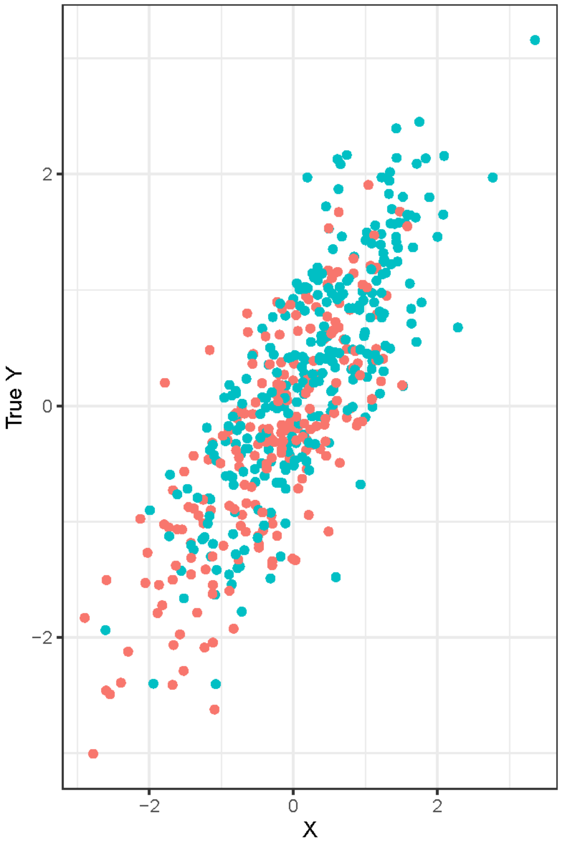
Single Mean Imputation

- Principle: Replace missing values with the average of all observed values

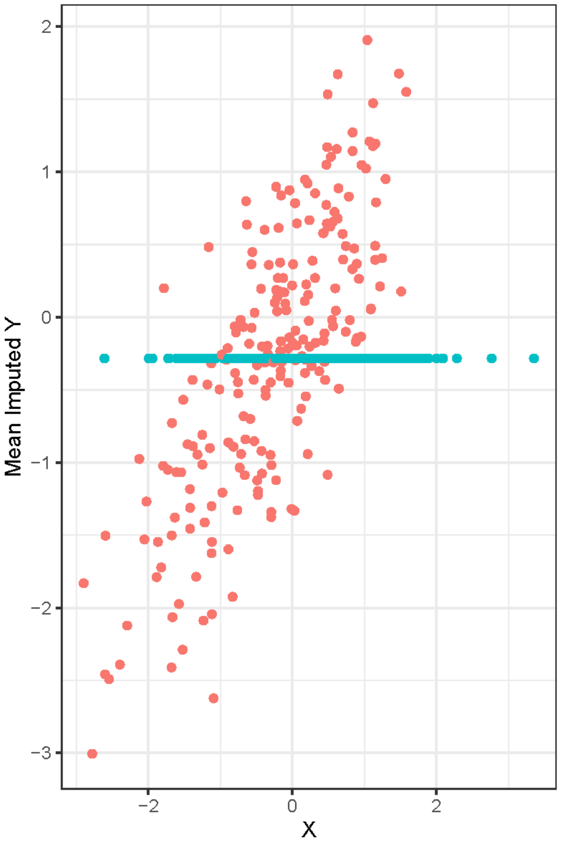
- Pros: Easy to implement

- Cons:
  - Will create a spike in the data distribution
  - **Underestimates the variance**
  - **Attenuates correlations between observed variables**

