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

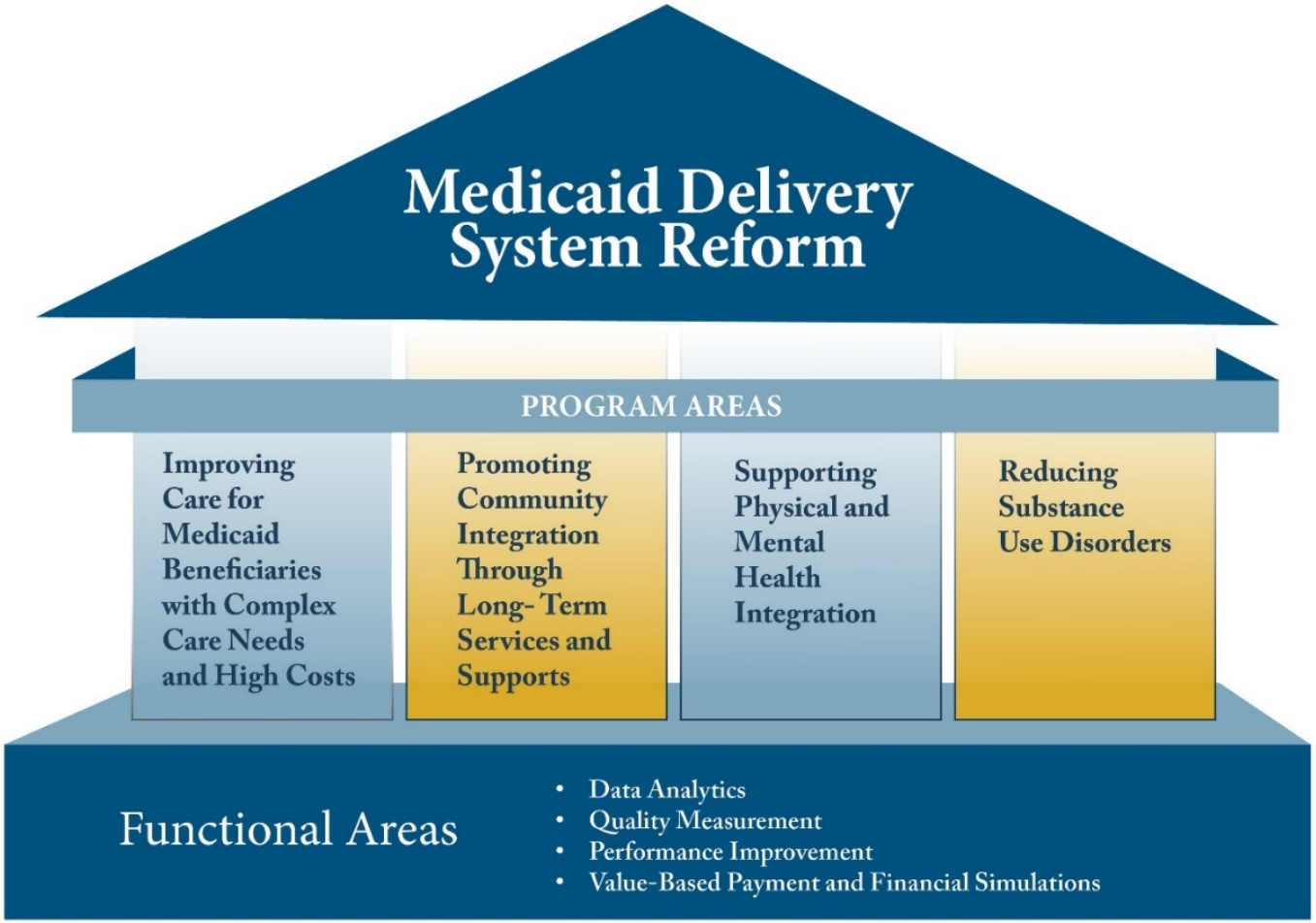
Agenda for Today's Webinar

- 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)



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

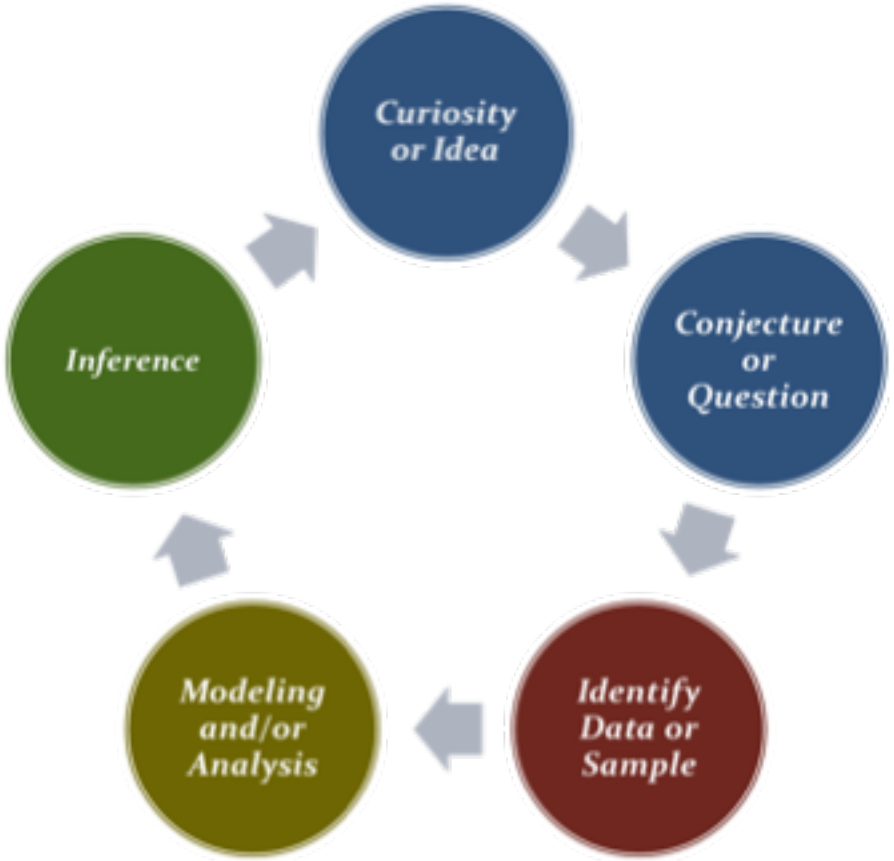
Thomas Flottemesch, Senior Research Leader, IBM Watson Health

