WELCOME TO THE 2024 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

NDACAN SUMMER TRAINING SERIES: BEST PRACTICES IN THE USE OF NDACAN DATA

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

ASSESSING REPORTING ISSUES IN NCANDS & AFCARS

JULY 17, 2024

NDACAN SUMMER TRAINING SERIES SCHEDULE

July 10 — NCANDS: Strengths & Limitations

July 17 — Assessing Reporting Issues in NCANDS & AFCARS

July 24 — AFCARS: Strengths & Limitations

July 31 — Survey Design & Using Weights

August 7 — NSCAW III for Experienced & New Users

August 14 — NYTD: Strengths & Limitations

SESSION AGENDA

- Missing data
- Measurement error
- Record linkage failure
- Demonstration in R

MISSING DATA

TYPES OF MISSING DATA MECHANISM

- Missing completely at random (MCAR): no systematic predictor
  - Listwise deletion okay (except for counting)
- Missing at random (MAR): missingness predicted only by observable factors
  - More sophisticated strategies necessary (imputation, ML or Bayesian methods, etc.)
- Missing not at random (MNAR): missingness predicted by unobservable factors
  - More data needed

SOURCES OF MISSING OBSERVATIONS

- State-level non-reporting

TABLE 1

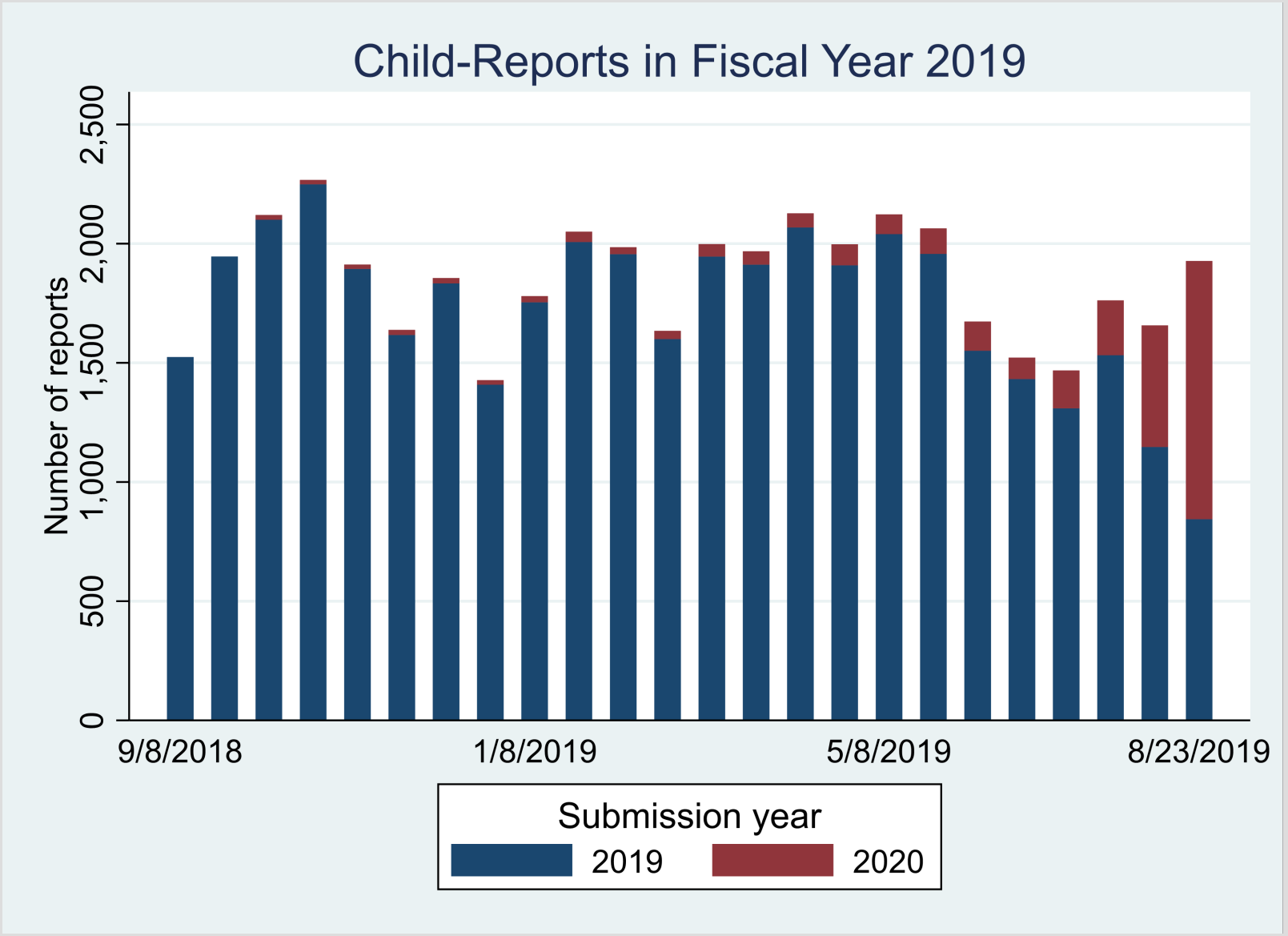
Data Year (Fiscal)	Version	Latest Release	Records	Variables	States Reporting	States Not Reporting	Missing States
2003	7	5/10/2023	3,092,437	143	45	7	AL,AK,GA,ND,OR,WI,PR
2004	5	5/10/2023	3,148,424	143	45	7	AL,AK,GA,ND,OR,WI,PR
2005	7	5/4/2023	3,453,095	143	49	3	ND, OR, PR
2006	6	5/10/2023	3,464,694	143	49	3	MD, ND, OR
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: NCANDS Child Files 2003–2022.