Single Mean Imputation



Scatterplot of observed data



Scatterplot of data after imputation

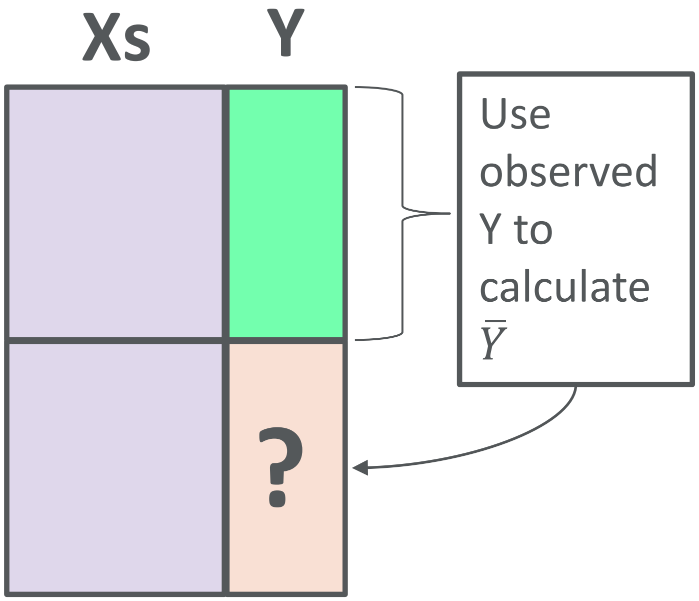
Single Mean Imputation

- The amount of missing data makes a big difference
  - In this example, half of the data were missing! Made a big impact
- This is true for all procedures for handling missing data

"[W]e should not waste a major portion of our resources fixing up a relatively minor problem (for example, do not spend 80% of the budget fixing up the 30% of information that is missing)…"

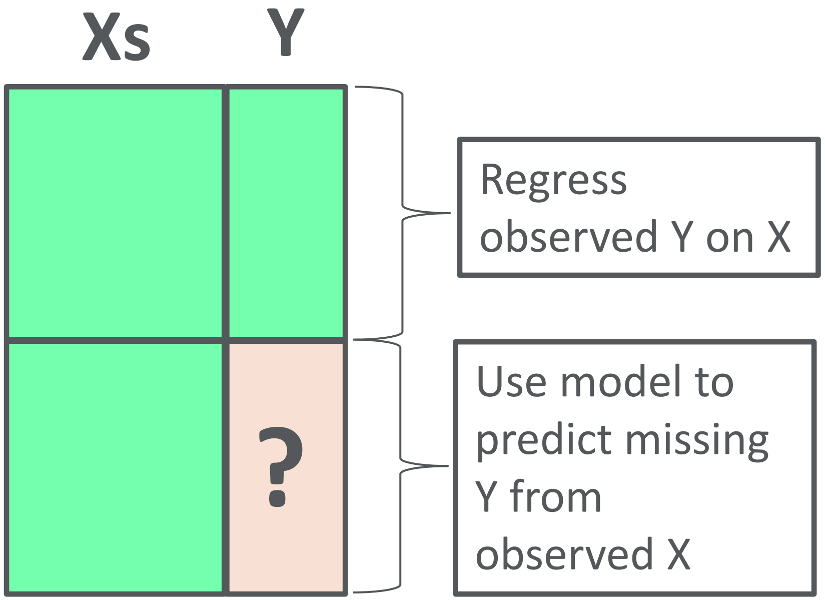
- Rubin & Schenker (1991)

