

**WELCOME
TO THE 2026
NDACAN
SUMMER
TRAINING
SERIES!**

The session will begin at 12pm EST.

This session is being recorded.

Please submit questions to the Q&A box!

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Duke University, Cornell University, UC San Francisco, & Mathematica



Children's Bureau

An Office of the Administration for Children & Families



LAYING THE GROUNDWORK

*FOUNDATIONAL SKILLS FOR USING ADMINISTRATIVE
DATA IN CHILD WELFARE RESEARCH*

NDACAN SUMMER TRAINING SERIES SCHEDULE

- **July 1st, 2026**
 - Overview of NDACAN administrative datasets
- **July 8th, 2026**
 - Data cleaning and management
- **July 15th, 2026**
 - Linking NCANDS and AFCARS
- **July 22rd, 2026**
 - Handling missing data
- **July 29th, 2026**
 - Data presentation and visualization

SESSION AGENDA

- Handling missing data
- Demonstration in R
- Q & A

HANDLING MISSING DATA

TYPES OF MISSING ADMINISTRATIVE DATA

- Record-level missingness
 - I.e. observations that should appear in your dataset don't
- Item-level missingness
 - I.e. values that should appear in your dataset don't

RECORD-LEVEL MISSINGNESS IN NDACAN ADMINISTRATIVE DATA

- Some NDACAN datasets have meaningful amounts of missing observations.
 - E.g. 100% of maltreatment reports in NCANDS are missing in Arizona in 2021.
 - Not all record-level missingness is accounted for in User Guides.

SOURCES OF RECORD-LEVEL MISSINGNESS

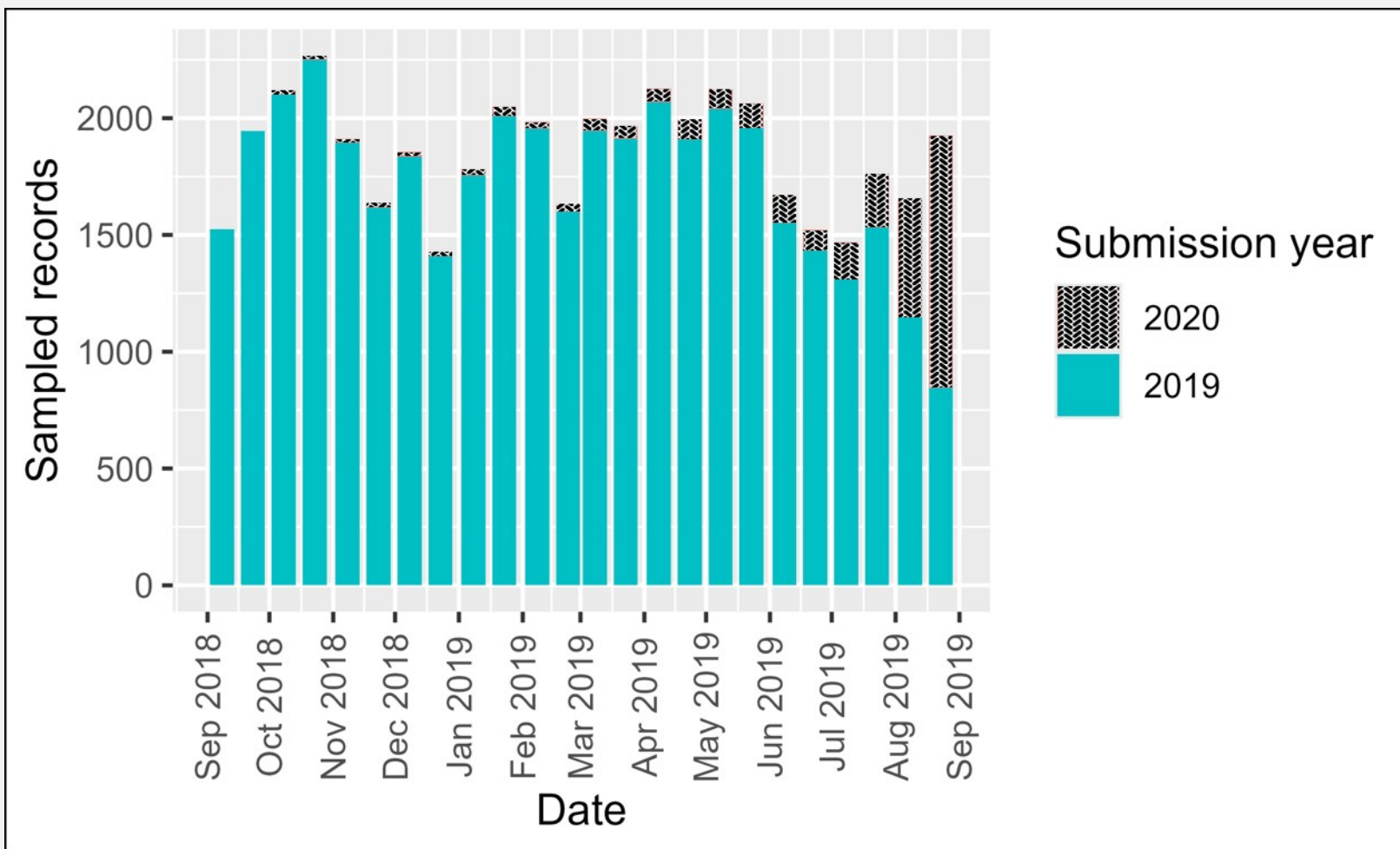
- State-level non-reporting
- Delayed reporting
- Data suppression (e.g. tribal agencies)