MISSING OBSERVATIONS

- State-level non-reporting
- Delayed reporting

FIGURE 1



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

SOURCES OF MISSING VALUES

- Missing data mechanisms aren't usually directly observable
  - May operate at state, county, caseworker, or child level
- What information can you use to make inferences about the missing data mechanism?
  - Patterns across time, states/counties, variables, etc.

TABLE 2

State	CdAlc	CdDrug	CdEmotnl	CdVisual	CdLearn	CdPhys	CdBehav	CdMedicl
Alabama	100	100	100	100	100	100	100	100
Alaska	5	5	1	1	1	1	1	1
Arizona								
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
Indiana	96	96	96	96	96	96	96	96

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

DEALING WITH MISSING VALUES

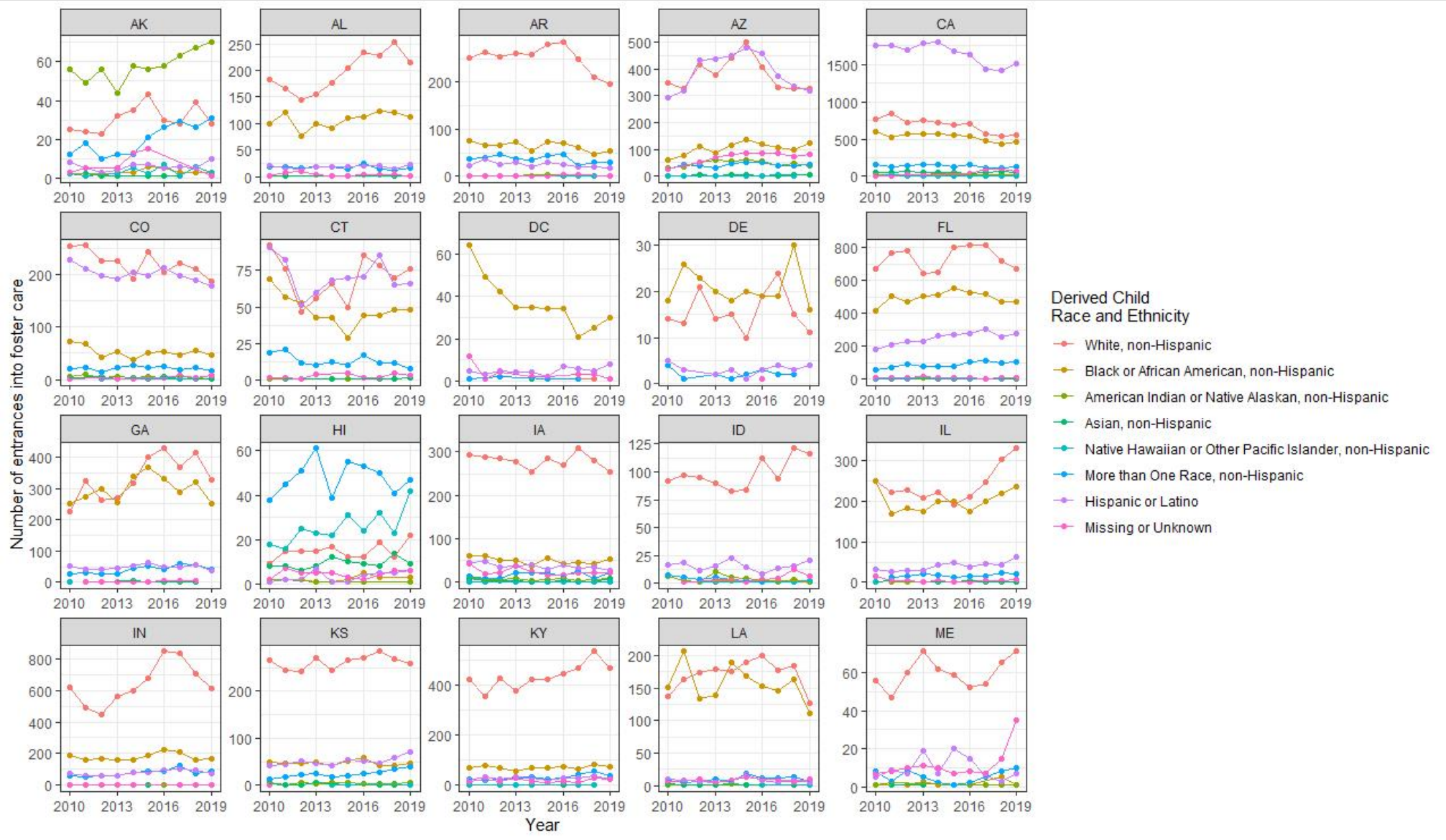
- Based on these inferences, what is the appropriate missing data strategy?
  - Limit scope to jurisdictions with good data (compromise external validity)
  - Employ established method for estimating missing data (potentially compromise internal validity)