Unconditional versus Conditional Mean Imputation



Use observed Y to calculate $\bar{Y}$

Unconditional Mean Imputation



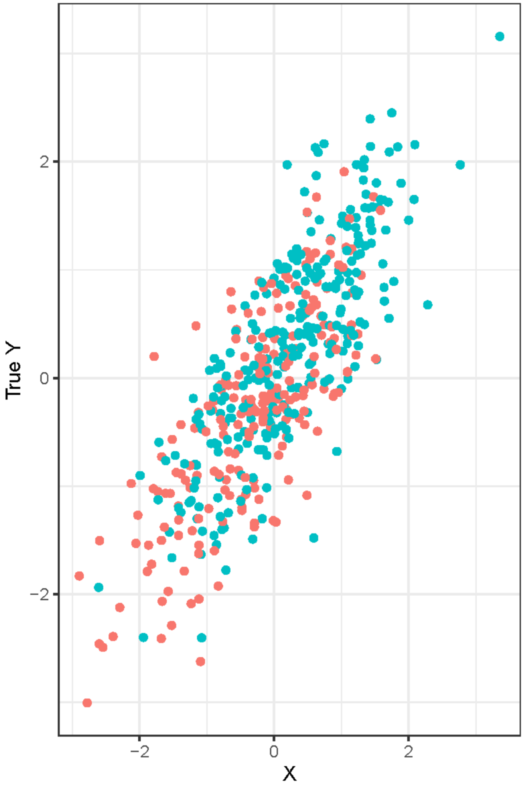
Regress observed Y on X

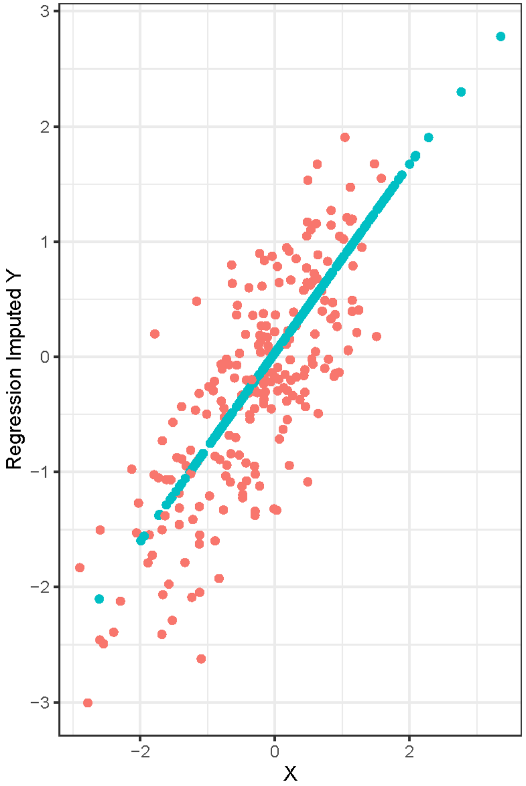
Use model to predict missing Y from observed X

Regression (Conditional Mean) Imputation

Regression (conditional mean) imputation

- Imputing conditional means is better than unconditional means but still not ideal
- We are conditioning on an observed variable (a positive!)



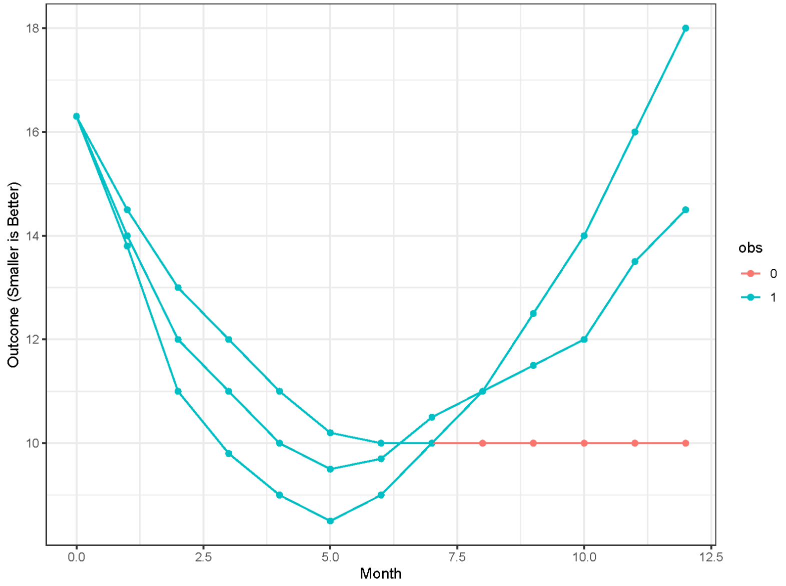


Last observation carried forward (LOCF)

