# WELCOME TO THE NDACAN SUMMER TRAINING SERIES!

National Data Archive on Child Abuse and Neglect

Duke University, Cornell University, University of California San Francisco, & Mathematica

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

# SUMMER TRAINING SERIES SCHEDULE

- **July 2nd, 2025**
  - Developing a research question & exploring the data
- **July 9th, 2025**
  - Data management
- **July 16th, 2025**
  - Linking data
- **July 23rd, 2025**
  - Exploratory Analysis
- **July 30th, 2025**
  - Visualization and finalizing the analysis

# LIFECYCLE OF AN NDACAN RESEARCH PROJECT

This session is being recorded.

Please submit questions to the Q&A box.

See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us

If issues persist and solutions cannot be found through Zoom, please contact Andres Arroyo at aa17@cornell.edu.

# SESSION AGENDA

- STS recap
- Linking data
- Demonstration in R

# STS RECAP

# DATA MANAGEMENT

- Develop workflow to guard against mistakes
- Observe data security protocols fastidiously
- Verify data features directly
- Format your data according to your unit of analysis

# RESEARCH QUESTION

*What is the relationship between lifetime incidence of removal and full-time employment among youth three years after aging out of foster care?*

# LINKING DATA

# WHAT IS RECORD LINKAGE?

- Linkage combines multiple data sources based on one or more shared variables
- Internal record linkage
  - NDACAN administrative data files (NCANDS, AFCARS, NYTD) can be linked to each other at the child level using unique (encrypted) child IDs
- External record linkage
  - Aggregated NDACAN data can be linked at the aggregate level to external sources using common variables:
    - Time: year, month, half-month
    - Place: state, county
    - Demographic groups: sex, race/ethnicity, age

# JOINING DATA

**NYTD**

| StFCID | CurrFTE_3 |
|---|---|
| A | 1 |
| B | 0 |

**AFCARS**

| StFCID | TOTREM |
|---|---|
| B | 3 |
| C | 4 |

**inner_join**

| StFCID | CurrFTE_3 | TOTREM |
|---|---|---|
| B | 0 | 3 |

**left_join**

| StFCID | CurrFTE_3 | TOTREM |
|---|---|---|
| A | 1 | *NA* |
| B | 0 | 3 |

**full_join**

| StFCID | CurrFTE_3 | TOTREM |
|---|---|---|
| A | 1 | *NA* |
| B | 0 | 3 |
| C | *NA* | 4 |

# LINKING NDACAN ADMINISTRATIVE RECORDS

The variable *RecNumbr* is an encrypted version of the youth's unique identifier used by the state agency. The ID may go by different names in the various linkable files. These are:

- NYTD Outcomes File: *RecNumbr*
- AFCARS Foster Care File: *RecNumbr*
- AFCARS Adoption File: *RecNum* (for some states)
- NCANDS Child File: *AFCARSID*

To facilitate linking data among this family of files, a common linking variable – *StFCID* has been added. It consists of concatenating the state's 2-character postal code to the *ChildID*, resulting in a 14-character variable.

*Source:* NYTD Outcomes File User's Guide

# CAUTION IN LINKING

The […] youth identifier is encrypted for all these datasets, but is encrypted consistently across datasets, so it serves as an indicator of the same youth across datasets and across years. Be careful, however. These commonalities are generally reliable, but are not applicable to all states in all years. Contact NDACAN Support for further information regarding which states can be linked across specific years.

*Source:* NYTD Outcomes File User's Guide

# JOINING MULTIPLE OBSERVATIONS

- **One to one**
  - Linking variable is unique in both datasets
- **One to many**
  - Linking variable is unique in one dataset
- **Many to many**
  - Linking variable is not unique in either dataset

# JOINING DATA: ONE-TO-MANY

**NYTD**

| StFCID | Wave | CurrFTE |
|---|---|---|
| A | 1 | 0 |
| A | 2 | 1 |
| A | 3 | 1 |
| B | 1 | 0 |
| B | 2 | 0 |
| B | 3 | 1 |

**AFCARS**

| StFCID | TOTREM |
|---|---|
| B | 3 |
| C | 4 |

**full_join**

| StFCID | Wave | CurrFTE | TOTREM |
|---|---|---|---|
| A | 1 | 0 | *NA* |
| A | 2 | 1 | *NA* |
| A | 3 | 1 | *NA* |
| B | 1 | 0 | 3 |
| B | 2 | 0 | 3 |
| B | 3 | 1 | 3 |
| C | *NA* | *NA* | 4 |

# BENEFITS TO LINKING

- Combining sources expands range of measures
  - Individual-level record linkage: within NDACAN administrative data
  - Aggregate-level record linkage: external data sources
- Repeated observations:
  - Enhance missing data solutions
  - Help identify and address measurement error
  - Enable longitudinal research designs

# PITFALLS OF LINKING

- Myriad errors arise from linking less-than-clean data
- Non-missing values of shared variables may not agree
- Linking may result in (systematic) measurement error
  - NDACAN child IDs are state-specific; interstate moves lead to false negative links
  - NDACAN data reflect variation and changes in record-keeping
- Data linkage can create/amplify missing data problems

# DEMONSTRATION IN R

# QUESTIONS?

**ALEX ROEHRKASSE**
AROEHRKASSE@BUTLER.EDU

**NOAH WON**
NOAH.WON@DUKE.EDU

**PAIGE LOGAN PRATER**
PAIGE.LOGANPRATER@UCSF.EDU

# NEXT WEEK…

**Date**: July 23rd, 2025

**Topic**: Exploratory Analysis

**Instructor**: Alex Roehrkasse & Noah Won

# R CODE PAGE 1 OF 7

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

# This program file demonstrates strategies discussed in
# session 2 of the 2025 NDACAN Summer Training Series
# "Data Management."