MEASUREMENT ERROR

IDENTIFYING MEASUREMENT ERROR

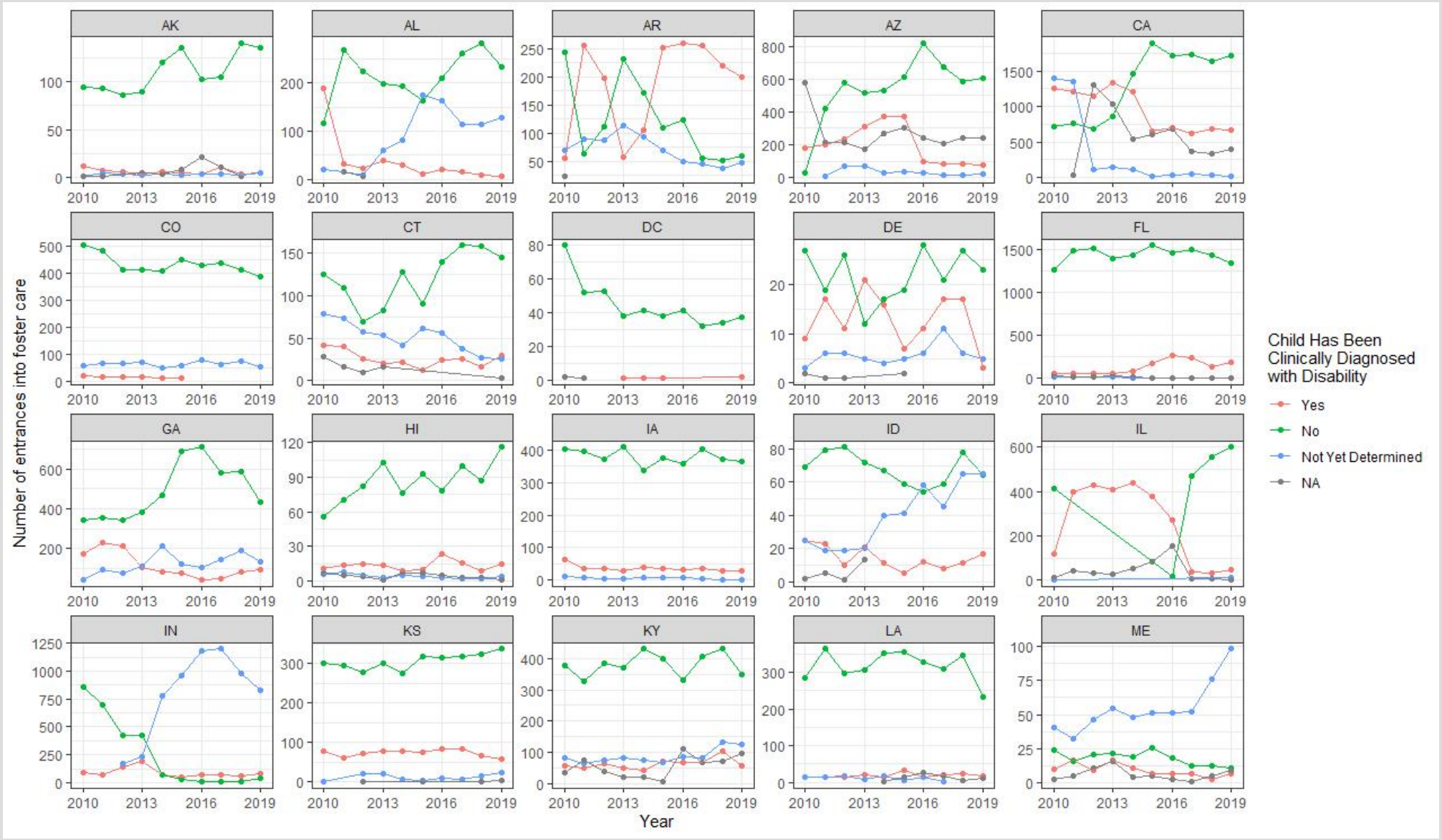
- Just because a value is observed doesn't mean it's true!
- Like missing data mechanisms, measurement error isn't directly observable
- Get clarity: what is the construct you are trying to measure?
  - E.g. if the definition of a value changes, is that a problem for you?
- What information can you use to make inferences about measurement error?
  - Patterns across time, states/counties, variables, repeated observations, etc.

FIGURE 2



Source: AFCARS Foster Care Files 2010-2019

FIGURE 3



Source: AFCARS Foster Care Files 2010-2019

RECORD LINKAGE FAILURE

TABLE 3

State	00-01	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20	20-21
Alabama						97	97	98	1	99	99	99	99	99	99	99	99	99	99	99	99
Arizona			0	99	99	99	99	99	100	99	99	99	99	99	99	99	99	99	99	100	
California			99	100	99	100	98	99	99	99	99	99	99	99	99	99	99	99	99	100	99