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

Group	Var 1	Var 2	Var 3
A	?	?	?
B		?	?
C			?
D			

- 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

Pattern	National Academy (US)	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*

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)

Costs ($)	Sex	Cancer	DM	Asthma	MI	PHQ-9	Smoking Status
$$$	M	?				?	
$$$	F						
$$$	M			?			
$$$	F		?		?		?
$$$	M						
$$$	F			?			
$$$	M		?		?		?
$$$	F	?					
$$$	M			?		?	
$$$	F						

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.

Jan-Mar	Apr-Jul	Aug-Dec
• Full Data Capture • *Complete Cases*	• Full or Partial • *Complete Cases Confounded with Complexity*	• Likely Partial • *Only Complex Complete Cases*

Strategy 2 (Month by Month Imputation):

- *Step1: 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

Full Cases (Gestational Month)
1
2
3
4
5
6
7
8?
9?
10?

Partial Cases (Gestational Month)
1
2
3
4
5
6?
7?
8?
9?
10?

- 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

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

Month by Month Imputation

Item	Total Spent	N	Average Costs Per Delivery	Standard Deviation
Preterm (<20 Wks)	$41,251	*4*	$10,312	*$15,469*
Preterm (20-36 Wks)	$4,884,010	*795*	$6,143	*$5,529*
Term (37+ Wks)	$37,009,159	*10065*	$3,677	*$2,206*
Total	**$41,934,420**	***10864***	**$3,859**	***$2,894***

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*

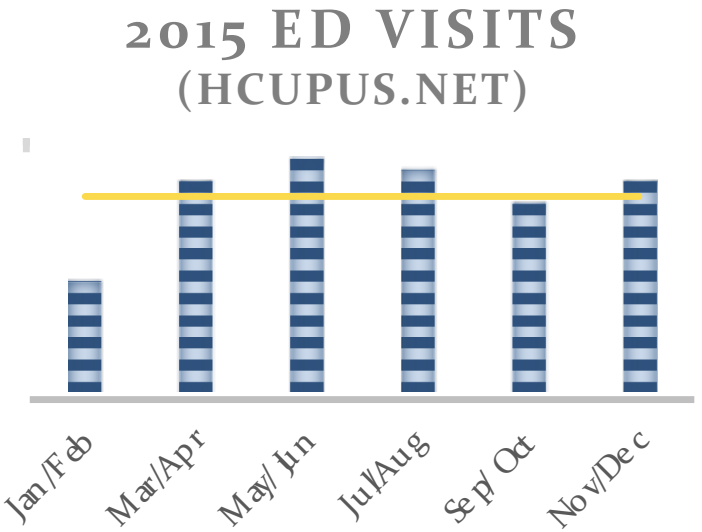


Illustration: MAR and ED Episodes

- **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

Item	Behavioral Health Centrality	2011 Predicted Mean	2012 Predicted Mean	2013 Predicted Mean
Visits	Jan-Mar	13.5	13.9	14.1
Visits	Apr-Jun	13.9	14.0	14.1
Visits	July-Sept	14.1	14.2	14.2
Visits	Oct-Nov	13.7	13.9	14.0

Strategy 2: Imputed (seasonally adjusted) ED Visits

Item	Behavioral Health Centrality	2011 Predicted Mean	2012 Predicted Mean	2013 Predicted Mean
Visits	Jan-Mar	*13.2*	*13.4*	*13.4*
Visits	Apr-Jun	**14.1**	**14.3**	**14.8**
Visits	July-Sept	**14.3**	**14.5**	**15**
Visits	Oct-Nov	*13.5*	*13.9*	*13.8*

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)

Item	Total Spent	N	Average Costs Per Delivery
Preterm (<20 Wks)	$17,715	*2*	$8,858
Preterm (20-36 Wks)	*$2,477,969*	*534*	*$4,640*
Term (37+ Wks)	$15,135,319	*4,574*	$3,309
Total	**$17,639,173**	***5,109***	**$3,938**

Managed Care Population

Item	Total Spent	N	Average Costs Per Delivery
Preterm (<20 Wks)	$1,965	*1*	$1,965
Preterm (20-36 Wks)	*$22,770,592*	*1,179*	*$19,313*
Term (37+ Wks)	$111,838,127	*11,020*	$4,688
Total	**$41,934,420**	***10,864***	**$3,859**

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

ALABAMA MEDICAID EFFICIENCY INNOVATION QUALITY

Number of Prenatal Visits

- Maximum number from any source
  - Vital Stats
  - Claims from physician
  - Case Management System

