An Introduction to Missing Data Problems

Medicaid Innovation Accelerator Program - Data Analytics National Webinar

June 7, 2018 3:00 – 4:00 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
- The Issue of Missing Data
- Types and Patterns of Missing Data
- Alabama Medicaid’s Experience with Missing Data
Today’s Speakers

- Thomas Flottemesch, Senior Research Leader, IBM Watson Health
- Chris McInnish, Director of Quality Analytics, and Drew Nelson, MPH, Director of Quality Assurance Division, Alabama Medicaid Agency
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:

- Challenges presented by missing data
- Types or patterns of missing data
- Alabama Medicaid’s approach to addressing missing data in their analysis of maternity care delivery
An Overview of Missing Data

Challenges, Patterns and Strategies

Thomas Flottemesch, Senior Research Leader, IBM Watson Health
The Issue of Missing Data

- Missing data refers to:
  - Variables within an observation that have no data when they should (*incomplete observations*).

- Missing data *does not refer to*:
  - Observations with no data (*missing observations*)
  - Unobserved and/or unobservable variables

- With missing data we need to know:
  - Are the missing data informative about the data we observe?
  - Could ignoring the missing data lead to incorrect conclusions?
Process of Statistical Analysis

- **Statistical Analysis:**
  1. Understand a *data generating process*
  2. Make *probabilistic inference(s)*
     - What happened?
     - How did it happen?
     - Did it happen differently?
     - Will it happen again?

- **Missing Data:**
  - **Pattern:** Which values are missing?
  - **Mechanism:** How is the pattern related to the observed data?
Structure of Missing Data

- **Univariate Missing Data:**
  - A single variable is missing data
  - Most critical when that variable is our outcome or factor of interest

- **Multivariate Missing Data:**
  - Multiple variables are missing data within and across observations
  - How data are missing informs our strategy:
    - Monotone or “Nearly” Monotone versus Arbitrary Missing

- *Different variables may have different patterns of missingness*
Dataset with a Monotone Structure

- The missingness can be addressed sequentially
  - “Conserve” information and use most effectively

- In the above table:
  1. Use Group D to inform Group C => C*
  2. Use Groups D and C* to inform Group B => B*
  3. Use Groups D, C*, and B* to inform Group A => A*

- How to do it:
  - **SAS:** PROC MI: PROC FREQ statement and Monotone Statement
  - **R:** MICE Library: md.pattern() AND mice() has visitSequence=“monotone” or “revmonotone” options
## Rubin’s (1976) typology of mechanisms

<table>
<thead>
<tr>
<th>Pattern</th>
<th>National Academy (US)</th>
<th>Takeaways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random (MCAR)</td>
<td>The missing data are unrelated to the study variables.</td>
<td>• Available data is an unbiased random sample.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <strong>Usually an unrealistically strong assumption.</strong></td>
</tr>
<tr>
<td>Missing At Random (MAR)</td>
<td>Whether or not data are missing does not depend on the values of the missing data.</td>
<td>• Need to address; do not need to understand mechanism.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data we observe can predict the data we cannot.</td>
</tr>
<tr>
<td>Missing Not At Random (MNAR)</td>
<td>Whether or not data are missing depends on the values of the missing data.</td>
<td>• The mechanism cannot be ignored</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The mechanism must be modeled.</td>
</tr>
</tbody>
</table>

Adapted from: [https://onbiostatistics.blogspot.com/2012/10/missingness-mechanism-mcar-mar-and-mnar.html](https://onbiostatistics.blogspot.com/2012/10/missingness-mechanism-mcar-mar-and-mnar.html)
Consider the following...

**Task:** Predict annual healthcare utilization (costs) adjusting for items on an outpatient clinic form.

- **Key Intake Items:**
  - **Family History:** Cancers, Diabetes, other risk factors
  - **PHQ-9:** Indicates depression risk
  - **Smoking Status:** Health risk and target for potential behavior intervention

- *Data were missing for several of these items across enrollees*
### Collected Data (Example)

<table>
<thead>
<tr>
<th>Costs ($)</th>
<th>Sex</th>
<th>Cancer</th>
<th>DM</th>
<th>Asthma</th>
<th>MI</th>
<th>PHQ-9</th>
<th>Smoking Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>$$$</td>
<td>M</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>F</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>M</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$$$</td>
<td>F</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DM: Diabetes Mellitus; MI: Mental Illness; PHQ-9 Depression Screen
Pertaining to our analysis…

- Family History, PHQ-9, and smoking status have missing values.
  - If patients do not know Family History, 
    - *Missing Completely At Random*
  - If men are less likely to complete a PHQ-9
    - *Missing At Random*
    - If Smokers tend not to share smoking status
      - *Missing Not at Random*
  - Missingness mechanism(s) are ignorable
  - Model the missingness mechanism(s)
    - Selection and/or Pattern Mixture models
Illustration: Missing Completely at Random (MCAR) & Maternal Episode

• **Task:** Develop risk-adjusted estimates of maternity episodes using *only* one year (Jan-Dec) of claims data
  
  – **Risk Adjusters:** Maternal health, gestational age, prenatal care initiation
  
  – **Missing Data Issue:** “Complete” episodes are available only for start of year

• **Why are the data MCAR?**
  
  – When pregnancy occurs is probably independent from the healthcare used.
  
