CENTERS FOR MEDICARE & MEDICAID SERVICES

Solving Missing Data Problems

Medicaid Innovation Accelerator Program - Data Analytics National Webinar

October 23, 2018
3:00 PM ET

Logistics for the Webinar

- All lines will be muted
- Use the chat box on your screen to ask a question or leave a comment
    - Note: chat box will not be seen in “full screen” mode
- Slides and a transcript will be posted online within a few weeks of the webinar
- Please complete the post-webinar survey with your feedback at the conclusion of the webinar!

Welcome!

- Jessie Parker, GTL and Analyst on Medicaid IAP Data Analytic Team, Data and Systems Group, CMCS

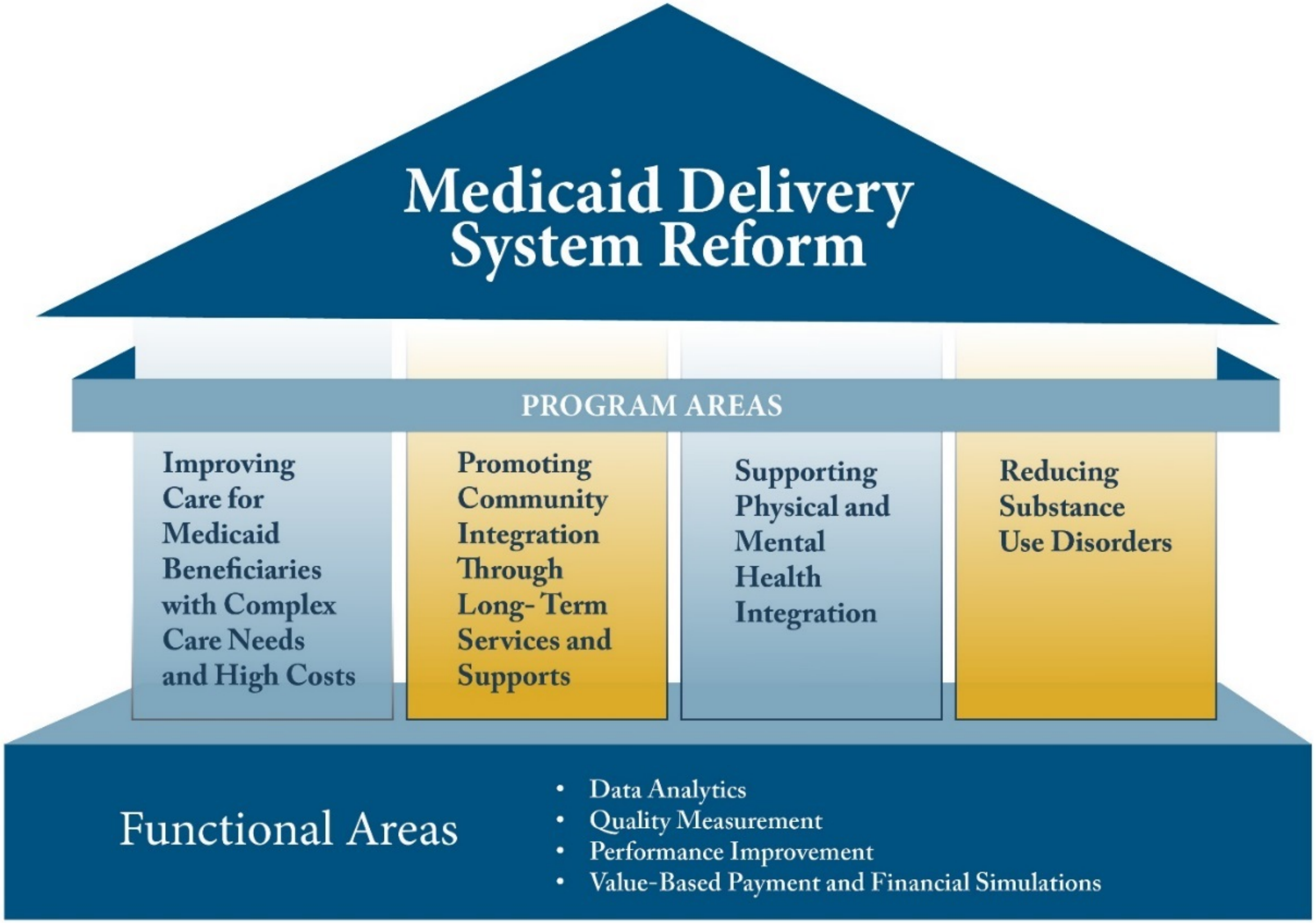
Agenda for Today’s Webinar

- Introduction
- Overview of the Medicaid Innovation Accelerator Program
- Review of Missing Data Problems
- Solving Missing Data Problems by Multiple Imputation
- Ohio Medicaid’s Experience

Today's Speakers

- Frank Yoon, Senior Statistician, IBM Watson Health
- Jonathan Barley, Chief, Bureau of Health Research and Quality, Ohio Medicaid
- Tim Sahr, Director of Research, Government Resource Center, Ohio Colleges of Medicine

Medicaid Innovation Accelerator Program (IAP)


Medicaid Delivery System Reform
PROGRAM AREAS
Improving Care for Medicaid Beneficiaries with Complex Care Needs and High Costs
Promoting Community Integration Through Long-Term Services and Supports
Supporting Physical and Mental Health Integration
Reducing Substance Use Disorders
Functional Areas
• Data Analytics
• Quality Measurement
• Performance Improvement
• Value-Based Payment and Financial Simulations

Goals for Today's Webinar

In this interactive webinar, states will learn about:

- Types and patterns of missing data
- Diagnosing missing data
- Solving missing data
- Ohio Medicaid's approach to addressing missing data in their annual assessment

Solving Missing Data Problems

Frank Yoon, Senior Statistician,
IBM Watson Health

Overview: Missing Data Analysis

- When missing, data element values are *missing*…
  - Completely at random (MCAR)
  - At random (MAR)
  - Not at random (MNAR)
- To solve the missing data problem, we first need to consider:
  - What does the pattern tell us about our data?
  - When is it appropriate to adjust for missing values?

Types of Missing Data

Pattern	National Academy of Science's Definition	Takeaways
Missing Completely At Random (**MCAR**)	The missing data are unrelated to the study variables	- Available data is an unbiased random sample - Usually an unrealistically strong assumption
Missing At Random (**MAR**)	Whether or not data are missing *does not depend on the values of the missing data*	- Need to address; do not need to understand mechanism. - Data we observe can predict the data we cannot
Missing Not At Random (**MNAR**)	Whether or not data are missing *depends on the values of the missing data*	- The mechanism cannot be ignored - The mechanism must be modeled

Adapted from: *https://onbiostatistics.blogspot.com/2012/10/missingness-mechanism-mcar-mar-and-mnar.html*

Example: Diagnosing Missing Data

Task: Predict annual healthcare costs adjusting for items in administrative and clinical screening data

- Key intake items:
  - **Sociodemographic:** race and ethnicity
  - **Clinical**: mental health and substance use disorders; other comorbidities
  - **PHQ-9:** depression risk