ALABAMA MEDICAID EFFICIENCY INNOVATION QUALITY

Birth Weight

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

ALABAMA MEDICAID EFFICIENCY INNOVATION QUALITY

Post-Partum

- Sources of Truth in Order
  - Claims from physician
  - Case Management System

ALABAMA MEDICAID EFFICIENCY INNOVATION QUALITY

Newborn Information

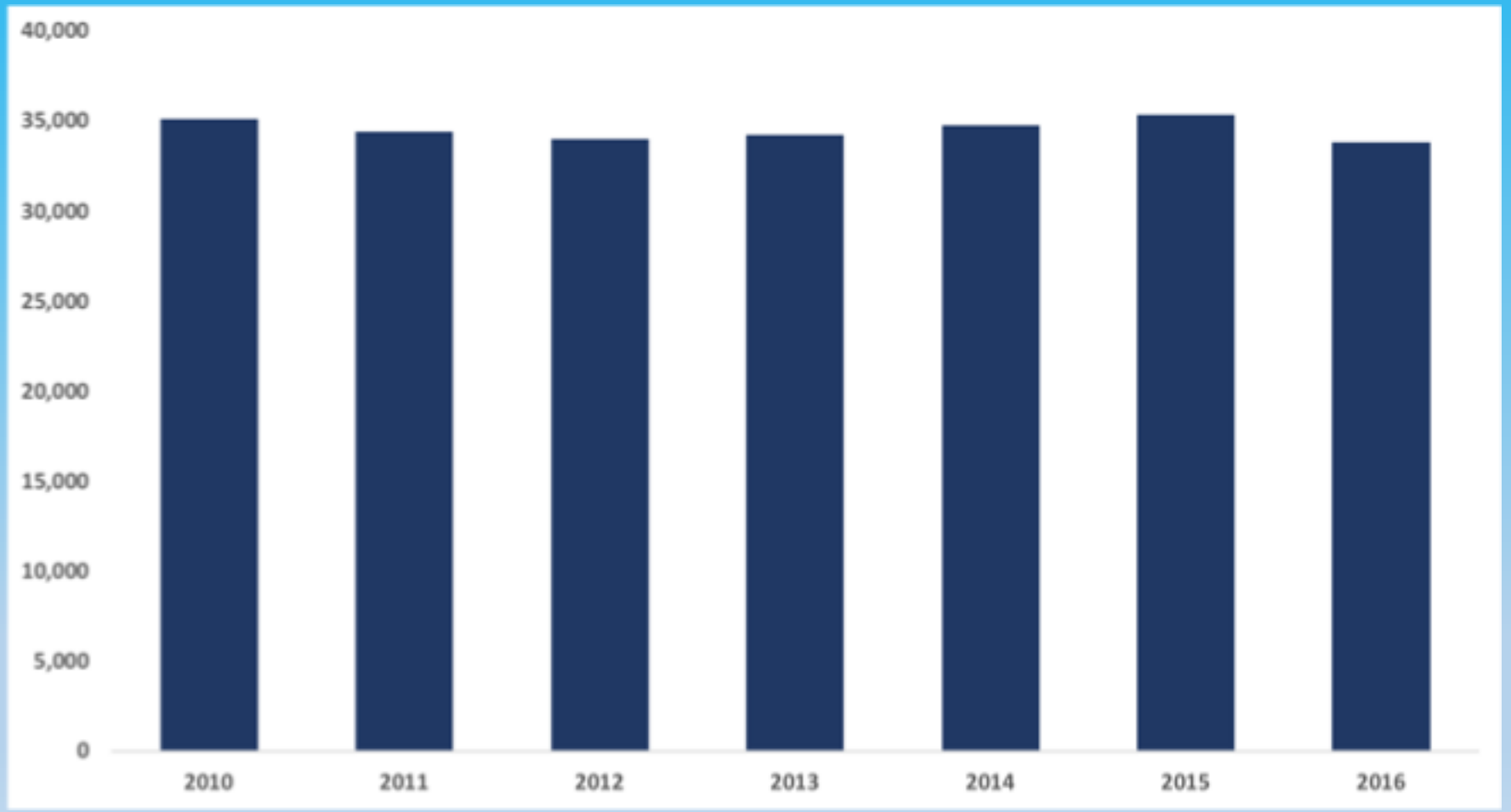
- Medicaid covers many newborns where the pregnancy and delivery were not paid for by Medicaid
- Claims for deliveries do not include infants ID
- Vital Stats and eligibility information used to match
- Vital stats data received for all infants with eligibility and all deliveries