# For questions, contact the presenter
# Alex Roehrkasse (aroehrkasse@butler.edu).

# Note that because of the process used to anonymize data,
# all unique observations include partially fabricated data
# that prevent the identification of respondents.
# As a result, all descriptive and model-based results are fabricated.
# Results from this and all NDACAN presentations are for training purposes only
# and should never be understood or cited as analysis of NDACAN data.


#########################
# TABLE OF CONTENTS #
#########################

# 0. SETUP
# 1. LINKING DATA
# 2. LINKING, SAMPLING, AND MISSING DATA
```

# R CODE PAGE 2 OF 7

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

# Clear environment
rm(list=ls())

# Installs packages if necessary, loads packages
if (!requireNamespace("pacman", quietly = TRUE)){
  install.packages("pacman")
}
pacman::p_load(data.table, tidyverse, mice)

# Defines filepaths working directory
project <- 'C:/Users/aroehrkasse/Box/Presentations/-NDACAN/2025_summer_series/'
data <- 'C:/Users/aroehrkasse/Box/NDACAN/'

# Set working directory
setwd(project)

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

# R CODE PAGE 3 OF 7

```
#######################
# 1. LINKING DATA #
#######################

# Let's read in our cleaned, anonymized
# versions of the 2020 NYTD and AFCARS files
nytd <- fread(paste0(data,'nytd_clean_anonymized.csv'))
afcars <- fread(paste0(data,'afcars_clean_anonymized.csv'))

# For demonstration purposes, let's also read in
# our long version of the NYTD file
nytd_long <- fread(paste0(data,'nytd_clean_anonymized_long.csv'))

# Let's create a variable that will indicate
# which record (row) corresponds to which dataset
nytd <- nytd |>
  mutate(data_nytd = 1)
nytd_long <- nytd_long |>
  mutate(data_nytd = 1)
afcars <- afcars |>
  mutate(data_afcars = 1)

# Let's link our data!
d <- nytd |>
  full_join(afcars, by = 'StFCID')

# Note that if we don't specify 'by',
# the function automatically joins on all shared variables.
d2 <- nytd |>
  full_join(afcars)
```

# R CODE PAGE 4 OF 7

```
# Let's inspect some aspects of the linkage.
# What number and proportion of records
# from each dataset were linked?
d |>
  group_by(data_nytd, data_afcars) |>
  summarize(n = n()) |>
  ungroup() |>
  mutate(prop_nytd = ifelse(!is.na(data_nytd),
                            n/sum(n[!is.na(data_nytd)]),
                            NA),
         prop_afcars = ifelse(!is.na(data_afcars),
                              n/sum(n[!is.na(data_afcars)]),
                              NA))

# Note an important thing about the above merge:
# The structure of the datasets resulted in a one-to-one merge.
# This is always good to check.

d |> filter(data_nytd == 1) |> nrow() == nrow(nytd)
d |> filter(data_afcars == 1) |> nrow() == nrow(afcars)

# This would *appear* to be because the value of the
# linking variable StFCID was distinct in each dataset.
# But was that actually the case?
d |>
  group_by(StFCID) |>
  filter(n()>1) |>
  group_by(data_nytd, data_afcars) |>
  summarize(n = n())
```

# R CODE PAGE 5 OF 7

```
# We can see that there were actually 32 records with
# duplicate IDs (i.e. values of StFCID).
# They didn't affect our linked sample because
# they weren't linked to NYTD,
# but we only avoided a major error here
# out of luck!

# It's good to verify the uniqueness of your linking variables
# before linking, and to inspect the structure of your linked data
# after linking.

#####################################################
# 2. LINKING, SAMPLING, AND MISSING DATA #
#####################################################

# Above we did a full join of our data.
# But we don't want all the AFCARS records that don't get linked.
# So we can left join to keep all records in NYTD,
# including ones that don't get linked.
d_left <- nytd |>
  left_join(afcars, by = 'StFCID')