Example: Outcomes and Covariates

Costs ($)	Sex	Age	Race/ Ethnicity	Mental Illness	Substance Use Disorder	PHQ-9
$$$	M					?
$$$	F					
$$$	M					?
$$$	F		?			
$$$	M					
$$$	F					?
$$$	M		?			?
$$$	F		?			
$$$	M					?
$$$	F					

Example: Choosing the Approach

- **Poll:** What type of missing data are we dealing with?

  a) Analyst spills coffee on client intake forms

  b) Men are less likely to complete a PHQ-9

  c) High utilizers of health services do not report race and ethnicity

Analytic Solutions

- Complete case analysis
  - When can we use it?
  - When should we definitely avoid it?
- Imputation
  - Mean imputation
  - Regression imputation
- Multiple imputation
  - Account for uncertainty in imputing missing values
  - Leverage computational power

Example: National Health and Nutrition Examination Survey (NHANES)

```
> library(mice)
> head (nhanes)
#      age    bmi   hyp    chl
1        1     NA    NA     NA
2        2   22.7     1    187
3        1     NA     1    187
4        3     NA    NA     NA
5        1   20.4     1    113
6        3     NA    NA    184
```

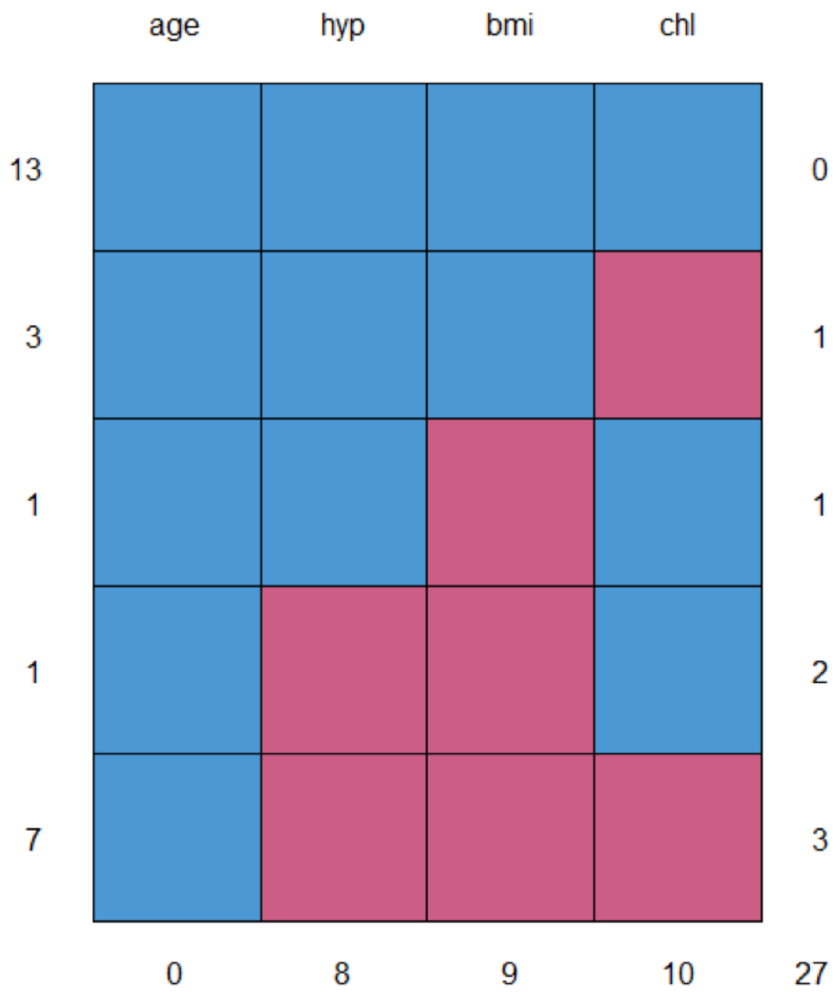
- Age Group
- Body Mass Index (bmi)
- Hypertension Status (hyp)
- Cholesterol (chl)

```
> dim (nhanes)
{1}  25  4
```

```
>summary (nhanes)
         age                    bmi                     hyp                     chl
Min      :1.00       Min      :20.40       Min      :1:000       Min      :113.0
1st Qu.  :1.00       1st Qu.  :22.65       1st Qu.  :1.000       1st Qu.  :185.0
Median   :2.00       Median   :26.75       Median   :1.000       Median   :187.0
Mean     :1.76       Mean     :26.56       Mean     :1.235       Mean     :191.4
3rd Qu.  :2.00       3rd Qu.  :28.93       3rd Qu.  :1.000       3rd Qu.  :212.0
Max.     :3.00       Max.     :35.30       Max.     :2.000       Max.     :284.0
                     NA's     :9           NA's     :8           NA's     :10
```

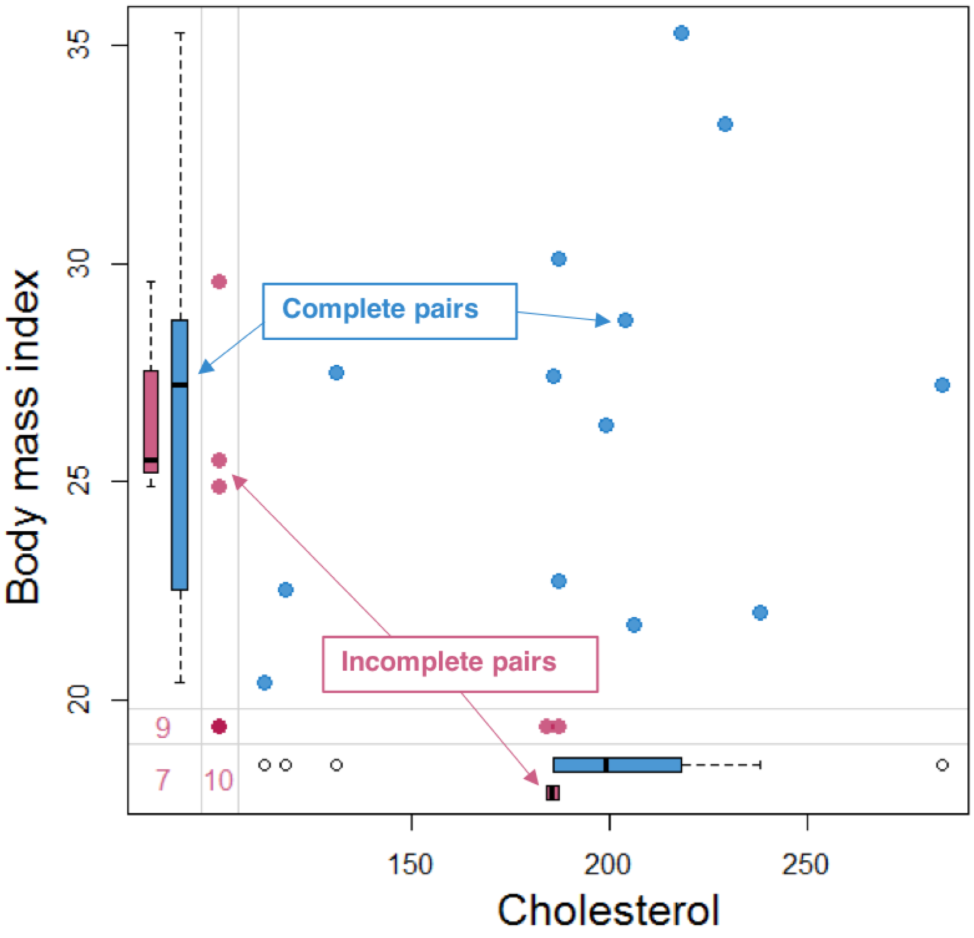
Examine Missing Data

```
> md.pattern(nhanes)
        age  hyp  bmi  chl
13        1    1    1    1    0
3         1    1    1    0    1
1         1    1    0    1    1
1         1    0    0    1    2
7         1    0    0    0    3
          0    8    9   10   27
```