STATE-LEVEL NONREPORTING

Data Year (Fiscal)	Version	Latest Release	Records	Variables	States Reporting	States Not Reporting	Missing States
2007	6	5/4/2023	3,323,128	143	49	3	MI, ND, OR
2008	6	5/4/2023	3,624,032	143	50	2	ND, OR
2009	7	5/10/2023	3,582,703	143	50	2	ND, OR
2010	6	5/10/2023	3,556,648	143	51	1	OR
2011	6	5/10/2023	3,655,951	143	51	1	OR
2012	6	5/10/2023	3,846,933	144	52	0	
2013	6	11/15/2023	3,863,014	147	52	0	
2014	5	11/15/2023	3,958,493	147	52	0	
2015	5	11/8/2023	4,063,137	147	52	0	
2016	4	11/15/2023	4,186,257	147	51	1	PR
2017	4	11/15/2023	4,279,060	149	52	0	
2018	5	9/7/2023	4,333,564	152	52	0	
2019	5	8/31/2023	4,256,572	152	52	0	
2020	3	7/23/2023	3,807,380	153	52	0	
2021	2	7/28/2023	3,592,284	153	51	1	AZ
2022	1	1/30/2024	3,732,871	153	52	0	

Source: NDACAN internal documentation for NCANDS Child Files 2007–2022.

DELAYED REPORTING



Note: Results based on 1% random samples of the 2019 and 2020 NCANDS Child Files.

ITEM-LEVEL MISSINGNESS IN NDACAN ADMINISTRATIVE DATA

- Most NDACAN datasets have meaningful amounts of missing values for variables of common interest.
 - E.g. 63% of maltreatment reports in NCANDS in 2019 have missing values for “Financial Problems.”
 - Not all missing values follow the formatting described in Codebooks.

SOURCES OF ITEM-LEVEL MISSINGNESS

- State- and agency-level processes
- Caseworker-, caregiver-, and child-level processes
- Data suppression (e.g. geographic identifiers)
- Linkage failure

IDENTIFYING MISSING DATA

- NDACAN staff can provide, upon request, internal documentation of record-level missingness.
- But the bread-and-butter techniques for identifying missing data are data cleaning and exploratory analysis

DATA CLEANING FOR IDENTIFYING MISSING DATA

- R may not understand correctly coded missing values as missing values.
- Missing values may not be correctly or consistently coded.
- Identifying all possible encodings of missing values is essential.
 - E.g. “”, “99,” “-1,” “Don’t know,” or “Decline to respond.”
- Encode missing values so that R understands them as such (*NA*).
 - If retaining available information about the missing-data mechanism is important, consider “tagging” *NA* values using packages like *haven*.

EXPLORATORY ANALYSIS FOR IDENTIFYING MISSING DATA

- Tabulate or visualize the frequency distribution of all variables.
 - Stratify by geographic unit of analysis.
- When possible, visualize trends and trend breaks in all variables.

HANDLING RECORD-LEVEL MISSINGNESS

- Limiting your sampling frame so that records are missing in a consistent way can be a helpful way of avoiding threats to internal validity.
- If properly accounted for, missing observations are therefore primarily a threat to external validity.
- Missing-data strategies can be applied to record-level missingness, but are not as common.
- Therefore, transparency about sampling and inference is often most important.

HANDLING ITEM-LEVEL MISSINGNESS

- A wide variety of approaches to handling missing values are feasible with NDACAN administrative data:
 - Complete-case analysis
 - Hot-deck imputation
 - Multiple imputation
 - Full-information maximum-likelihood
 - Bayesian methods
 - Additional data collection

THE STAKES OF HANDLING MISSING DATA APPROPRIATELY

- Without an appropriate strategy, inferences from data with missing values and/or missing observations may be biased and/or inefficient.
- The appropriate missing data strategy depends on both the missing-data mechanism and your analytic goals.