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

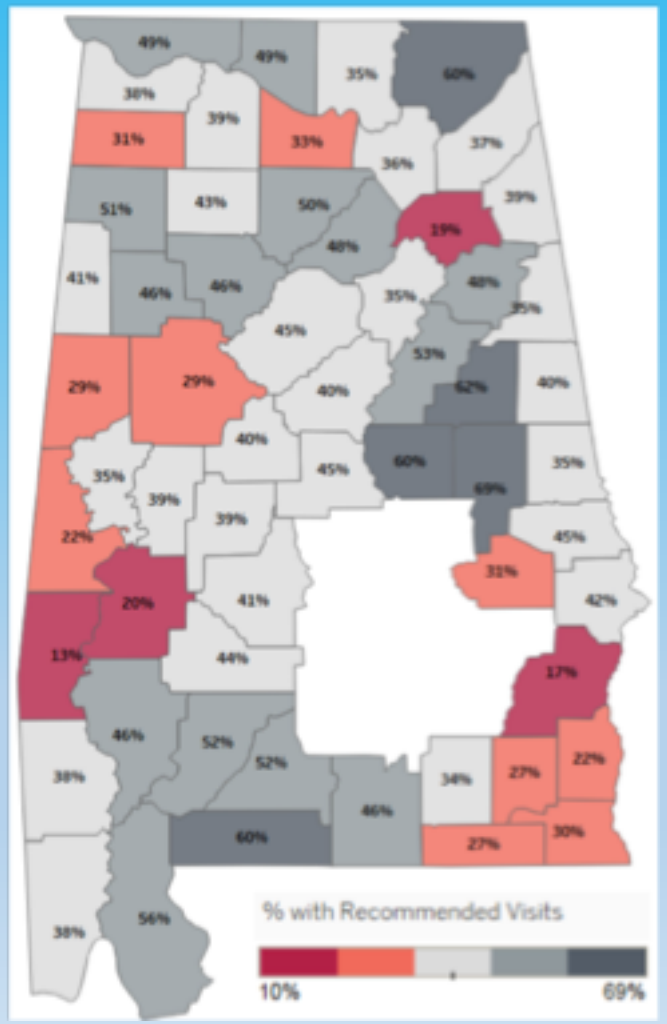
Examples

Medicaid Births by Year

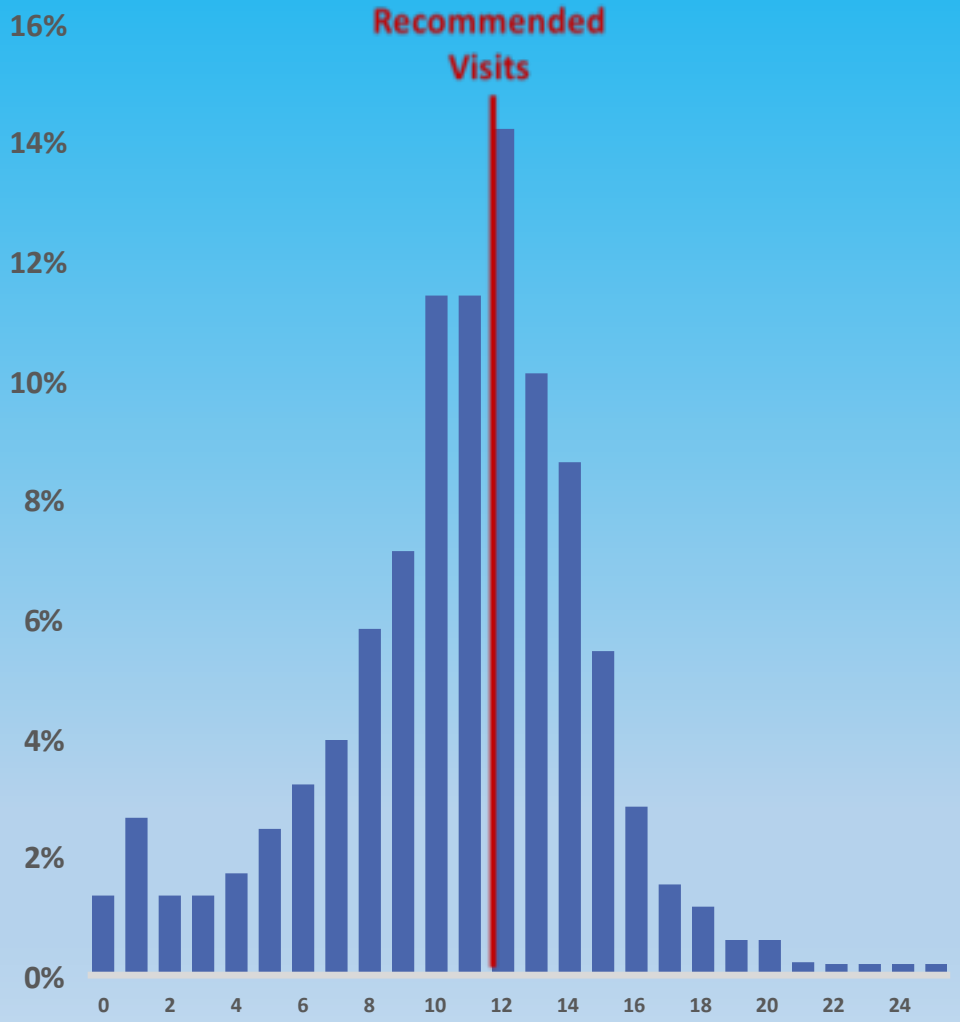


Prenatal Care

Percentage of Mothers receiving Recommended Prenatal Care (2016)