Examine Missing Data *cont.*

- Assess complete and incomplete pairs
- Use *margin plot* (`VIM` R package)

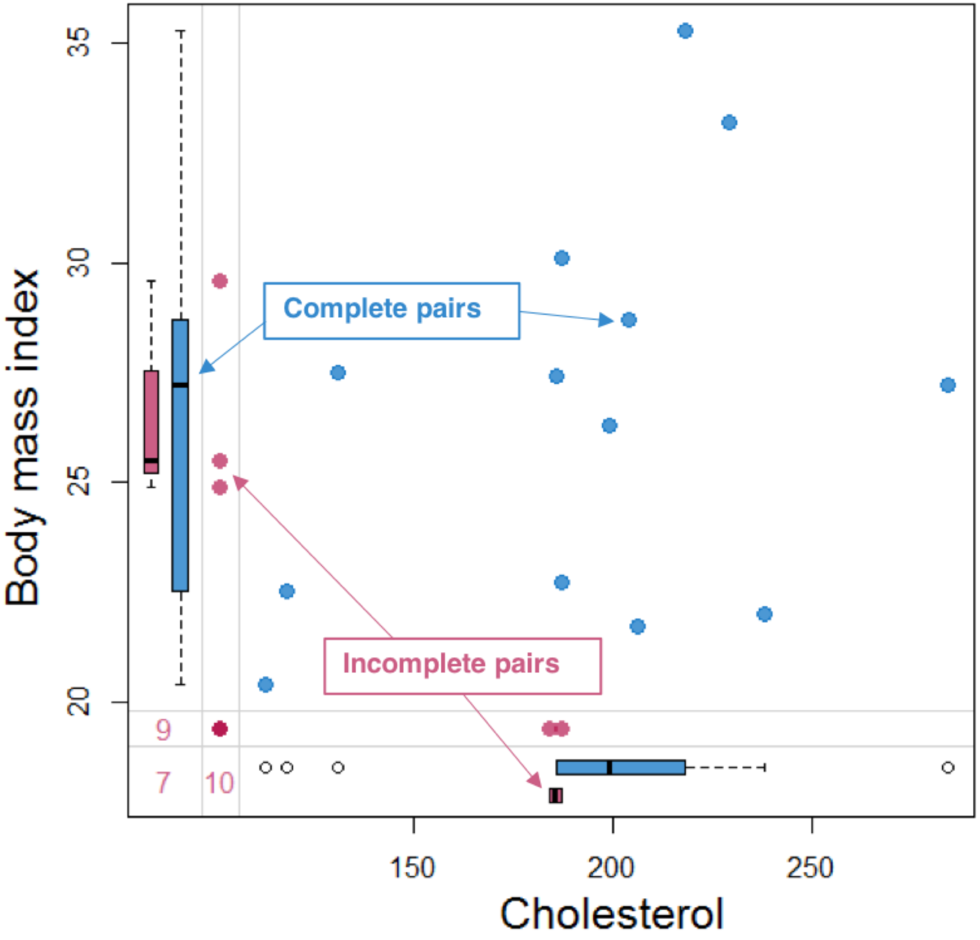


Complete pairs

Incomplete pairs

Examine Missing Data - Poll

- **Poll:** How should the boxplots look under MCAR?
  a) Alike
  b) Different
  c) Not sure; need more information



Elementary Solutions

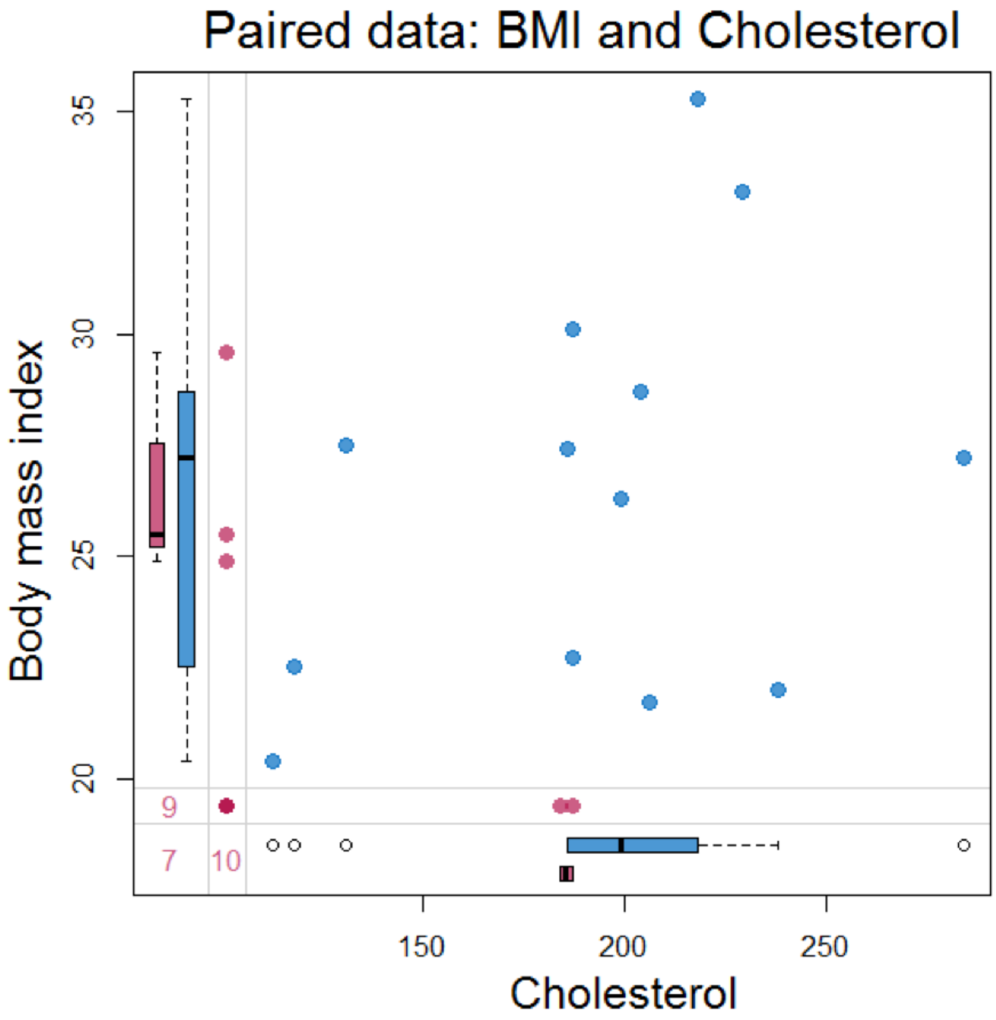
- Complete case analysis
  - Listwise deletion
  - Pairwise deletion
- Single imputation
  - Mean imputation
  - Regression imputation
- What could go wrong?