MISSING-DATA MECHANISMS

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

MISSING COMPLETELY AT RANDOM (MCAR)

- The missing-data mechanism is unrelated to the data-generating process.
 - In other words: missingness is not predicted by observed or missing values.
- E.g. when they haven't had enough coffee, caseworkers sometimes forget to record information about children's emotional disturbance.
- Candidate solutions: complete-case analysis, hot-deck imputation, multiple imputation, full-information maximum-likelihood, Bayesian methods, additional data collection

MISSING AT RANDOM (MAR)

- The missing-data mechanism is related to the data-generating process in observable ways.
 - In other words: missingness is predicted only by observed values.
- E.g. states vary in their procedures for recording children's emotional disturbance.
- Candidate solutions: ~~complete-case analysis~~, hot-deck imputation, multiple imputation, full-information maximum-likelihood, Bayesian methods, additional data collection

MISSING NOT AT RANDOM (MNAR)

- The missing-data mechanism is related to the data-generating process in unobservable ways.
 - In other words: missingness is predicted by the missing values themselves.
- E.g. caseworkers are less likely to record information about children's emotional disturbance when they do not experience emotional disturbance.
- Candidate solutions: ~~complete-case analysis, hot-deck imputation, multiple imputation, full-information maximum-likelihood, Bayesian methods~~, additional data collection

PATTERNS OF ITEM-LEVEL MISSINGNESS

State	CdAlc	CdDrug	CdEmotnl	CdVisual	CdLearn	CdPhys	CdBehav	CdMedicl
Alabama	100	100	100	100	100	100	100	100
Arkansas	12	14	11	11	11	11	11	11
California	5	6	3	1	0	0	0	8
Colorado	100	100	100	100	100	100	100	100
Connecticut	79	79	79	79	79	79	79	79
Delaware	86	86	86	86	86	86	86	86
District of Columbia	0	2	0	0	0	0	0	0
Florida	13	13	1	1	1	1	100	1
Georgia	0	7	1	0	0	0	2	1
Hawaii	100	100	31	31	0	31	1	31
Idaho	0	0	0	0	0	0	0	0
Illinois	100	100	100	100	100	100	0	100

Note: Numbers are percentages of records with non-missing values in the 2021 NCANDS Child File.

YOUR ANALYTIC GOALS

- Is your estimand a population parameter such as a count or an incidence rate?
- Is your estimand a model parameter such as a causal effect?
- Is your estimand a prediction or classification?

BEST PRACTICES FOR HANDLING MISSING DATA

- Define an estimand clearly.
- Use available data to make inferences about missing-data mechanisms.
- Choose an appropriate, primary missing-data strategy.
- Conduct sensitivity analysis using reasonable alternative strategies.
- Report information about data missingness, key assumptions, diagnostics, and sensitivity analysis.

DEMONSTRATION IN R

ADDITIONAL RESOURCES

- *Flexible Imputation of Missing Data* (2e), Stef van Buuren
- *The Missing Book*, Nicholas Tierney and Allison Horst

QUESTIONS?

USER SUPPORT

NDACANSUPPORT@CORNELL.EDU

ALEX ROEHRKASSE

AROHRKASSE@BUTLER.EDU

ALYSSA LINDSEY

ALYSSA.LINDSEY@UCSF.EDU

NEXT WEEK...