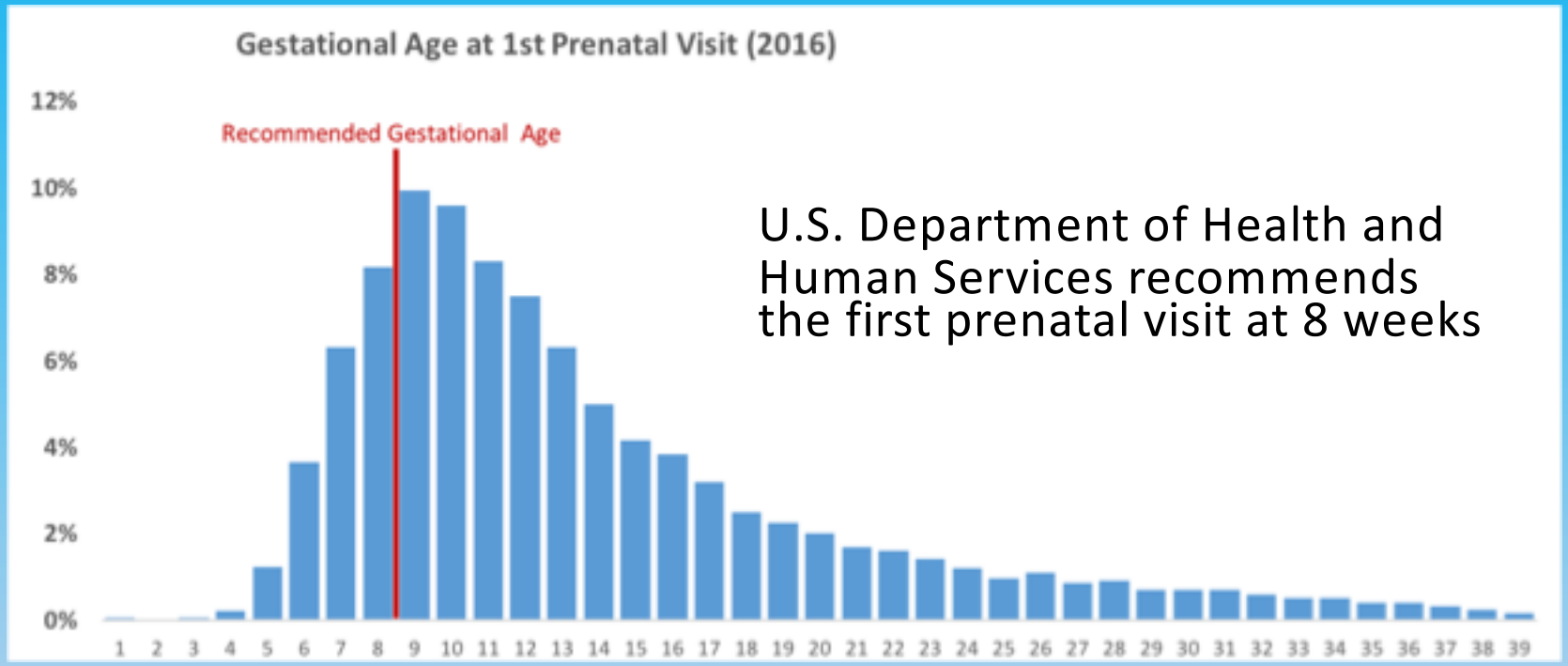


Number of Prenatal Visits (2016)



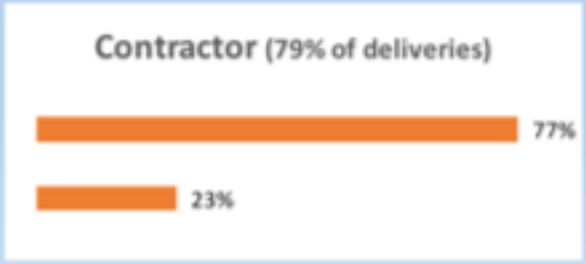
U.S. Department of Health and Human Services recommends a total of 12 prenatal visits

Prenatal and Post Partum Care



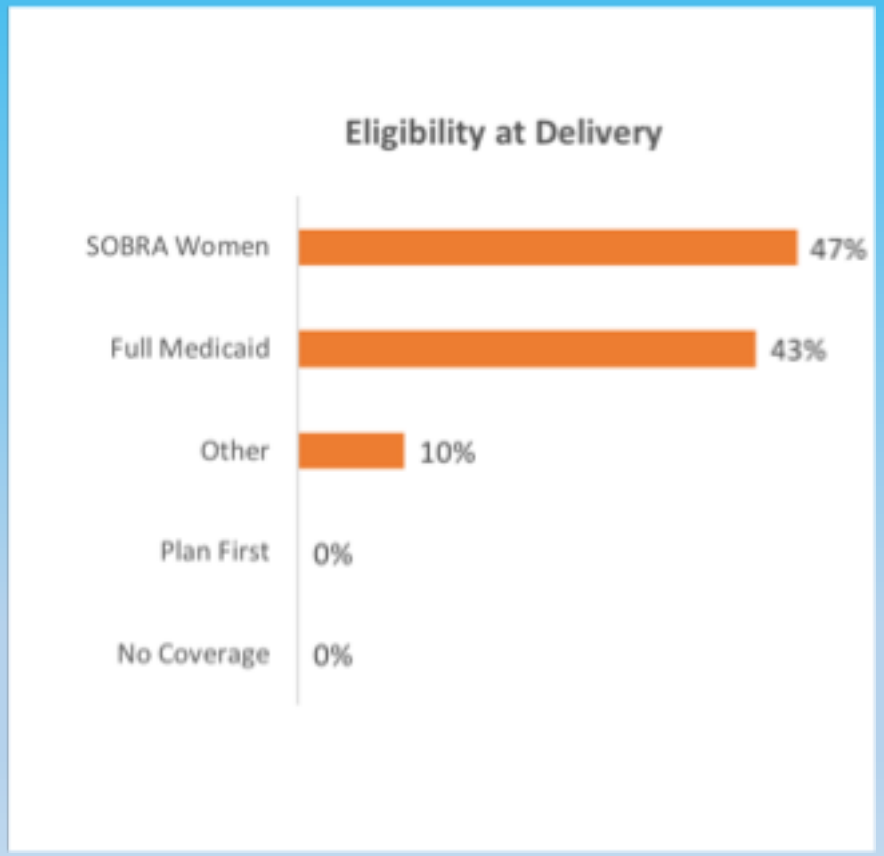
U.S. Department of Health and Human Services recommends the first prenatal visit at 8 weeks