Elementary Solutions *cont.*

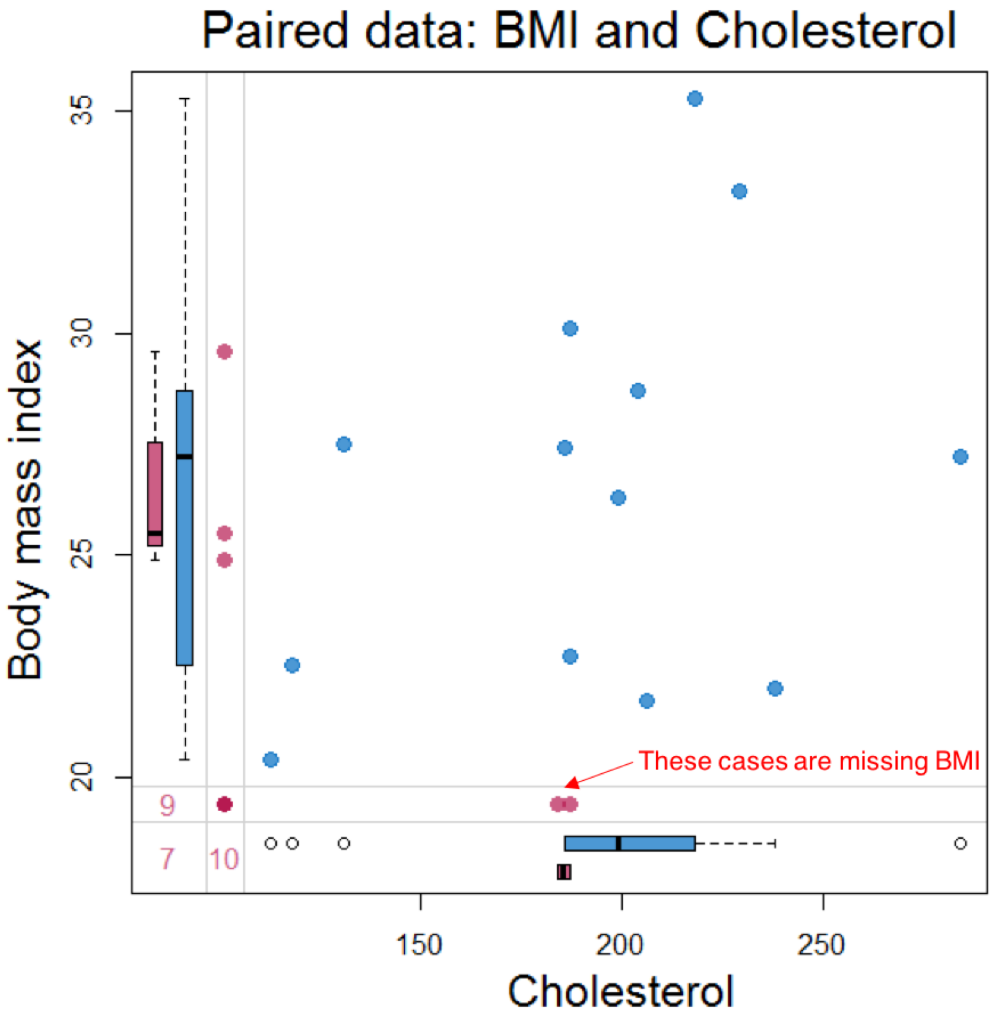
- Complete case analysis
  - Assumes complete cases accurately represent incomplete cases (with missing data)
  - For what pattern of missing data is this appropriate?

- Single imputation
  - Assumes that we can calculate missing values from available data – i.e., observed predictors from complete cases
  - For what pattern of missing data is this appropriate?
  - What is the inherent statistical limitation?