  – *Rubin:* Missing data are not related to the study variables.
**Illustration: MCAR and Maternal Episode**

**Strategy 1 (Complete Case Analysis):** Use only the complete episode.

<table>
<thead>
<tr>
<th>Jan-Mar</th>
<th>Apr-Jul</th>
<th>Aug-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Full Data Capture</td>
<td>• Full or Partial Complete Cases Confounded with Complexity</td>
<td>• Likely Partial Only Complex Complete Cases</td>
</tr>
</tbody>
</table>

**Strategy 2 (Month by Month Imputation):**

- **Step 1:** Model expected utilization for each gestational month including month of expected care initiation
- **Step 2:** Use model to impute missing values from unobserved months

NOTE: Allows us to use data from Jan-Jul. Aug-Dec becomes less reliable.
Imputing the Missing Months

- The full cases are used to develop a month-to-month model of maternity costs.
- The model is used to impute unobserved months of data.
### Results – Imputation of Missing Months

#### Complete Case Analysis

<table>
<thead>
<tr>
<th>Item</th>
<th>Total Spent</th>
<th>N</th>
<th>Average Costs Per Delivery</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preterm (&lt;20 Wks)</td>
<td>$41,251</td>
<td>4</td>
<td>$10,312</td>
<td>$15,469</td>
</tr>
<tr>
<td>Preterm (20-36 Wks)</td>
<td>$3,256,007</td>
<td>530</td>
<td>$6,143</td>
<td>$6,757</td>
</tr>
<tr>
<td>Term (37+ Wks)</td>
<td>$16,822,345</td>
<td>4575</td>
<td>$3,677</td>
<td>$3,309</td>
</tr>
<tr>
<td>Total</td>
<td>$20,119,603</td>
<td>5109</td>
<td>$3,938</td>
<td>$3,741</td>
</tr>
</tbody>
</table>

#### Month by Month Imputation

<table>
<thead>
<tr>
<th>Item</th>
<th>Total Spent</th>
<th>N</th>
<th>Average Costs Per Delivery</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preterm (&lt;20 Wks)</td>
<td>$41,251</td>
<td>4</td>
<td>$10,312</td>
<td>$15,469</td>
</tr>
<tr>
<td>Preterm (20-36 Wks)</td>
<td>$4,884,010</td>
<td>795</td>
<td>$6,143</td>
<td>$5,529</td>
</tr>
<tr>
<td>Term (37+ Wks)</td>
<td>$37,009,159</td>
<td>10065</td>
<td>$3,677</td>
<td>$2,206</td>
</tr>
<tr>
<td>Total</td>
<td>$41,934,420</td>
<td>10864</td>
<td>$3,859</td>
<td>$2,894</td>
</tr>
</tbody>
</table>
In our Study...

- Family History, PHQ-9, and smoking status have missing values.
  - If patients do not know Family History, 
    - Missing Completely At Random
  - If men are less likely to complete a PHQ-9
    - Missing At Random
  - If Smokers tend not to share smoking status
    - Missing Not at Random

Missingness mechanism(s) are **ignorable**

- Model the missingness mechanism(s)
  - Selection and/or Pattern Mixture models
Illustration: MAR and ED Utilization

- **Task:** Estimate the relationship between BH integration and ED use among Substance Abuse/Mental Health patients
  - **Inclusion Criteria:** Medicaid enrollees with 10 or more months of semi-continuous enrollment
  - **Missing Data Issue:** ED utilization follows known cyclical patterns by month.