Date: July 29th

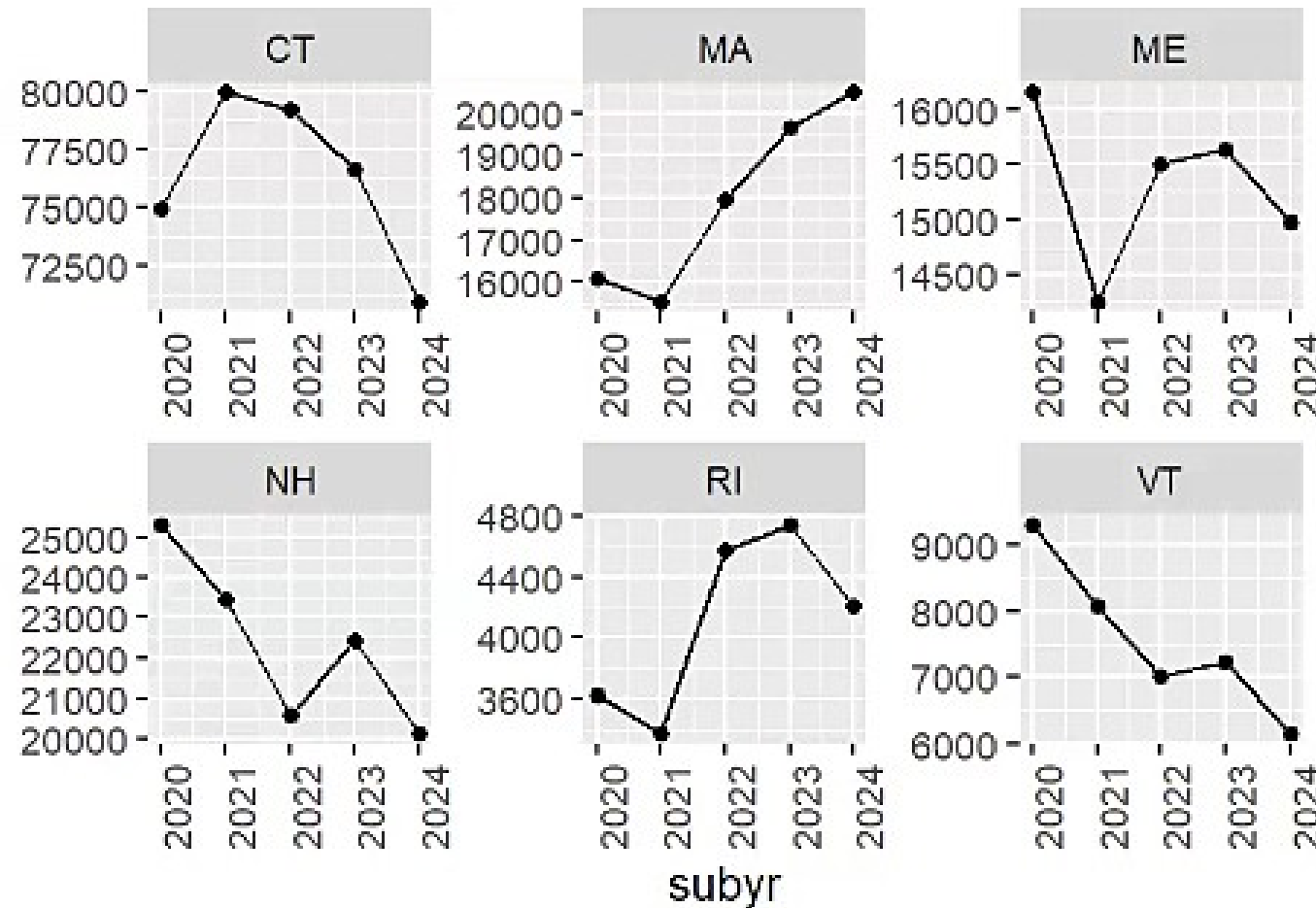
Topic: Data presentation and visualization

Instructor: Alex Roehrkasse

COMMENTS IN THE ZOOM CHAT

- Email the NDACAN archive at NDACAN@cornell.edu with general questions, and email NDACANsupport@cornell.edu with technical support questions.
- The NDACAN archive website <https://www.ndacan.acf.hhs.gov> contains the following:
 - The R code and output of this presentation.
 - Information and instructions for joining the Child-Maltreatment-Research-L (CMRL) list serve.
 - The FAQ Frequently Asked Questions about using the NDACAN archive data including AFCARS, NCANDS, and NYTD.
 - Instructions for ordering restricted datasets