Single Imputation

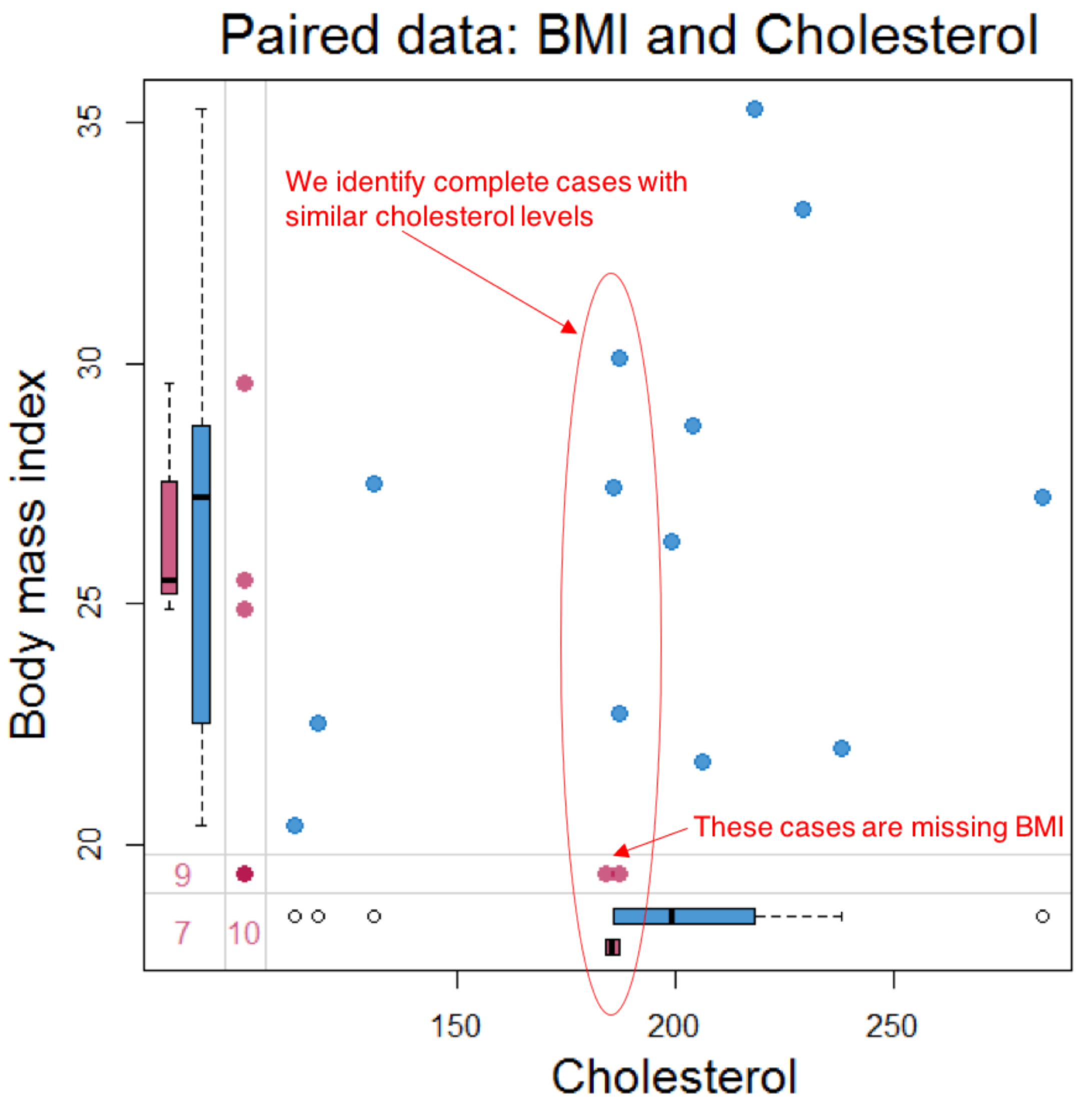


Single Imputation

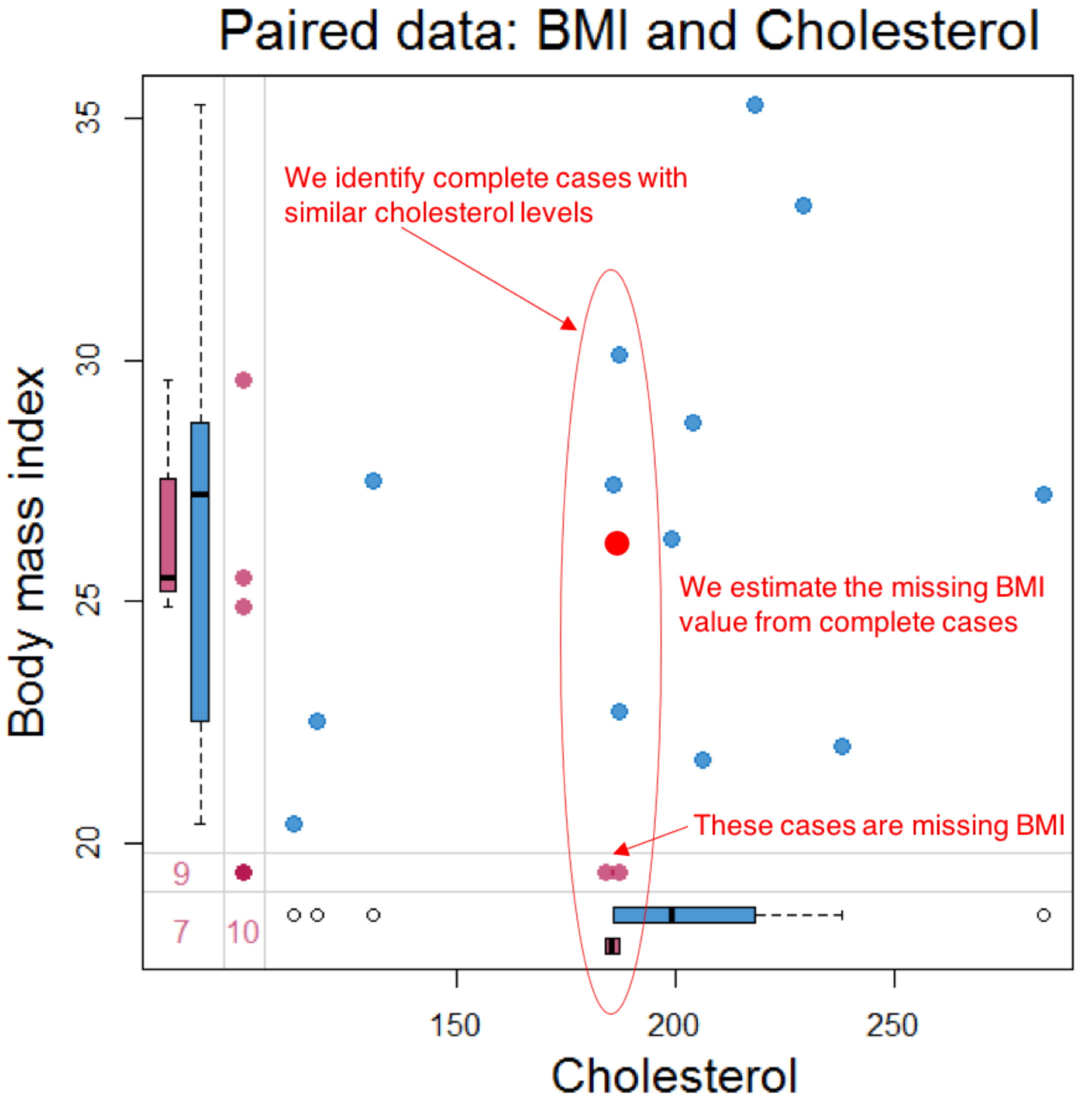


These cases are missing BMI

Single Imputation



Single Imputation



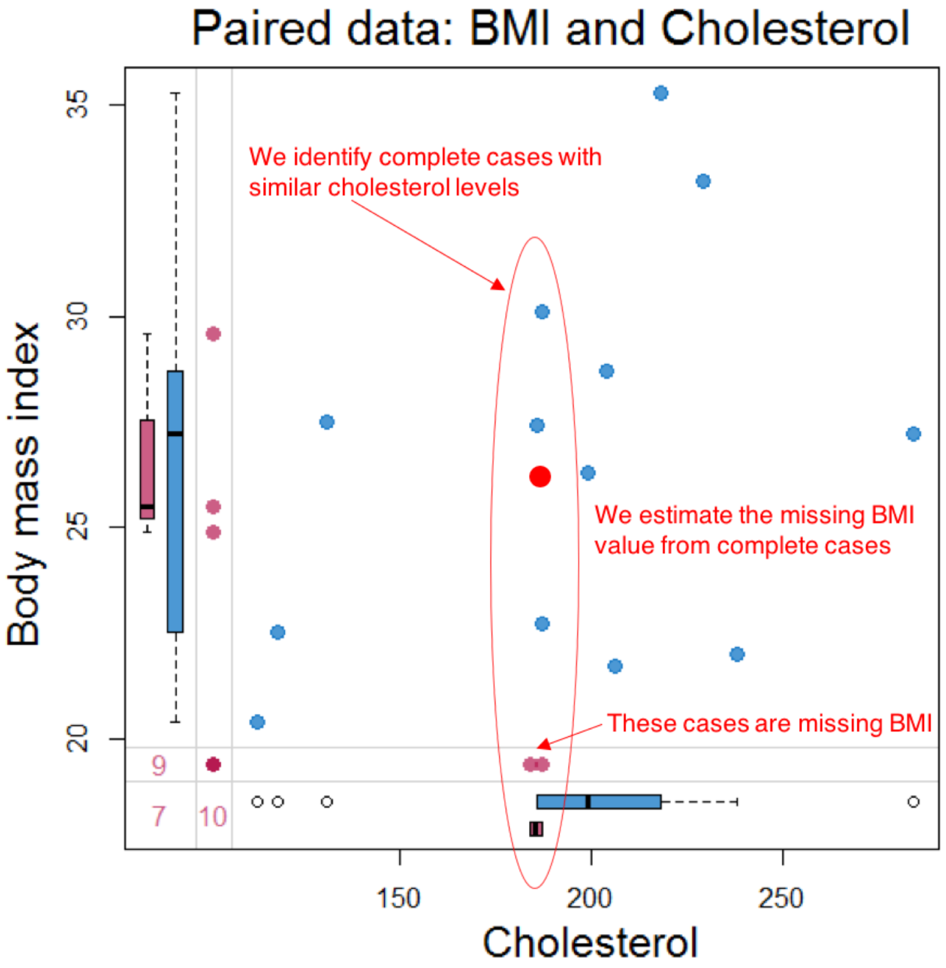
We identify complete cases with similar cholesterol levels