Florida	0	0	100	98	99	99	0	99	99	0	99	0	0	99	99	98	98	98	98	99	99
Illinois			0	98	99	0	0	0	0	100	100	0	0	100	100	99	99	99	99	99	99
Indiana			94	99	99	99	99	99	100	99	99	0	98	98	99	98	98	98	98	98	98
Massachusetts	0	0	95	100	97	95	96	97	98	96	96	97	97	96	96	95	100	100	98	98	98
Michigan			0	99	99	99	0		96	97	97	97	97	96	98	99	100	100	99	100	100
Montana		0	0	94	96	94	88	90	94	93	92	94	94	87	94	96	97	97	97	96	99
New York			0	98	96	98	98	98	98	98	98	98	96	96	99	97	98	97	98	98	98
North Dakota											99	100	99	98	99	97	98	98	98	97	98
Oregon												0	100	100	100	99	100	100	99	99	99
Pennsylvania	0	0	0	98	0	0	0	0	0	0	0	94	0	95	0	0	0	0	0	0	97
Tennessee				0	95	96	96	97	98	80	0	0	98	97	97	97	97	99	98	97	98
Texas	0	0	0	98	96	97	97	99	98	98	98	98	98	98	98	98	95	95	95	96	96

Note: Numbers are percentages of children in NCANDS Child Files, linked by ChID in adjacent years, for whom DOB and sex match (true positive link).