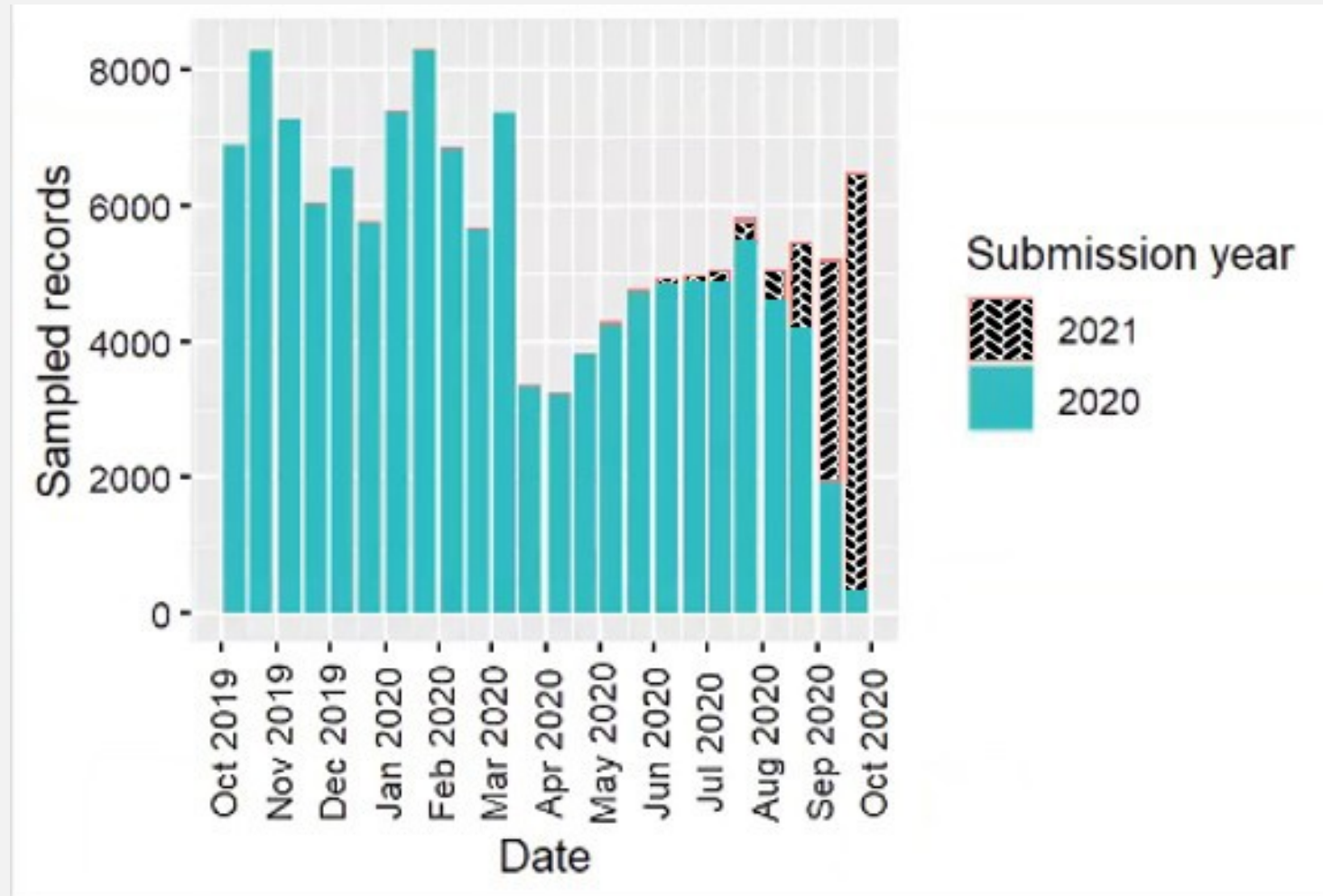
R CODE IMAGE I



Exploring record-level missingness. First, inspect plots of records counts by reporting unit (usually the state) and reporting period for trend breaks which indicate record-level missingness.

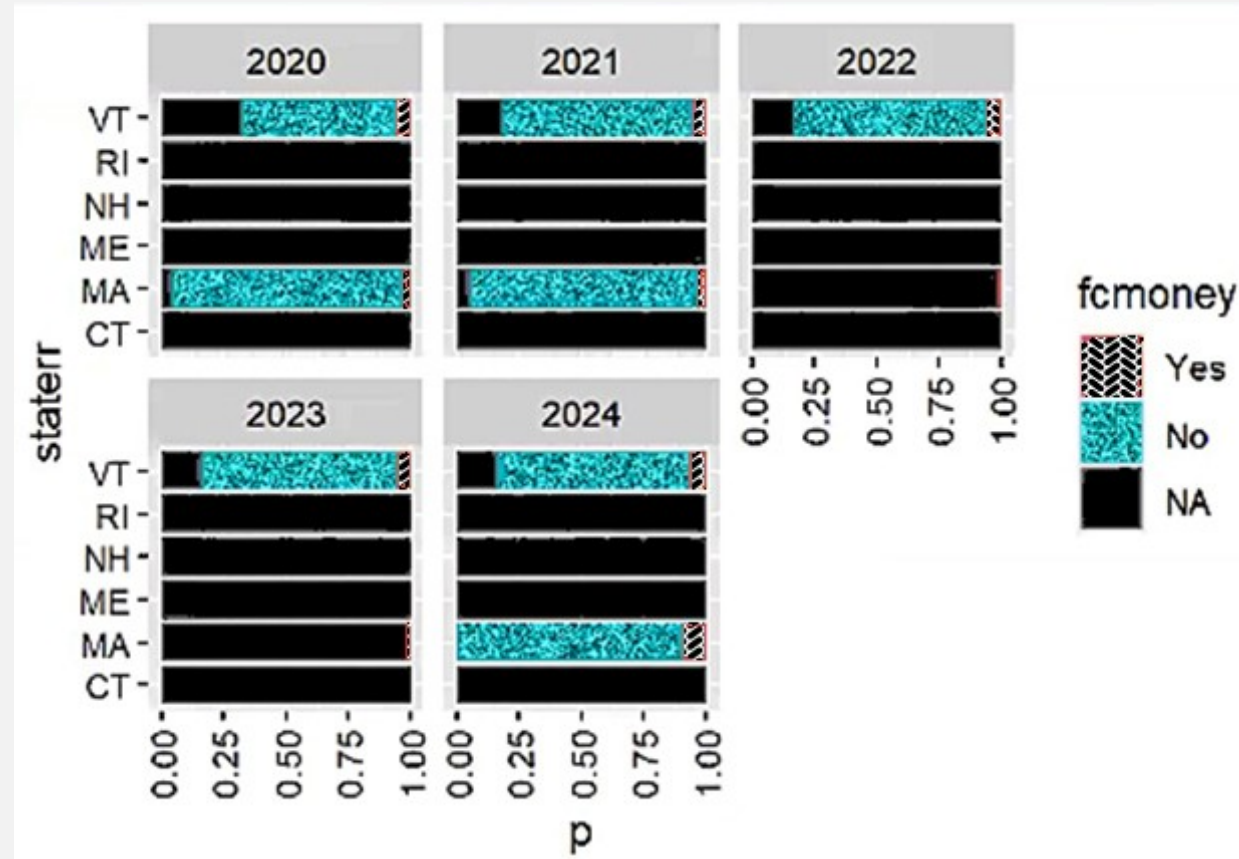
Data in all charts are simulated and not actual NCANDS data.