**Why MAR?**

- *ED visits in unobserved months are predictable by when they are missing*
Illustration: MAR and ED Episodes

• **Strategy 1 (K-Means/Nearest Neighbors):**
  – Neighborhood 1: Populate missing months with that person’s average from observed months (*Common Approach*)
    • *Issue*: Ignores month to month seasonality
  – Neighborhood 2: Populate missing months with average from age, gender matched peers
    • *Issue*: Lowers variation in outcome (*false positives*)

• **Strategy 2 (Multiple Imputation with Month Covariate/Fixed Effect):**
  – Directly implementable with SAS PROC MI, LCMD (R), etc.
    • *NOTE*: MICE (R) would work but it assumes normality
### Results – K-Means vs. Imputed

#### Strategy 1: K-Means ED Visits

<table>
<thead>
<tr>
<th>Item</th>
<th>Behavioral Health Centrality</th>
<th>2011 Predicted Mean</th>
<th>2012 Predicted Mean</th>
<th>2013 Predicted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits</td>
<td>Jan-Mar</td>
<td>13.5</td>
<td>13.9</td>
<td>14.1</td>
</tr>
<tr>
<td>Visits</td>
<td>Apr-Jun</td>
<td>13.9</td>
<td>14.0</td>
<td>14.1</td>
</tr>
<tr>
<td>Visits</td>
<td>July-Sept</td>
<td>14.1</td>
<td>14.2</td>
<td>14.2</td>
</tr>
<tr>
<td>Visits</td>
<td>Oct-Nov</td>
<td>13.7</td>
<td>13.9</td>
<td>14.0</td>
</tr>
</tbody>
</table>

#### Strategy 2: Imputed (seasonally adjusted) ED Visits

<table>
<thead>
<tr>
<th>Item</th>
<th>Behavioral Health Centrality</th>
<th>2011 Predicted Mean</th>
<th>2012 Predicted Mean</th>
<th>2013 Predicted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits</td>
<td>Jan-Mar</td>
<td>13.2</td>
<td>13.4</td>
<td>13.4</td>
</tr>
<tr>
<td>Visits</td>
<td>Apr-Jun</td>
<td>14.1</td>
<td>14.3</td>
<td>14.8</td>
</tr>
<tr>
<td>Visits</td>
<td>July-Sept</td>
<td>14.3</td>
<td>14.5</td>
<td>15</td>
</tr>
<tr>
<td>Visits</td>
<td>Oct-Nov</td>
<td>13.5</td>
<td>13.9</td>
<td>13.8</td>
</tr>
</tbody>
</table>

Strategy 1 vs 2: The nearest neighbors approach reduces the impact of seasonal patterns
In our Study… (continued)

- Family History, PHQ-9, and smoking status have missing values.
  - If patients do not know Family History, 
    - *Missing Completely At Random*
  - If men are less likely to complete a PHQ-9 
    - *Missing At Random*
  - If Smokers tend not to share smoking status 
    - *Missing Not at Random*

- Missingness mechanism(s) are ignorable

- Model the missingness mechanism(s)
  - Selection and/or Pattern Mixture models
Illustration: MNAR and Maternal Episode

• **Task:** Estimate total amount spent on delivery by gestational age
  
  – **Missing Data Issue:** Emergency transport costs is tracked in a different system and not available for Fee For Service (FFS)

• **Why are the data MNAR?**
  
  – The amount/type of data collected is associated with payer type

  – *Rubin: The pattern of missing data dependent upon the pattern of missingness*
# Billed Amounts (Maternal Episodes)

## Fee For Service (No emergency transport costs)

<table>
<thead>
<tr>
<th>Item</th>
<th>Total Spent</th>
<th>N</th>
<th>Average Costs Per Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preterm (&lt;20 Wks)</td>
<td>$17,715</td>
<td>2</td>
<td>$8,858</td>
</tr>
<tr>
<td>Preterm (20-36 Wks)</td>
<td>$2,477,969</td>
<td>534</td>
<td>$4,640</td>
</tr>
<tr>
<td>Term (37+ Wks)</td>
<td>$15,135,319</td>
<td>4,574</td>
<td>$3,309</td>
</tr>
<tr>
<td>Total</td>
<td>$17,639,173</td>
<td>5,109</td>
<td>$3,938</td>
</tr>
</tbody>
</table>

## Managed Care Population

<table>
<thead>
<tr>
<th>Item</th>
<th>Total Spent</th>
<th>N</th>
<th>Average Costs Per Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preterm (&lt;20 Wks)</td>
<td>$1,965</td>
<td>1</td>
<td>$1,965</td>
</tr>
<tr>
<td>Preterm (20-36 Wks)</td>
<td>$22,770,592</td>
<td>1,179</td>
<td>$19,313</td>
</tr>
<tr>
<td>Term (37+ Wks)</td>
<td>$111,838,127</td>
<td>11,020</td>
<td>$4,688</td>
</tr>
<tr>
<td>Total</td>
<td>$41,934,420</td>
<td>10,864</td>
<td>$3,859</td>
</tr>
</tbody>
</table>

A comparison of Fee-for Service and Managed Care Organization episode costs:
- Similar overall average episode costs
- Contracted MCOs appear to managed complex cases poorly
- A Heckman adjustment or pattern-mixture model is needed
Summary of Intro to Missing Data