We estimate the missing BMI value from complete cases

These cases are missing BMI

Single Imputation

- **Poll:** What can go wrong here?
  a) Nothing, it’s perfect
  b) There’s just one imputed data point
  c) We have to assume MAR and not MNAR



Paired data: BMI and Cholesterol

We identify complete cases with similar cholesterol levels

We estimate the missing BMI value from complete cases

These cases are missing BMI

Elementary Solutions

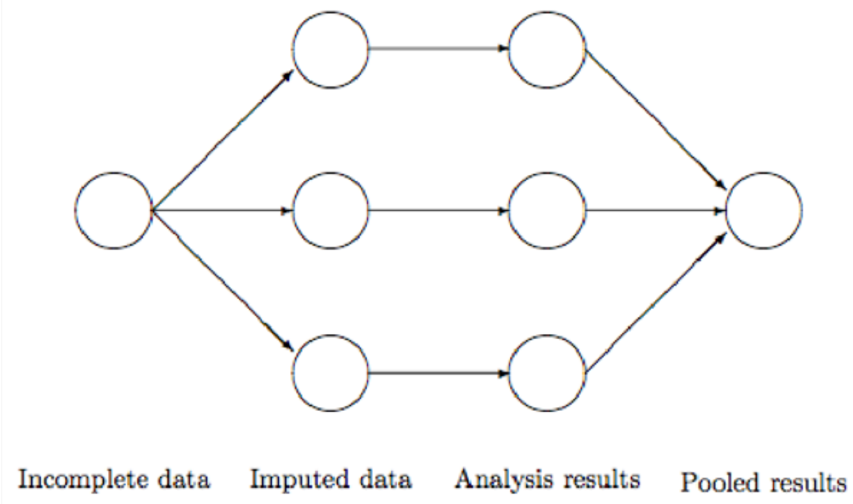
- Reminder: assess the pattern!

age	bmi	hyp	chl
1	NA	NA	NA
1	NA	1	187
1	20.4	1	113
1	22.5	1	118
1	30.1	1	187
1	NA	NA	NA
1	29.6	1	NA
1	NA	NA	NA
1	35.3	1	218
1	NA	NA	NA
1	33.2	1	229
1	27.5	1	131

age	bmi	hyp	chl
2	22.7	1	187
2	22.0	1	238
2	NA	NA	NA
2	NA	NA	NA
2	28.7	2	204
2	26.3	2	199
2	27.4	1	186
3	NA	NA	NA
3	NA	NA	184
3	21.7	1	206
3	27.2	2	284
3	25.5	2	NA
3	24.9	1	NA

Applying Multiple Imputation

1. “Fill in” missing values, multiple times from a data distribution
2. Analyze the multiple data sets, each containing filled-in values
3. Pool analyses, combining filled-in values to incorporate statistical uncertainty



Source: https://www.stefvanbuuren.name/mice/

Applying Multiple Imputation (in R)

- Multivariate imputation by chained equations
- mice package in R provides functionalities:
  - Inspect the missing data pattern
  - Impute the missing data *m* times
  - Diagnose the quality of the imputed values
  - Analyze each completed data set
  - Pool the results of the repeated analyses
  - Store and export the imputed data in various formats
  - Generate simulated incomplete data
  - Incorporate custom imputation methods

Source: https://cran.r-project.org/web/packages/mice/index.html

Impute Missing Values

```
> imp <- mice(nhanes, m = 5, seed = 35472)

> print(imp)

Class: mids
Number of multiple imputations:  5
Imputation methods:
  age   bmi   hyp   chl
   "" "pmm" "pmm" "pmm"
PredictorMatrix:

           age     bmi     hyp     chl

 age         0       1       1       1
 bmi         1       0       1       1
 hyp         1       1       0       1
  chl        1       1       1       0
```

- Default is *predictive mean matching* (pmm)

Diagnose Imputed Values

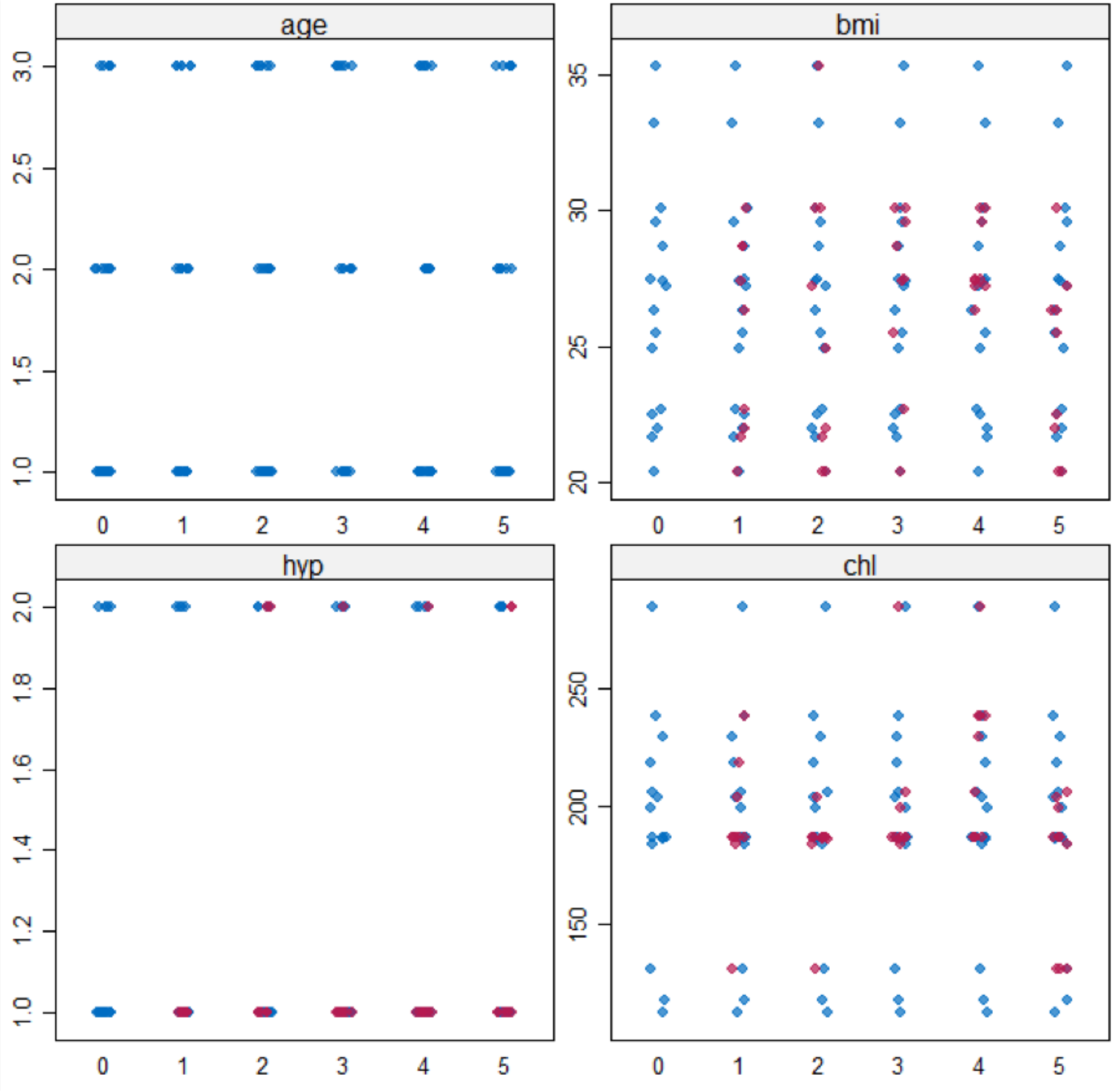
- Imputed values should “look like” the original data – i.e., have similar scales, distributions
- In other words, imputed values should be plausible
  - Negative BMI?
  - Outlying age?

```
 > imp$imp$bmi

item     1    2    3    4    5
1      27.4 35.3 30.1 30.1 27.2
3      30.1 30.1 29.6 27.2 26.3
4      21.7 24.9 22.7 27.5 20.4
6      20.4 20.4 25.5 27.4 22.5
10     22.7 20.4 27.4 27.2 25.5
11     26.3 30.1 28.7 27.5 26.3
12     22.0 22.0 20.4 26.3 20.4
16     28.7 21.7 30.1 29.6 30.1
21     28.7 27.2 27.5 30.1 22.0
```