R CODE IMAGE 2



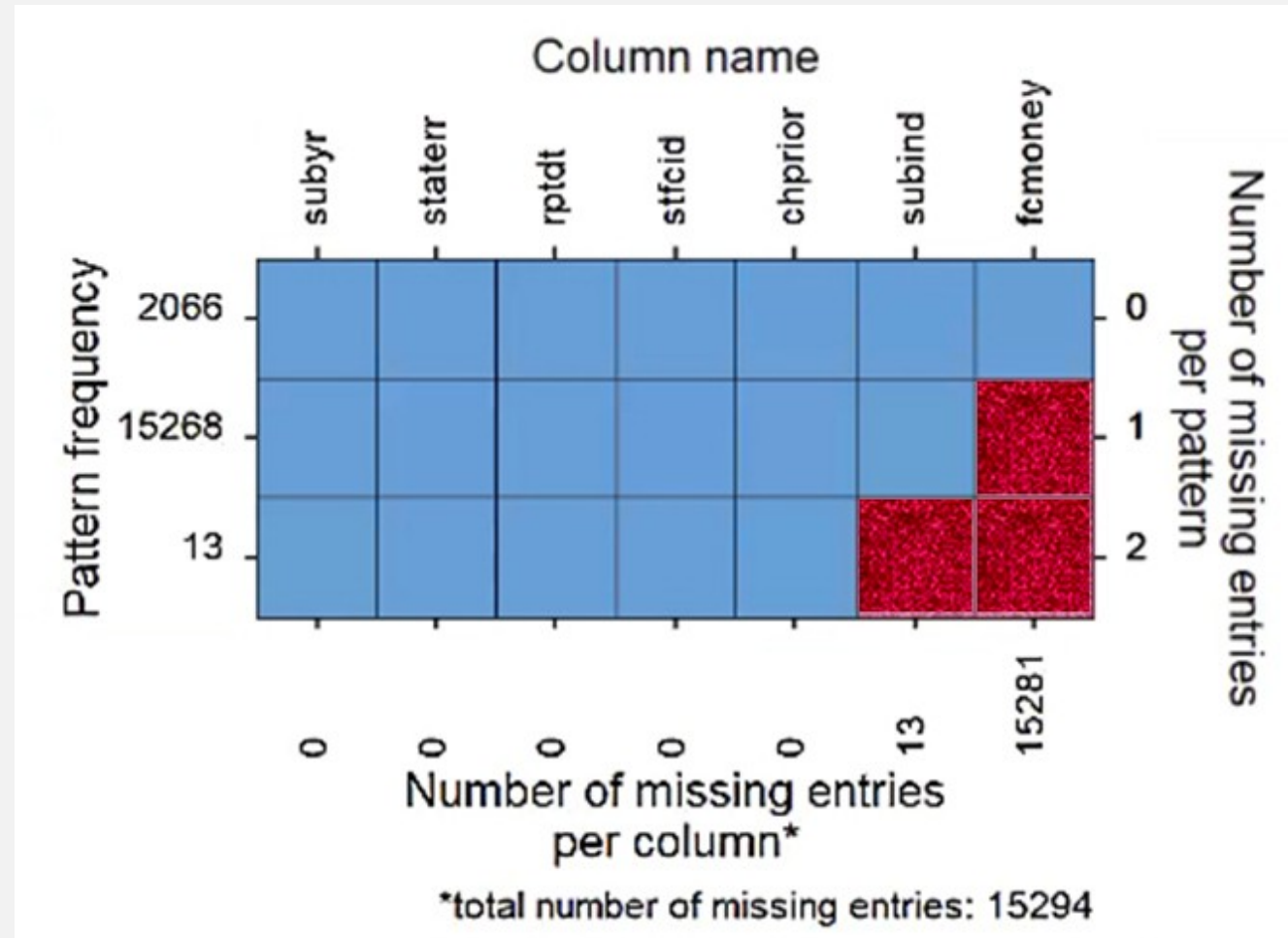
Delayed reporting means records missing at the end of one fiscal year can appear in the subsequent fiscal year.