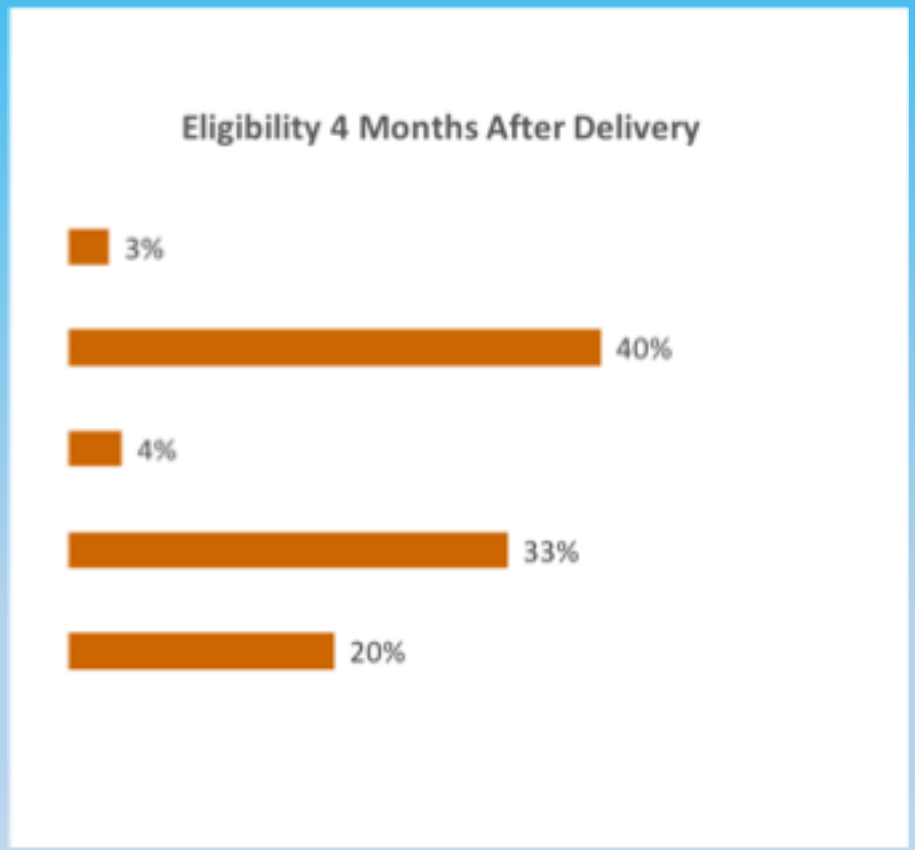
Post Partum Visit Rates (2016



Eligibility

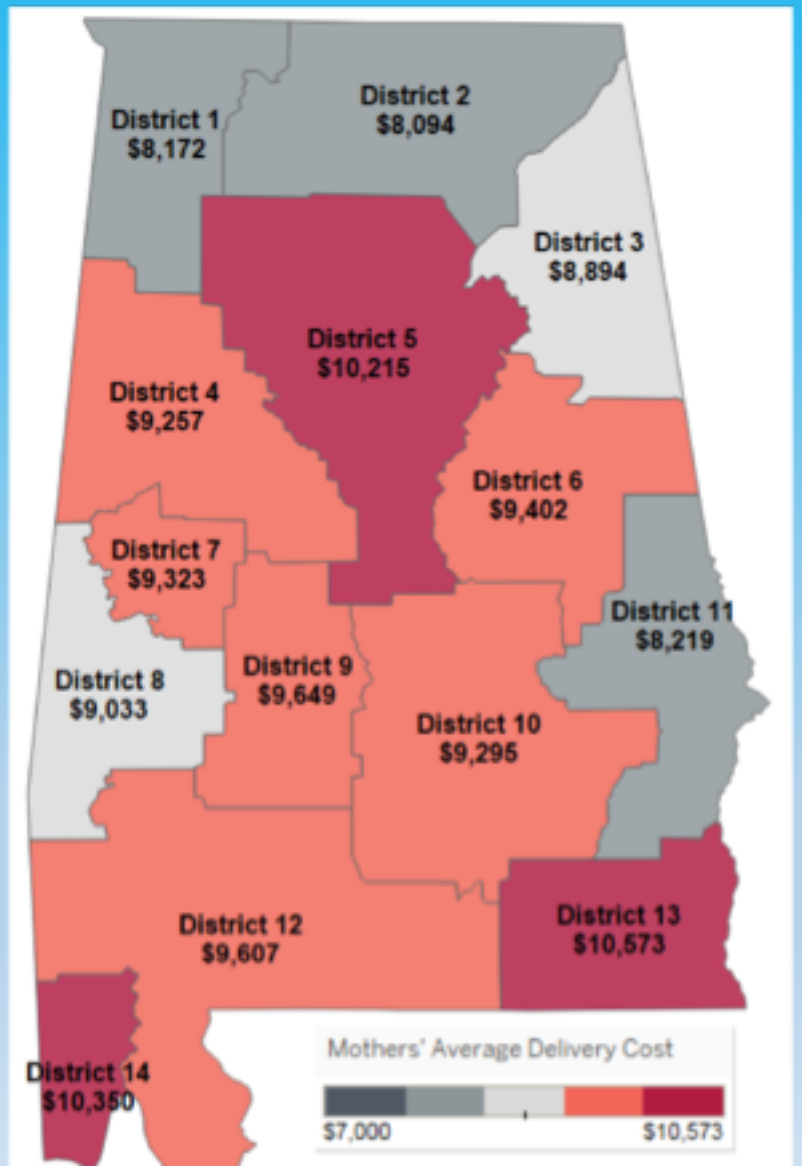
Mother's Eligibility Status (2016)