# Let's inspect this linkage:
d_left |>
  group_by(data_nytd, data_afcars) |>
  summarize(n = n())
```

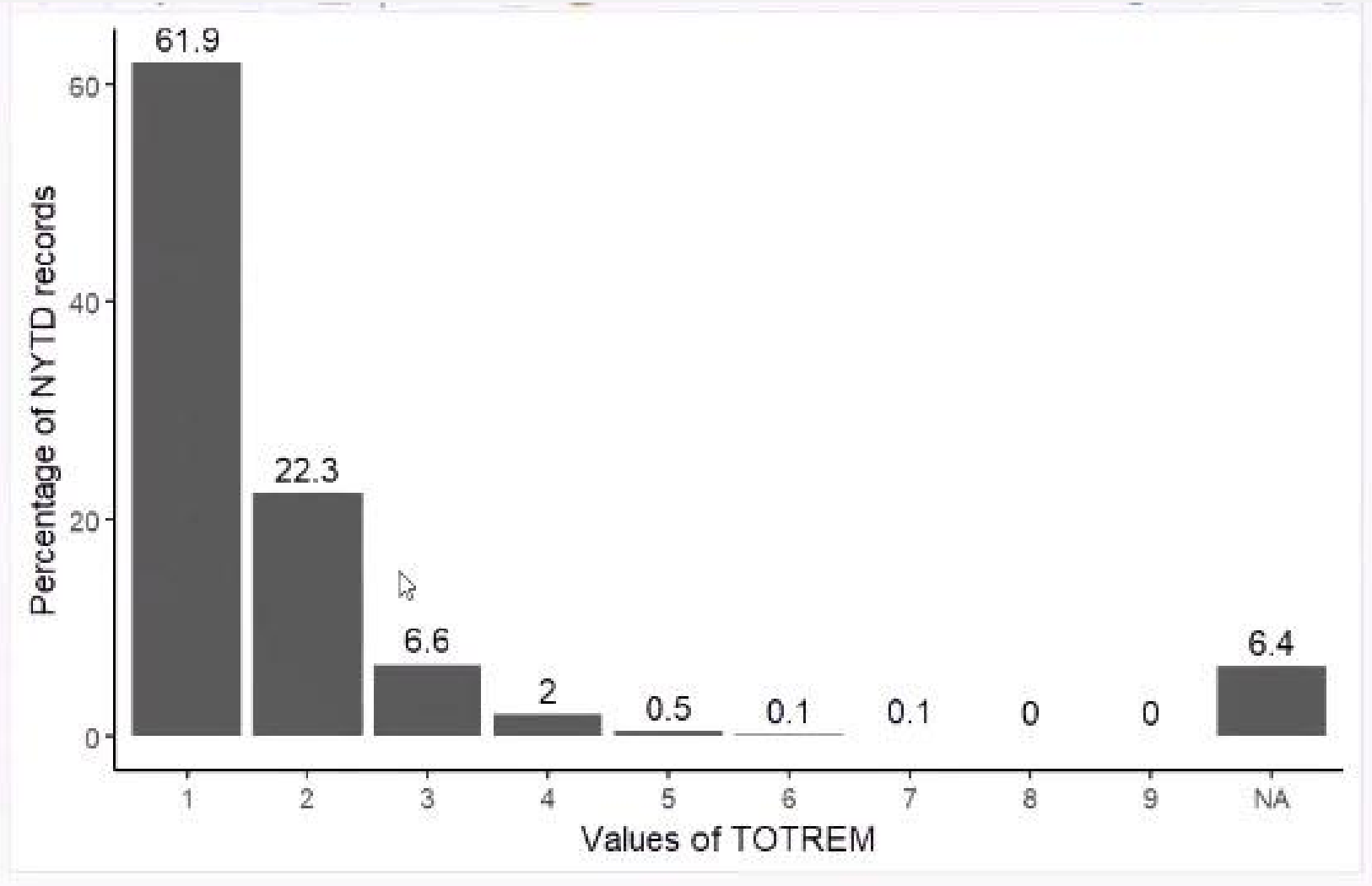
# R CODE PAGE 6 OF 7

```
# The unlinked records from NYTD we want to treat
# as a missing-data problem.
# Let's visualize how much missing data
# for the TOTREM variable that we have
# as a result of failed links:
d_left |>
  group_by(TOTALREM) |>
  summarize(n = n()) |>
  ungroup() |>
  mutate(pct = n/sum(n)*100) |>
  ggplot(aes(x = factor(TOTALREM), y = pct, label = as.character(round(pct,1)))) +
  geom_bar(stat = 'identity') +
  geom_text(vjust = -.5) +
  labs(x = 'Values of TOTREM', y = 'Percentage of NYTD records') +
  theme_classic()

# Let's consider this in the broader context of missing data.
# The mice package is R's best package for multiple imputation.
# It also has helpful diagnostic functions.
d_left |>
  select(State,
         Sex_3, RaceEthn_3,
         CurrFTE_3, CurrenRoll_3,
         TOTALREM) |>
  md.pattern(rotate.names = T)

# Analyzing these patterns helps us choose the best
# missing-data strategy going forward.
# This is beyond the scope of this year's STS,
# but check out NDACAN's various trainings on missing data.
```

# R CODE PAGE 7 OF 7

```
# Lastly, let's save our linked data for next week's session
# on exploratory analysis.
fwrite(d_left, paste0(data,'d_linked_anonymized.csv'))
```

# R CODE IMAGE 1



# R CODE IMAGE 2