TABLE 4

	FY 2021				FY 2020			
State	Foster File Adoption Exits	Adoption File Records	Common Child IDs	Percentage Exits Matched	Foster File Adoption Exits	Adoption File Records	Common Child IDs	Percentage Exits Matched
Alabama	790	792	0	0	807	813	0	0
Alaska	335	339	334	100	351	354	351	100
Arizona	2,320	1,977	1,975	85	2,898	2,902	2,889	100
Arkansas	753	768	0	0	770	777	0	0
California	5,963	6,241	5,833	98	5,282	5,562	5,187	98
Colorado	606	790	26	4	661	832	41	6
Connecticut	437	451	0	0	377	427	0	0
Delaware	86	86	86	100	102	116	102	100
District of Columbia	108	110	0	0	98	98	0	0
Florida	3,873	3,937	3,873	100	4,431	4,525	4,431	100
Georgia	1,262	1,394	1,249	99	1,400	1,583	1,384	99

Source: AFCARS Foster Care and Adoption Files, 2020-2021

DEMONSTRATION IN R

The program, written in R, is included in the downloadable files for the slides and the transcript.

Link to R Code:
https://drive.google.com/file/d/1ZVQyrc6mK32oCp8hOYaRRmM3DnKuMiAi/view?usp=sharing

R CODE PAGE 1 OF 5

```
###########
# NOTES #
###########

# THIS PROGRAM FILE DEMONSTRATES SOME STRATEGIES DISCUSSED IN
# SESSION 2 OF THE 2024 NDACAN SUMMER TRAINING SERIES
# "ASSESSING REPORTING ISSUES IN NCANDS & AFCARS"

# FOR QUESTIONS, CONTACT ALEX ROEHRKASSE
# (AROEHRKASSE@BUTLER.EDU; ALEXR.INFO)

##############
# 0. SETUP #
##############

# Clear environment
rm(list=ls())

# Install packages (only necessary once)
#install.packages(c('data.table', 'tidyverse'))

# Loads packages
library(data.table)
library(tidyverse)

# Set filepaths
afrpc <- 'C:/Users/aroehrkasse/Box/Presentations/-NDACAN/2024_summer_series/'
setwd(afrpc)

# Set seed
set.seed(1013)
```

R CODE PAGE 2 OF 5

```
##########################################
# 1. DELAYED REPORTING IN NCANDS #
##########################################

# NCANDS data files are organized by the year in which
# maltreatment reports are *submitted* to NCANDS.
# A very common oversight is that this is *not* the same
# as the year in which reports *occur.*
# How do we measure the number of reports occurring in Fiscal Year (FY) 2019?
# (FYs are Sep. 1 - Aug. 31)

# Let's read in an anonymized 1% sample of the 2019 NCANDS Child File
d19 <- fread('cf_2019_anon_samp.csv')

# Let's reformat NCANDS's report date variable as a
# date format variable that R can understand.
d19 <- d19 %>%
  mutate(rptdt = as.Date(rptdt))

# And let's plot a histogram of the report dates for the 2019 NCANDS Child File
d19 %>%
  group_by(rptdt) %>% # tell R to group the data by half-month
  summarize(n = n(), .groups = 'keep') %>% # count child-reports in each group
  ggplot(aes(x = rptdt, y = n)) +
  geom_bar(stat = 'identity') + # plot a bar chart of the counts
  geom_vline(aes(xintercept = as.numeric(as.Date('2018-09-01'))),
             color = 'red', size = 1.5) +
  geom_vline(aes(xintercept = as.numeric(as.Date('2019-08-31'))),
             color = 'red', size = 1.5) +
  labs(x = 'Report date', y = 'Number of child-reports') +
  theme_bw()
```