- Commonly used in longitudinal studies, especially with dropout

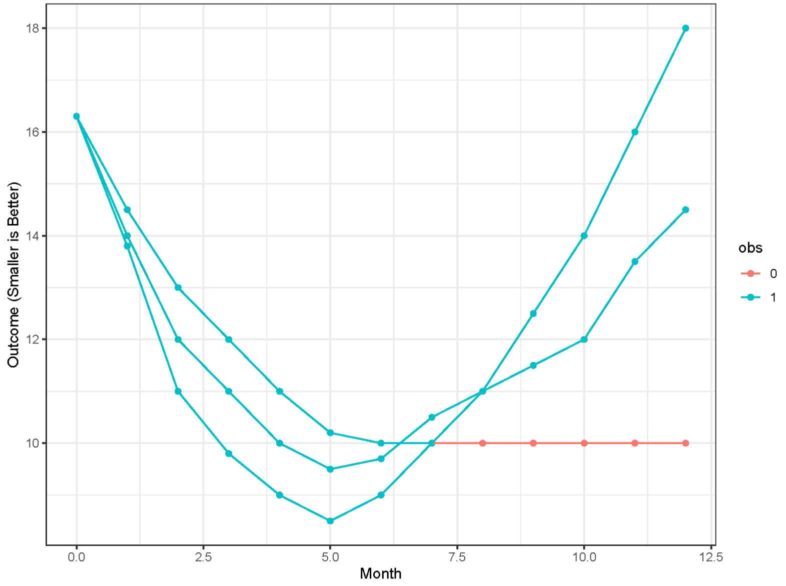
- Idea: After a participant drops out of a study, impute their remaining time points with their last observed value.

- Cons:
  - No variability in imputation (imputes the exact same value at each time point).
  - Ignores any trends (overall or within individual) in the data. Assumes that any progress or deterioration stops at the time of dropout.



Last observation carried forward (LOCF)

- Historically LOCF was recommended by the FDA to handle missing data
- Considered “conservative:" assumes that participants will continue to improve so fixing their follow up values is a “worst case scenario."
- But, in many studies (depression, weight loss, physical activity, etc.) outcomes will initially improve then get worse (regress to the mean)
- Underestimating variability can increase Type 1 error
- If participants in Arm A tend to drop out later than Arm B, between-group comparisons can be confounded with time



Good imputation practices…

- Leverage observed information
  - Condition on observed variables
  - Reduces bias, improves precision
  - Preserves associations between variables with missing and full observed values

- Are multivariate
  - Preserves associations between variables with missing values

- Draw from the predictive distribution of the missing values, not just the means
  - Incorporate information about variability too
  - Estimates are valid over a wide range of estimands (scientific questions)

Good imputation practices…

- Leverage observed information ✓
  - Condition on observed variables
  - Reduces bias, improves precision
  - Preserves associations between variables with missing and full observed values
- Are multivariate ✓
  - Preserves associations between variables with missing values
- Draw from the predictive distribution of the missing values, not just the means ☒
  - Incorporate information about variability too
  - Estimates are valid over a wide range of estimands (scientific questions)

Northwestern Medicine®
Feinberg School of Medicine

Multiple Imputation Methods

Reiterate: Imputation

- **Imputation** is the process of replacing missing values with plausible values
- Single imputation is the easiest way to move forward
- Pros:
  - Once missing values are "filled-in," the data set is rectangular, allowing for analysis using standard methods
- Cons:
  - Single imputation makes no distinction between observed values and imputed values
  - Single imputation only incorporates a one-number summary of the observed data distribution (mean, mode, etc.)