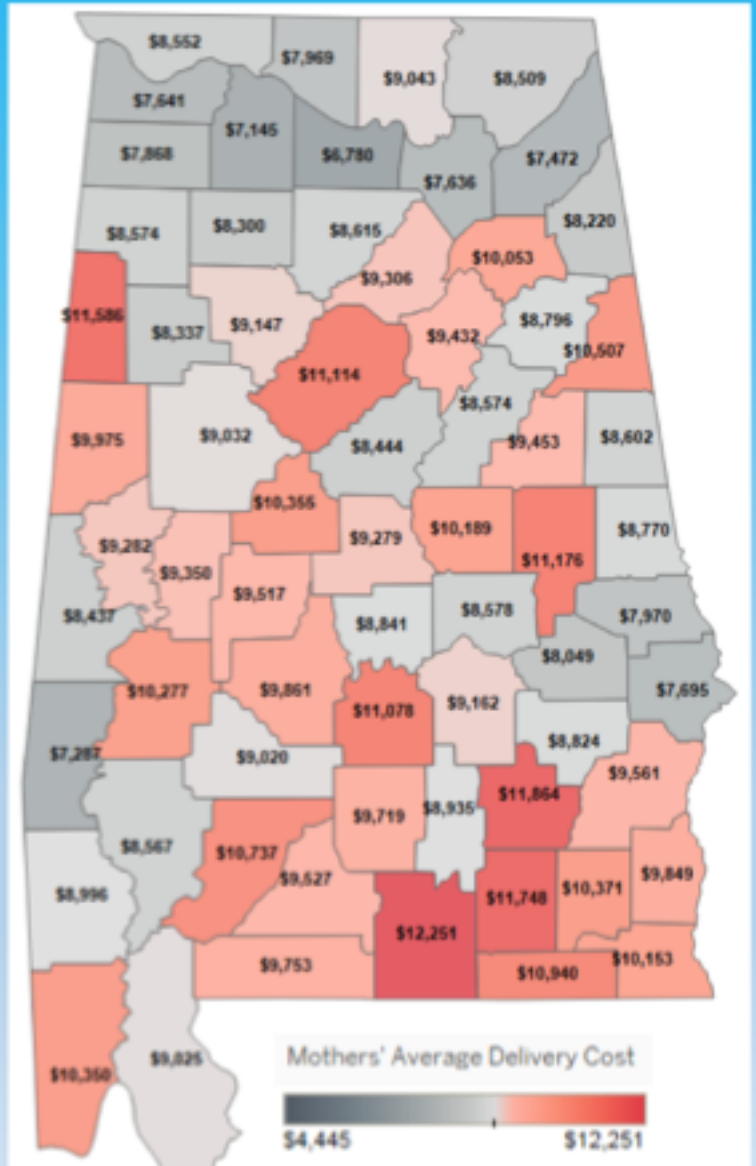


Delivery Costs

Average Delivery Cost by District (2016)



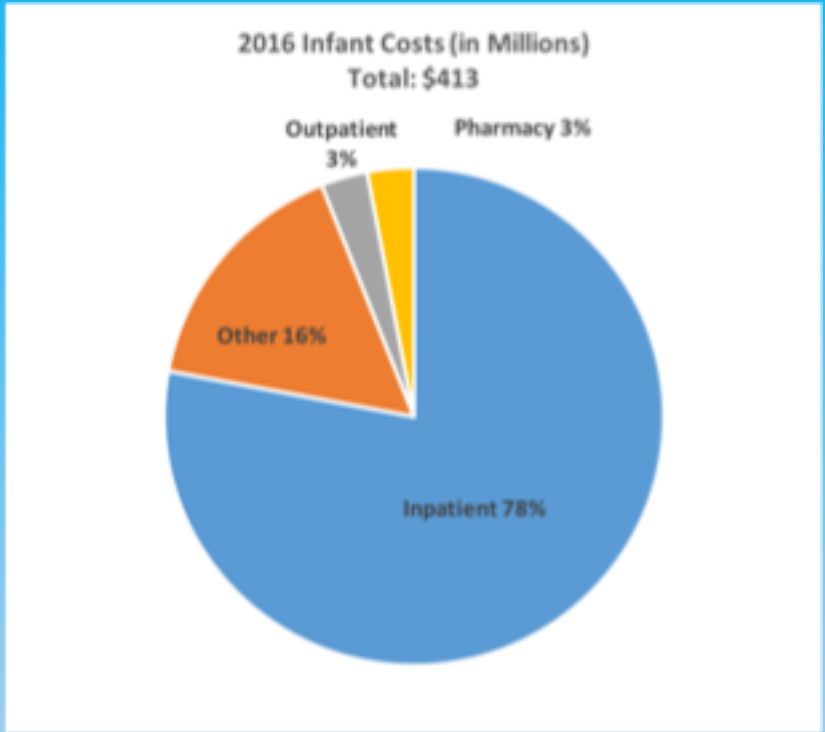
Average Delivery Cost by County (2016)