R CODE IMAGE 3



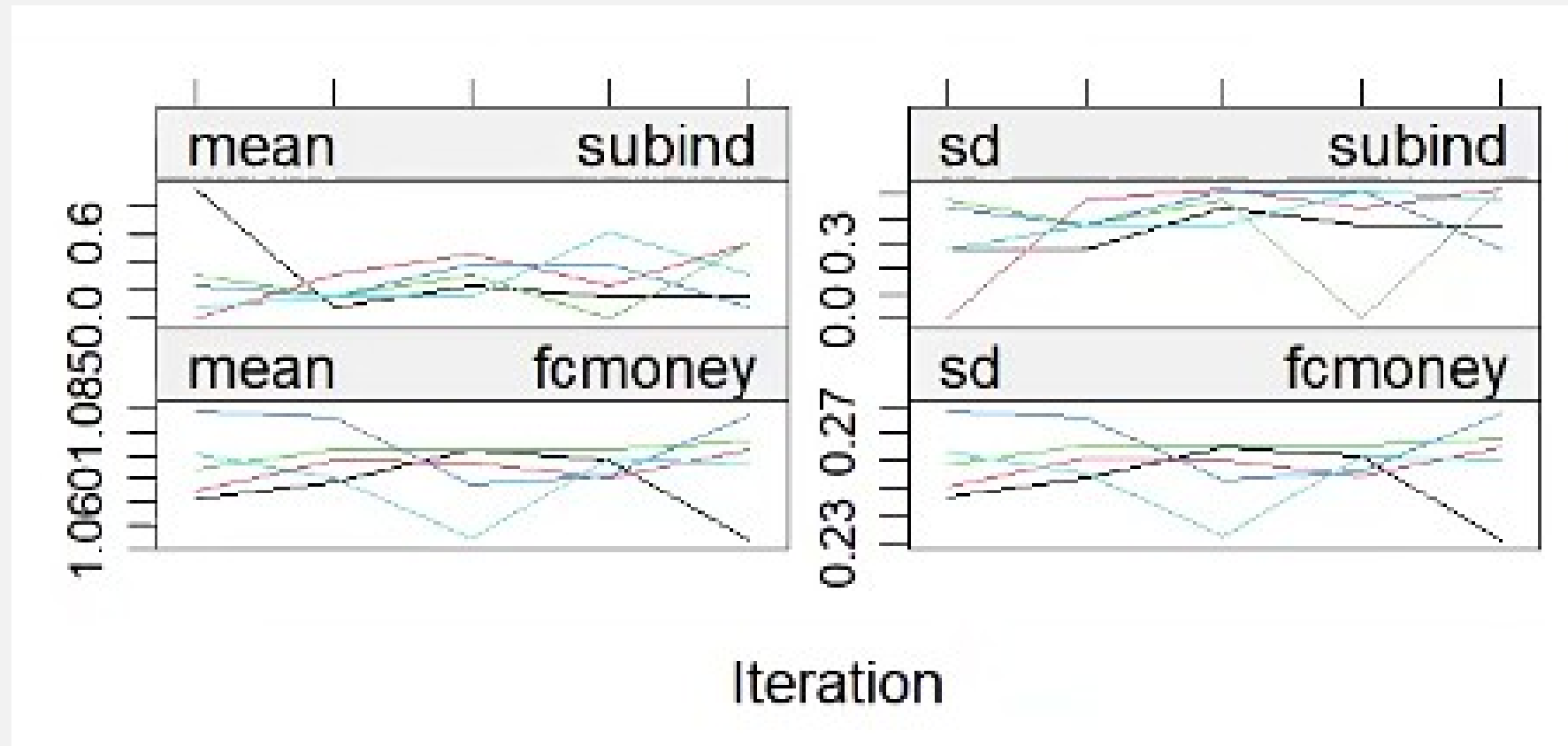
While it's not usually possible to observe the missing-data mechanism directly from the data alone, we can make inferences about it using missing-data patterns. For example, plotting variable distributions by state and year can help identify how much missingness is a function of the state or state-year of report.