Diagnose Imputed Values *cont.*

- Strip plot
- Compare observed vs. imputed values



Polling Time

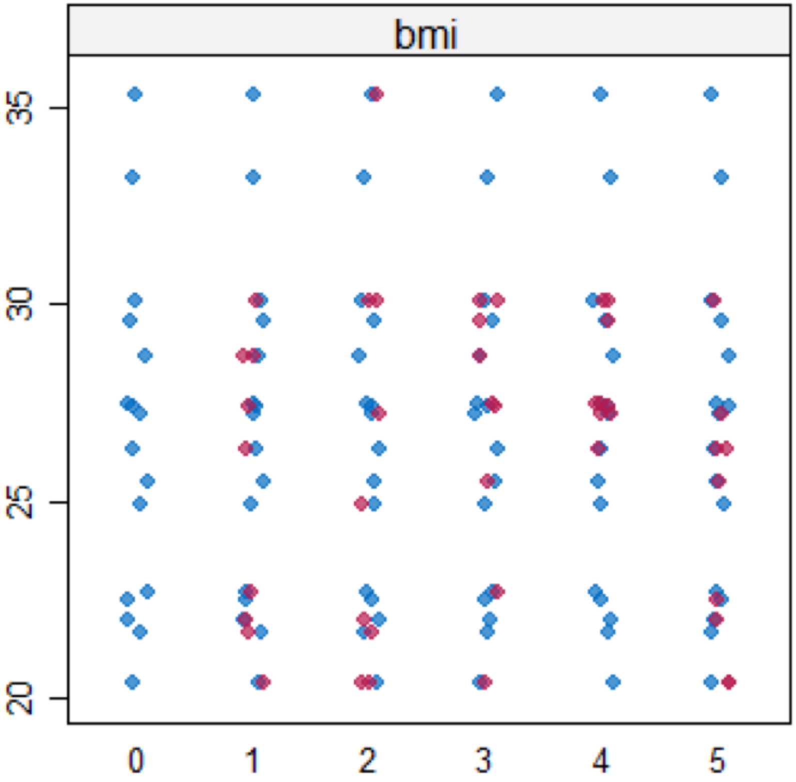
- **Poll:** What should we see here?

  a) Imputed look similar to observed values

  b) There is variation in the imputations

  c) Imputed values appear stable over iterations

  d) All of the above



Analysis of Imputed Data

- Linear regression of cholesterol on age and BMI

```
> imp <- mice(nhanes, m = 5, seed = 35472)
> fit <- with(imp, lm(chl ~ age + bmi))
> round(summary(pool(fit)), digits = 2)
```

```
item          estimate   std. error   statistic      df   p. value
{intercept}      -26.5        55.25       -0.48   18.49       0.64
age              34.71         8.36        4.15   18.19       0.00
bmi               5.97         1.79        3.34   17.89       0.00
```

- Increase the number of imputations

```
> imp <- mice(nhanes, m = 50, seed = 35472)
> fit <- with(imp, lm(chl ~ age + bmi))
> round(summary(pool(fit)), digits = 2)
```

```
item          estimate   std. error   statistic      df   p. value
{intercept}      -8.78        72.51       -0.12   11.76       0.91
age              33.12        11.92        2.78   10.70       0.02
bmi               5.46         2.25        2.43   12.87       0.03
```

Analysis of Imputed Data *cont.*

- **Poll:** What do we need to do differently when regression modeling with multiply imputed data?

  a) We need to double check our coefficient estimates to make sure they agree over multiple imputations

  b) We need to throw out results that don't look right

  c) We need to combine results over multiple imputations in order to calculate standard errors of estimated coefficients in the regression model

Further Study

- Predictive mean matching uses the observed data distribution

1. Regress “missing” variable on complete data
2. With estimated regression, predict new values for missing variable (across all observations)
3. For *each* missing case, identify complete cases with predicted values that are close to its predicted values
4. Randomly choose one complete case to impute or fill in the value for each missing case
5. Repeat *m* times

Practical Considerations

- Modeling choices
  - Options include: classification and regression trees, Bayesian, bootstrapping, interaction terms…
  - Consider your predictive needs, choose a few methods, and evaluate those candidates
- Standing by your diagnosis
  - Key assumption is *ignorability*
  - Are values MAR or MNAR?
- Other software
  - SAS – IVEware
  - Stata – mi
  - Check default settings!

Recap: Solving Missing Data Problems

- Why do we care about missing data?
- How do we diagnose it?
- How do we solve it?