Multiple Imputation (Rubin 1978)

1. For each missing value in the data set, generate *D* imputations to create *D* imputed data sets

2. Analyze each of the *D* datasets separately, generating *D* copies of each parameter estimate

3. Use "Rubin’s rules" to combine the *D* parameter estimates into one final estimate that incorporates the uncertainty due to multiple imputations

Multiple Imputation Example, D = 4

X_1	X_2	X_3	X_4
0	3	8	2
0	4	9	3
0	3	NA	3
0	4	NA	NA
1	5	7	NA
1	NA	5	NA
1	NA	7	NA
1	NA	6	7
1	4	8	7
1	5	NA	9

Original Data

X_1	X_2	X_3	X_4
0	3	8	2
0	4	9	3
0	3	8	3
0	4	8	2
1	5	7	12
1	5	5	9
1	4	7	8
1	3	6	7
1	4	8	7
1	5	7	9

X_1	X_2	X_3	X_4
0	3	8	2
0	4	9	3
0	3	7	3
0	4	8	1
1	5	7	9
1	4	5	9
1	5	7	8
1	5	6	7
1	4	8	7
1	5	6	9

X_1	X_2	X_3	X_4
0	3	8	2
0	4	9	3
0	3	8	3
0	4	7	2
1	5	7	10
1	5	5	9
1	4	7	8
1	5	6	7
1	4	8	7
1	5	8	9

X_1	X_2	X_3	X_4
0	3	8	2
0	4	9	3
0	3	8	3
0	4	9	3
1	5	7	11
1	4	5	9
1	4	7	9
1	5	6	7
1	4	8	7
1	5	7	9

Multiply-Imputed Data ($D = 4$)

Multiple Imputation: Pros and Cons

- Pros:
  - Maintains entire dataset and uses all available information
  - Weak (more plausible) assumptions about the missing data mechanism
    - *Still can only handle MAR (ignorable) mechanisms*
  - Incorporates variability in imputation, resulting in correct confidence intervals
  - Maintains relationships between variables
  - One set of imputed datasets can be used for many analyses

- Cons:
  - Have to rely on modeling assumptions
  - Complex to implement
    - Current software makes this less of an issue!

How many imputations is enough?

- Historically, a small number
  - Rubin initially proposed 3 as sufficient

- Several researchers have recently looked at the influence of *D* on bias and precision
  - Initial conclusions are that benefits can be seen with larger *D* (say, 20-100).

- Theoretically, it is always better to use larger *D,* but this is more computationally intensive and requires lots of storage (usually RAM). Setting *D* high may not be worth the extra wait.

- If calculation is not prohibitive, a rule of thumb is to *set D to the average percentage of missing data*.

Multiple Imputation by Chained Equations (MICE)

- AKA fully conditional specification or sequential regressions
- Imputations are created by drawing from iterated conditional models one-by-one
- Requires a specification of an imputation model for each incomplete variable
- One “iteration” consists of one cycle through all incomplete variables
- The number of iterations is usually low (recommendation is 20)

- Can be implemented in standard software!
  - R (MICE)
  - SAS (PROC MI, PROC MIANALYZE)
  - Stata (mi command)

MICE in “real life” data

- In practice, you will encounter missing values on several different variables
  - These variables may be of different types (continuous vs binary)
  - MICE can handle this!

- The imputation model should:
  - Account for the process that created the missing data
  - Preserve the relationships that already exist in the data
  - Preserve uncertainty about these relationships

- Selecting predictors for imputation models
  - Include variables that will appear in the analysis model to preserve relationships (this includes the outcome!)
  - Include variables related to nonresponse
  - Include variables that explain variance (to get more precise imputations)

Steps to implement MICE in “real life” data

1. Decide what variables to include as predictors into imputation model

2. Decide whether to impute variables that are functions of other variables
   - Sum scores, interactions
   - Can use “passive” imputation to maintain integrity

3. Decide variable imputation order
   - Least missing to most missing
   - Longitudinal data

4. Specify number of iterations
   - At least 20

5. Specify number of imputations
   - At minimum, use average percent missing across all variables

Things to check after using MICE in “real life” data

1. What were the patterns of missing data?
   - Are there variables that were missing in tandem?

2. Are imputed data plausible?
   - What are the ranges?
   - Are the variable types correct (e.g. not continuous when should be binary)?