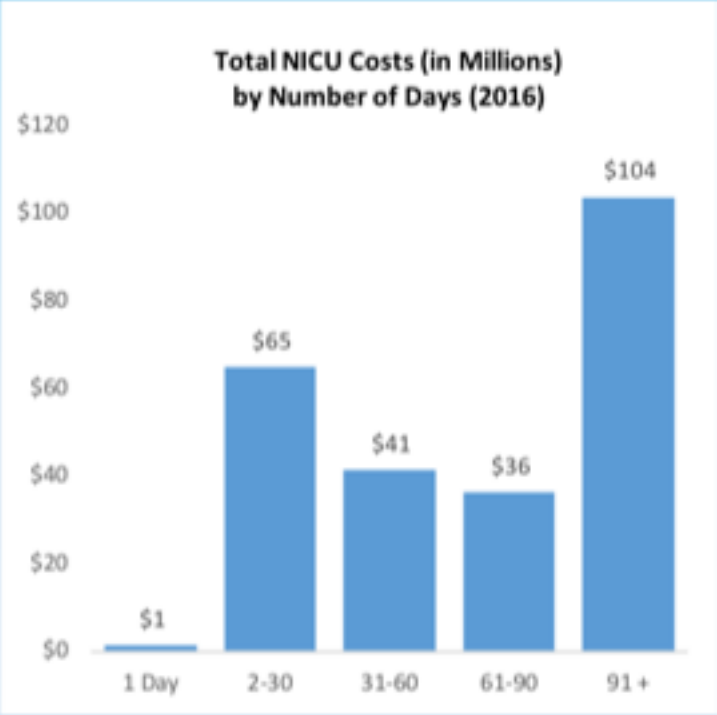


Infant Costs

Medicaid NICU Information (2016)

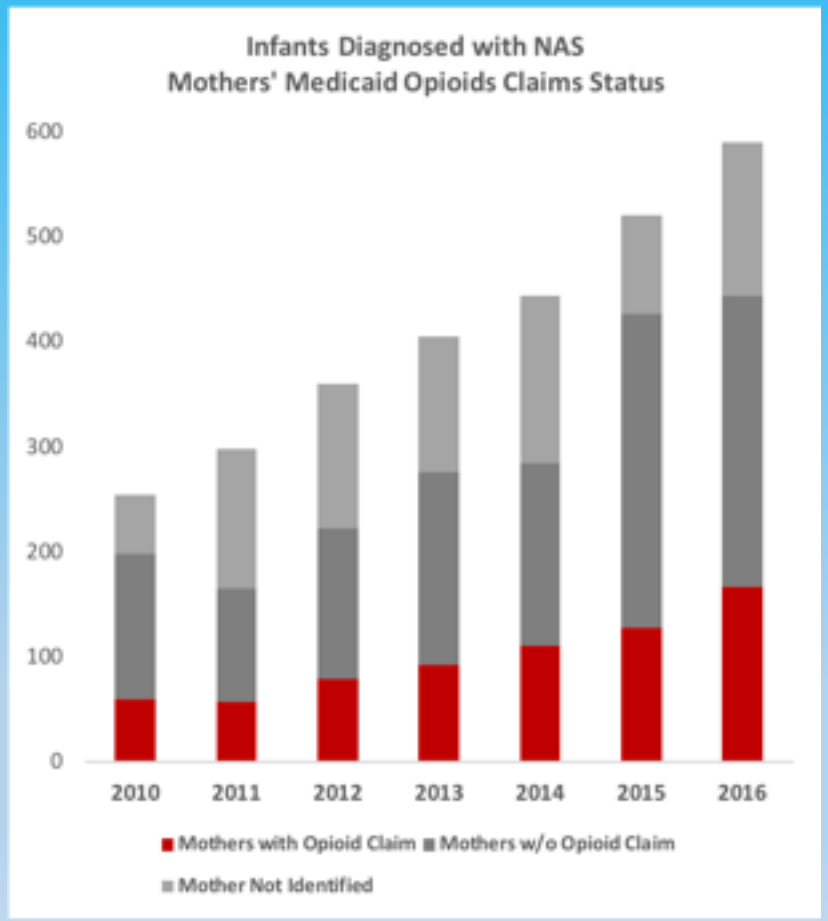
- 16% Requires NICU
- 19 Days Average NICU Stay
- $2069 Average Cost/Day



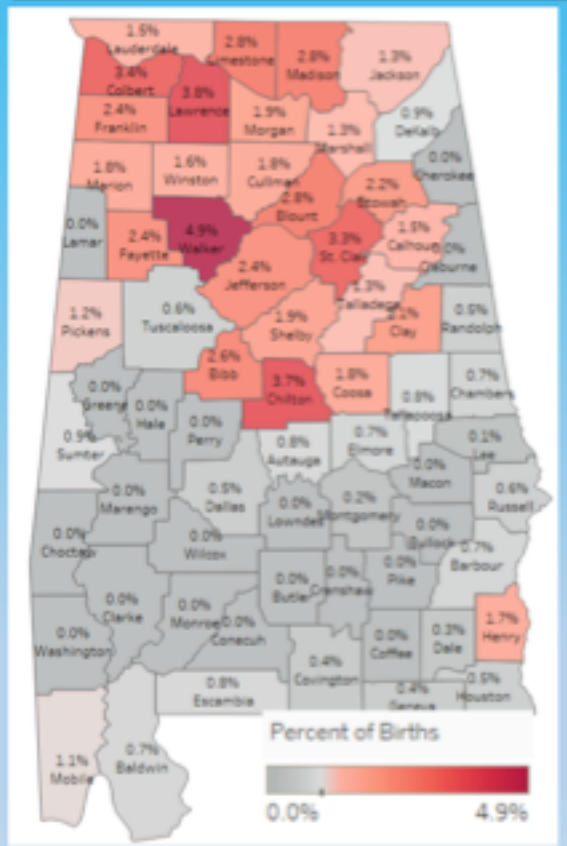


Item	NICU	No NICU
Number of Infants	6,310	32,427
1st Year Costs	$325,962,477	$86,891,048
Cost Per Infant	$51,658	$2,680

Infants – Neonatal Abstinence Syndrome



NAS Infants by County (2016)



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