- Why do we care about Missing Data?
  - Impact analytic precision (MCAR)
  - Effect analysis and interpretation (MAR)
  - Introduce potential bias (MNAR)
- How does it appear in datasets?
- What can be done about it?
  - *Ignore/Complete Case: MCAR*
  - *Impute*
  - *Model Directly*
Filling in the Gaps in Maternity Data

Chris McInnish, Director of Quality Analytics, and

Drew Nelson, MPH, Director of Quality Assurance Division

Alabama Medicaid
Problem

• Deliveries identified from claims data and vital statistics
• Not all deliveries have vital stats match
  • Out of state
  • Name differences
  • Source of payment issues
• Claims data inconsistent for quality indicators
Alabama Approach

Multiple Sources of Semi-Truth. Looking for at least two to agree.

- Build a hierarchy of data sources for maternity indicators
Identify Deliveries

• Claims
  • Maternity case management providers
  • Hospitals
  • Physician
• Deduplication
• Validate
• Determine where possible mother/baby match
Validation

1. Vital Stats Match
2. Out of State Deliveries
3. Two Independent Sources
4. Exclude for multiple deliveries in 6 months
5. Exclude for age <9 and >60
Delivery Date

- Vital Stats
- Case Management Claim
- Physician Claim
- Hospital Date of Admission
Gestational Age

- Sources of truth in order
  - Vital Stats
  - Claims from hospital/physician
  - Case Management System
Number of Prenatal Visits

• Maximum number from any source
  • Vital Stats
  • Claims from physician
  • Case Management System
Birth Weight

• Sources of truth in order (with reasonableness bounds)
  • Vital Stats
  • Claims from Physician / Hospital
  • Case Management System
Post-Partum

• Sources of Truth in Order
  • Claims from physician
  • Case Management System
Newborn Information

• Medicaid covers many newborns where the pregnancy and delivery were not paid for by Medicaid
• Claims for deliveries do not include infants ID
• Vital Stats and eligibility information used to match
• Vital stats data received for all infants with eligibility and all deliveries
Product

- Maternity file with all deliveries
  - One line per baby delivered

- Infant File
  - One line per infant with eligibility within first 3 months of life
Examples
Medicaid Births by Year
Prenatal Care

Number of Prenatal Visits (2016)

Percentage of Mothers receiving Recommended Prenatal Care (2016)

U.S. Department of Health and Human Services recommends a total of 12 prenatal visits.
Prenatal and Post Partum Care

U.S. Department of Health and Human Services recommends the first prenatal visit at 8 weeks.

Post Partum Visit Rates (2016)

- All Deliveries:
  - Yes: 66%
  - No: 34%

- Contractor (79% of deliveries):
  - Yes: 77%
  - No Contractor: 23%

- No Contractor:
  - Yes: 77%
Eligibility

Mother’s Eligibility Status (2016)

Eligibility at Delivery

- SOBRA Women: 47%
- Full Medicaid: 43%
- Other: 10%
- Plan First: 0%
- No Coverage: 0%

Eligibility 4 Months After Delivery

- 3%
- 4%
- 33%
- 20%
Delivery Costs

**Average Delivery Cost by District (2016)**

- District 1: $8,172
- District 2: $8,094
- District 3: $8,894
- District 4: $9,257
- District 5: $10,215
- District 6: $9,402
- District 7: $9,323
- District 8: $9,033
- District 9: $9,649
- District 10: $9,295
- District 11: $8,219
- District 12: $9,607
- District 13: $10,573
- District 14: $10,350

**Average Delivery Cost by County (2016)**

- Mother's Average Delivery Cost Range: $4,445 to $12,251
Infant Costs

Medicaid NICU Information (2016)

- 16% Requires NICU
- 19 Days Average NICU Stay
- $2069 Average Cost/Day

<table>
<thead>
<tr>
<th>Item</th>
<th>NICU</th>
<th>No NICU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Infants</td>
<td>6,310</td>
<td>32,427</td>
</tr>
<tr>
<td>1st Year Costs</td>
<td>$325,962,477</td>
<td>$86,891,048</td>
</tr>
<tr>
<td>Cost Per Infant</td>
<td>$51,658</td>
<td>$2,680</td>
</tr>
</tbody>
</table>
Infants – Neonatal Abstinence Syndrome

![Infants Diagnosed with NAS Mothers' Medicaid Opioids Claims Status](image1)

![NAS Infants by County (2016)](image2)
Questions?
Takeaways

• Ignoring (deleting) missing data may lead to incorrect conclusions

• Strategy to address missing data is determined by its pattern and structure

• Minimizing the presence of missing data is the best solution, but there are also analytical approaches that we will highlight in our next webinar which will be in September 2018
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.go