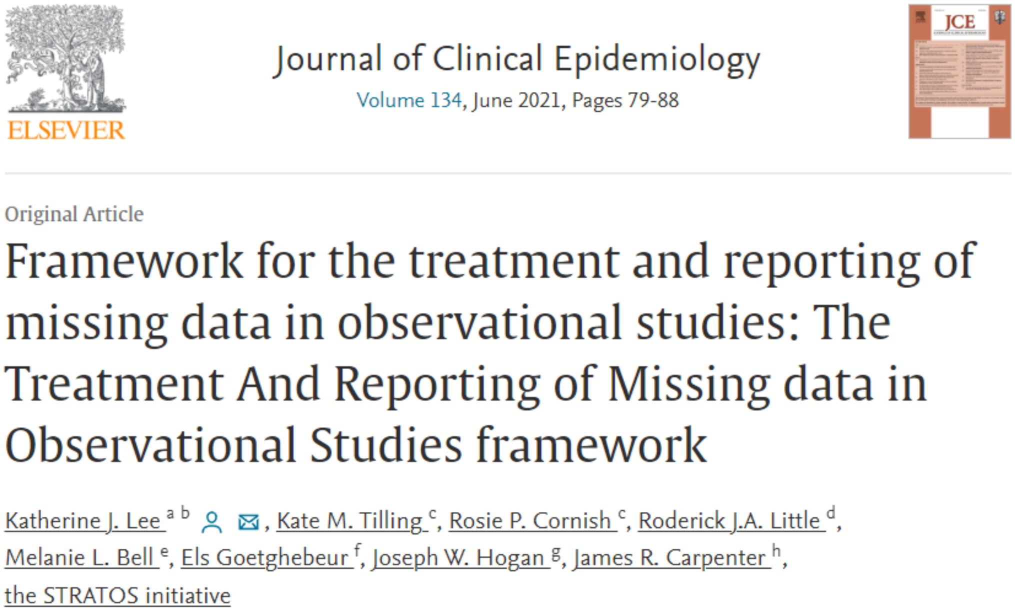
3. Are imputed data consistent with observed data?
   - Check distributions

Reporting Recommendations

- Table that includes
  - What variables were included in the imputation model
  - What percentage missing each had
  - What type of model was used for each variable
  - Any variables created passively afterwards

- Table that compares complete case data to imputed data

- Typically mentioned in "Methods" and details included in supplemental information



Journal of Clinical Epidemiology

Volume 134, June 2021, Pages 79-88

Original Article

Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework

Katherine J. Lee[a] [b], Kate M. Tilling[c], Rosie P. Cornish[c], Roderick J.A. Little[d], Melanie L. Bell[e], Els Goetghebeur[f], Joseph W. Hogan[g], James R. Carpenter[h], the STRATOS initiative

References

- Newgard CD, Lewis RJ. Missing Data: How to Best Account for What Is Not Known. *JAMA*. [https://doi.org/10.1001/jama.2015.10516]