2016 Ohio Medicaid Group VIII Assessment Methodology

General Design and Development Considerations, 1

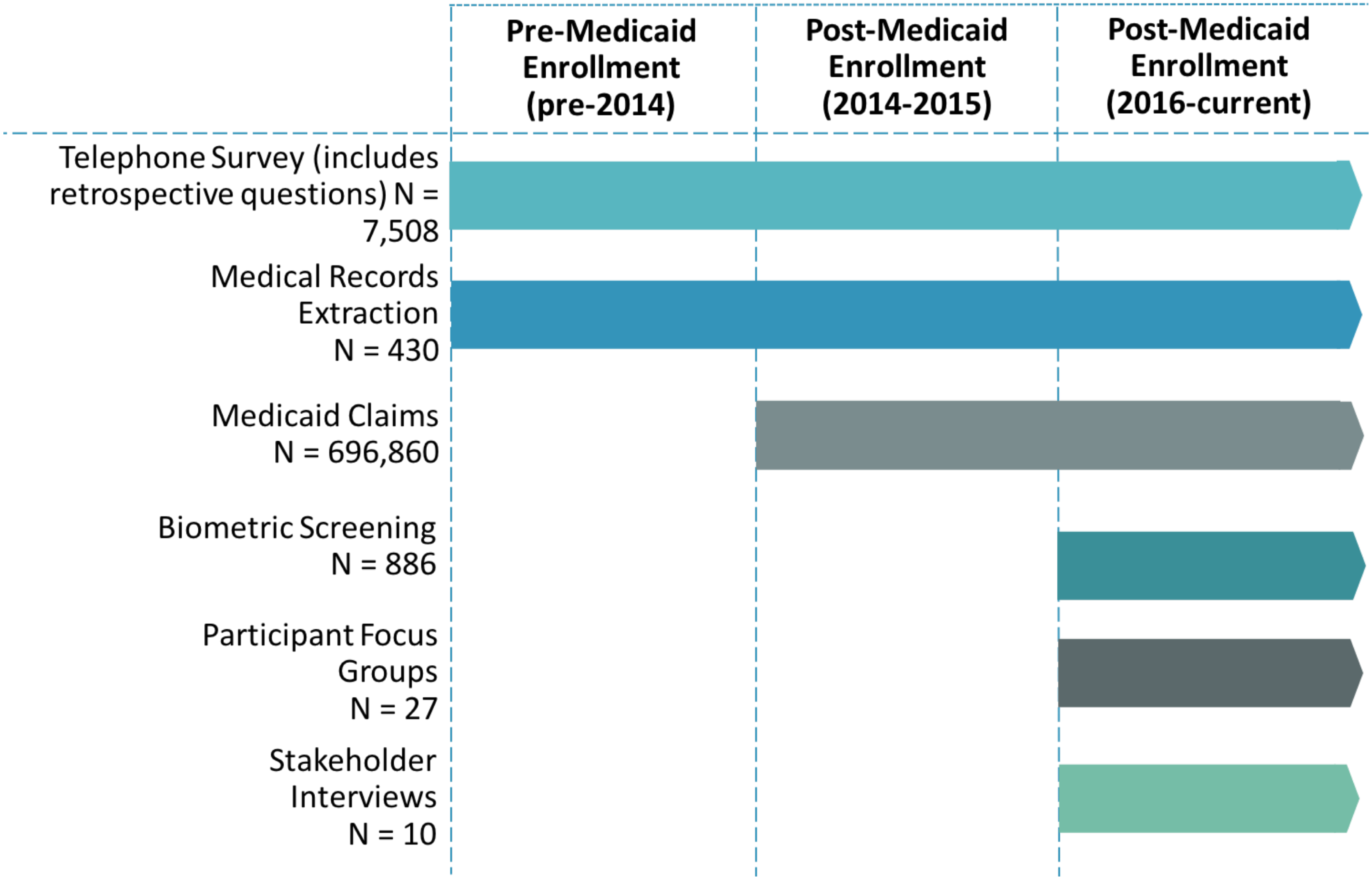
Population comparison, 2016 G-VIII: To examine the differences in health and socioeconomic statuses for the newly Medicaid enrolled, comparing pre-expansion health and socioeconomic statuses to post G-VIII enrollment statuses – over a period going back to 2013 (primarily within group analyses).

General Design and Development Considerations, 2

- **Selection of study design:** G-VIII data collection was setup to afford verification of Medicaid administrative data results using multiple data source validation – how well measures from varying collection modes indicate common findings (e.g., health statuses, access to care, socioeconomic benefit). (Salkind, 2010; Ansari et al., 2014; Carmine & Zeller, 1979)

- **The modes of data collection were:**
  - » Medicaid administrative data/billing data
  - » A G-VIII survey of approximately 20 minutes with sample drawn from Medicaid data – domains included access to care, utilization of health services, self-reported chronic and acute health conditions, self-rated physical health, self-rated mental health, family stressors and issues, employment, resources, and demographics
  - » Biometric measures – heart, weight, height, mental health screener, health check, etc.
  - » Medicaid records extractions – primarily medical history and current conditions
  - » Qualitative interviews of G-VIII enrollees – perceptions of benefits, access to care issues, barriers to care (i.e., transportation, appointment waiting times)
  - » Stakeholder interviews

2016 G-VIII Study Design



Pre-Medicaid Enrollment (pre-2014)

Post-Medicaid Enrollment (2014-2015)

Post-Medicaid Enrollment (2016-current)

Telephone Survey (includes retrospective questions) N = 7,508

Medical Records Extraction N = 430

Medicaid Claims N = 696,860

Biometric Screening N = 886

Participant Focus Groups N = 27

Stakeholder Interviews N = 10

2016 G-VIII Weighting

- **Target population: The target population is all pre-expansion and post-expansion Medicaid enrollees in Ohio – the sampling population consisted of Medicaid enrollees who met the eligibility criteria as defined in the sampling plan.**

- **The survey was selected using a stratified simple random sample. There were 400 strata levels. The base weight was the inverse probability of selection within each stratum.**

UWE for the telephone survey

Telephone Weights	**Group VIII**	**Pre-expansion**
Design-based	1.19	1.35
Nonresponsive adjustment	1.22	1.29
Post-stratification	1.28	1.34

- **All Analyses were weighted to represent the entirety of the Group VIII study eligible population.**

2016 G-VIII Imputation: For Select Variables

- Imputation was conducted on survey variables needed for weighting as well as a few derived variables. All variables that were required in the weighting process had less than 5% missing data. Because of the low level of item nonresponse, a conditional stochastic imputation was conducted. Each variable imputed was conditioned on the age category, and sex of the respondent. The following variables were imputed:
  - Race (most missing and categorical)
  - Ethnicity (Hispanic origin)
  - Marital Status
  - Education
  - Chronic health conditions status
  - Smoking status

2016 G-VIII Methods Summary

- The 2016 Ohio Medicaid Group VIII Assessment Study (G-VIII) was a mixed data study using Medicaid administrative data, survey data, biometric measures, medical records data, and qualitative data.

- The 2016 G-VIII data editing processes enabled analyses across strata and adjusted for non-response bias, missing data, and coverage bias. Validity and reliability checks compared content (across data sets) and measurements (quantitative techniques that primarily checked for design errors).

- Missing data with the 2016 G-VIII data sets was very minor. For survey data stochastic and modeled imputation was used to assist weighting and completeness of select demographic, chronic disease, and health risk behavior variables.

- Data was analyzed using SAS Enterprise, Stata 15 MP, and R-system analytical software.

- Main findings were checked for content validity and interpretation using qualitative interviews of Group VIII enrollees and variation post analyses.

Q&A

Takeaways

- To solve the missing data problem, we first need to consider:
  - What does the pattern tell us about our data?
  - When is it appropriate to adjust for missing values?
- Analytic solutions may include complete case analysis, mean or regression imputation, or multiple imputation, depending on the pattern of missing data.
- Findings may be checked for content validity and interpretation using qualitative interviews and variation post analyses.

Thank You

Thank you for joining today’s webinar!

Please take a moment to complete the post-webinar survey.

We appreciate your feedback!

For more information & resources, please contact MedicaidIAP@cms.hhs.gov