R CODE PAGE 3 OF 5

```
# If our sampling frame is FY2019, we
## (i) have some extra observations we don't want and
## (ii) are seemingly missing some observations that we do want.
# We can solve (ii) by appending submission-year 2020 data, and
# solve (i) by dropping data outside our sampling frame.
d20 <- fread('cf_2020_anon_samp.csv')
d1920 <- d19 %>%
  bind_rows(d20) %>%
  mutate(rptdt = as.Date(rptdt))
d1920 <- d1920 %>%
  filter(rptdt %in% as.Date('2018-09-01'):as.Date('2019-08-31'))

# Now let's replot our histogram as a stacked bar graph
# that illustrates how much each submission year
# contributes reports to any given reporting year.
# Each bar represents a half-month interval.
d1920 %>%
  group_by(rptdt, subyr) %>%
  summarize(n = n(), .groups = 'keep') %>%
  ggplot(aes(x = rptdt, y = n, fill = fct_rev(as.factor(subyr)))) +
  geom_bar(stat = 'identity') +
  labs(x = 'Report date',
       y = 'Number of child-reports',
       fill = 'Submission year') +
  theme_bw()

# This is not so much an issue with AFCARS data if your
# unit of analysis is the FISCAL year,
# but it is an important (and similar) challenge if your
# unit of analysis is the CALENDAR year.
```

R CODE PAGE 4 OF 5

```
#########################################################
# 2. EXAMINING REPORTING OF RACE IN AFCARS #
#########################################################

# Let's read a 5% sample of the AFCARS Foster Care files for 2010-2019
fc <- fread('fc.csv')
head(fc)

# Then let's restructure our data as
# counts of entrances into foster care
# by state, year, and ethnoracial group.
fccount <- fc %>%
  filter(Entered == 1 & St <= 'MA') %>%
  group_by(St, FY, RaceEthn) %>%
  summarize(n = n(), .groups = 'keep') %>%
  mutate(RaceEthn = factor(RaceEthn,
                           levels = c(1:7,99),
                           labels = c('White, non-Hispanic',
                                      'Black or African American,\nnon-Hispanic',
                                      'American Indian/Native Alaskan,\nnon-Hispanic',
                                      'Asian, non-Hispanic',
                                      'Native Hawaiian/Other Pacific Islander,\nnon-Hispanic',
                                      'More than One Race,\nnon-Hispanic',
                                      'Hispanic or Latino',
                                      'Missing or Unknown')))
```

R CODE PAGE 5 OF 5

```
# Now let's plot a series of trend lines
# to examine any possible reporting issues
fccount %>%
  ggplot(aes(x = factor(FY), y = n, color = RaceEthn, group = RaceEthn)) +
  #geom_point() +
  geom_line() +
  scale_x_discrete(breaks = seq(2010, 2019, 3)) +
  labs(color = 'Derived Child\nRace and Ethnicity',
       x = 'Year',
       y = 'Number of entrances into foster care') +
  facet_wrap(~St, scales = 'free') + # faceting can help examine multiple trends
  theme_bw()

# It can be helpful to try different scales
# depending on what you're trying to suss out.
fccount %>%
  ggplot(aes(x = factor(FY), y = n, color = RaceEthn, group = RaceEthn)) +
  #geom_point() +
  geom_line() +
  scale_x_discrete(breaks = seq(2010, 2019, 3)) +
  scale_y_continuous(trans = 'log2') +
  labs(color = 'Derived Child\nRace and Ethnicity',
       x = 'Year',
       y = 'Number of entrances into foster care\n(logarithmic scale)') +
  facet_wrap(~St) + # faceting can help examine multiple trends at once
  theme_bw()
```

QUESTIONS?

ALEX ROEHRKASSE
ASSISTANT PROFESSOR
BUTLER UNIVERSITY

AROEHRKASSE@BUTLER.EDU

NEXT WEEK…

July 24, 2024

at 12pm (Eastern)

Presenter:

Sarah Sernaker, M.S.

Topic:

AFCARS: Strengths & Limitations