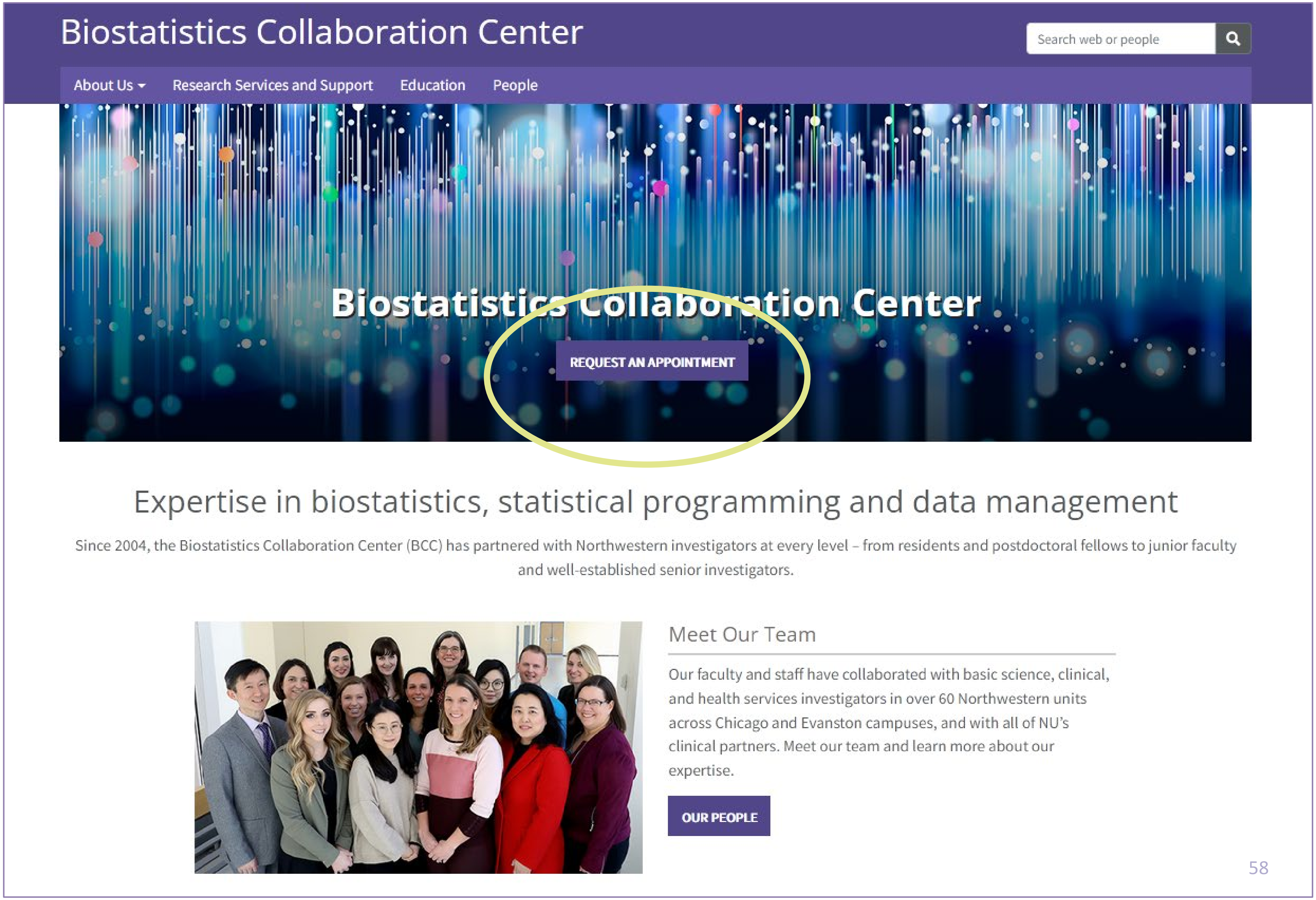
- Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*. [https://doi.org/10.1001/jama.2015.15281]

- Sterne JAC, White IR, et al. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*. [https://doi.org/10.1136/bmj.b2393]

- Little RJ, D'Agostino R, et al. The Prevention and Treatment of Missing Data in Clinical Trials. *NEJM.* [https://doi.org/10.1056/NEJMsr1203730]

- Lee KJ, Tilling KM, et al. Framework for the Treatment and Reporting of Missing Data in Observational Studies: The Treatment And Reporting of Missing Data in Observational Studies Framework. *J Clin Epi*. [https://doi.org/10.1016/j.jclinepi.2021.01.008]

Thank you!



Biostatistics Collaboration Center

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 have collaborated with basic science, clinical, and health services investigators in over 60 Northwestern units across Chicago and Evanston campuses, and with all of NU's clinical partners. Meet our team and learn more about our expertise.