R CODE IMAGE 4



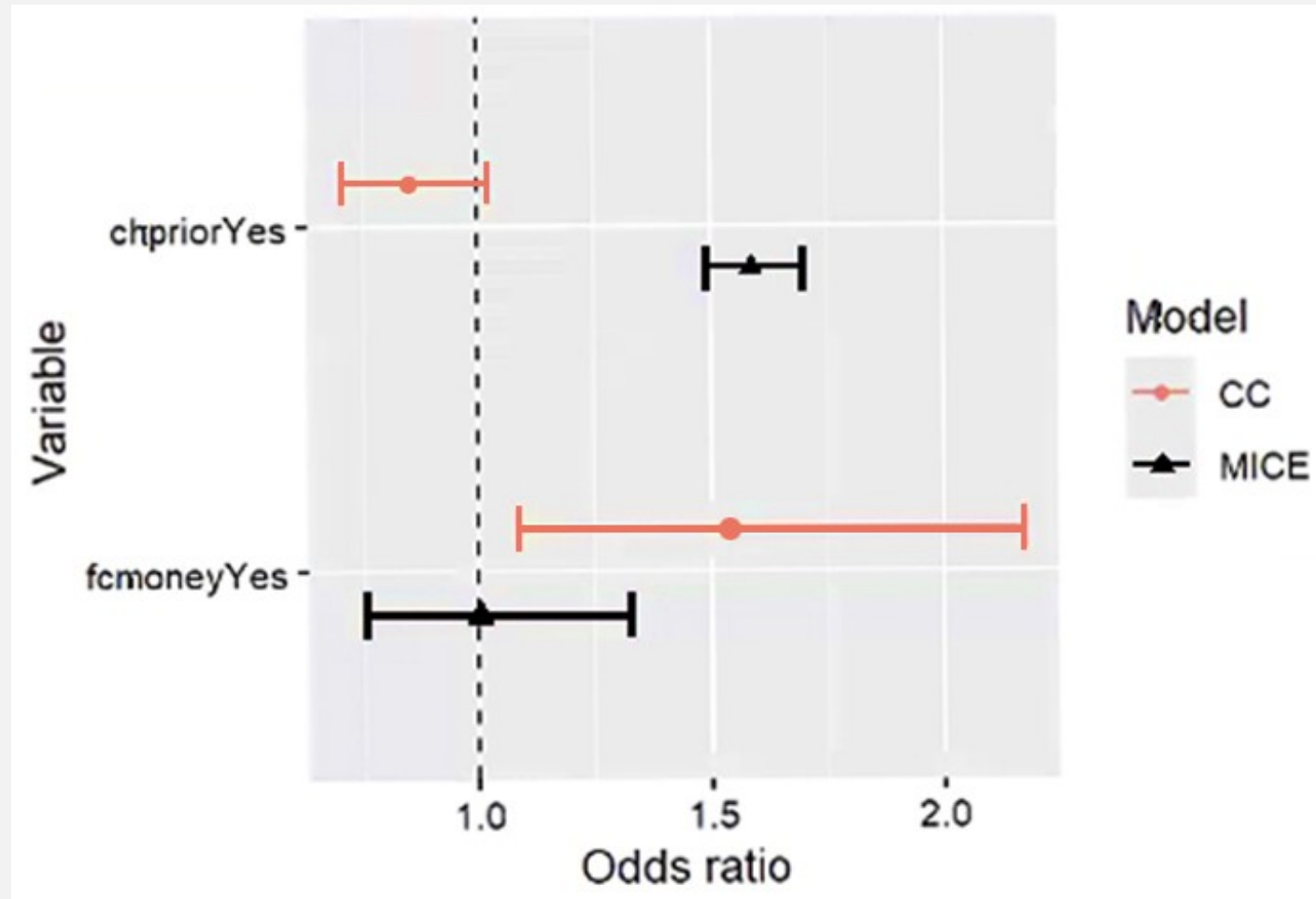
We examine patterns of missingness in the sample, which is important for assessing the feasibility of multiple imputation.

R CODE IMAGE 5



We can easily diagnose the convergence of our imputation model, essentially seeing whether certain statistical assumptions are satisfied for valid imputation (see additional resources).

R CODE IMAGE 6



Forest plot comparing odds ratios for two variables, `chpriorYes` and `fcmoneyYes`, under two models: CC and MICE.