# Best Practices for Finding and Using Add Health Data

ROBERT A. HUMMER

HOWARD W. ODUM DISTINGUISHED PROFESSOR OF SOCIOLOGY

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

ADD HEALTH USERS CONFERENCE JULY 12, 2022

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

# Major Takeaways:

1) Use the Add Health Navigator to find measures you may be interested in

2) Work with our contracts team or with ICPSR to acquire the data

3) Add Health has substantial online resources (e.g., User Guides) to help with analyses

4) Complex Add Health survey design necessitates use of appropriate sampling weights to make population estimates

5) Add Health has expertise on hand to help with questions that may not be addressed in online resources

# Finding Add Health Data

# Exploring Add Health Data: The Add Health Navigator

Go to our homepage:

https://addhealth.cpc.unc.edu

… and click on “Add Health Navigator”

1) Add Health Series: Examine available data in each wave and by topic.
2) Search: Type in a concept to find.
3) Explore: See a list of Add Health topics and explore them.
4) Baskets: build your own codebook.

# Obtaining Add Health Data

1) Restricted Use:

Through User Contracts

GWAS Data … through dbGaP

For info, go to:

https://addhealth.cpc.unc.edu/data/

https://www.cpc.unc.edu/projects/addhealth/documentation

2) Public Use: through ICPSR … go to:

https://www.icpsr.umich.edu/web/DSDR/studies/21600

# Other Major Add Health Data Sets

Omics: https://addhealth.cpc.unc.edu/about/omics/
- GWAS
- Candidate Genes
- Polygenic Risk Scores

Add Health Parent Study (PIs: Kathleen Mullan Harris & Joseph Hotz): https://addhealth.cpc.unc.edu/about/#studies-satellite

Sexual Orientation, Gender Identity, Socioeconomic Status, and Health Across the Life Course (PIs: Carolyn Tucker Halpern & Kerith Conron) … data coming soon:

https://www.cpc.unc.edu/research-themes/projects/sexual-orientation-gender-identity-socioeconomic-status-and-health-across-the-life-course/

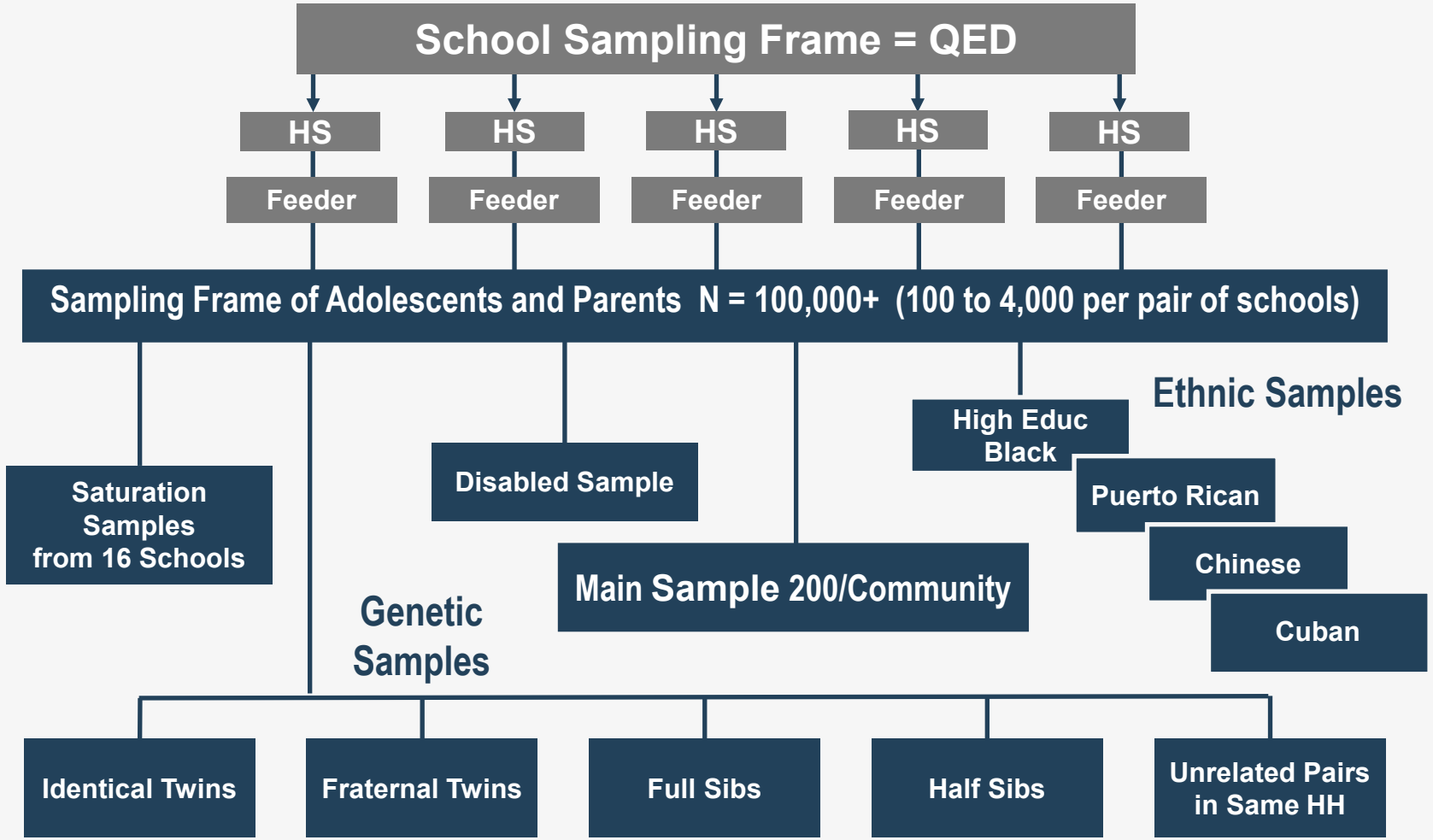
# Using Add Health Data

# Add Health Design and Analysis: Key Resources

1) Harris, Kathleen Mullan, Carolyn Tucker Halpern, Eric A. Whitsel, Jon M. Hussey, Ley KilleyaJones, Joyce Tabor, and Sarah C. Dean. **Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health).** International Journal of Epidemiology (published online June 29, 2019) https://doi.org/10.1093/ije/dyz115

2) User Guides: https://www.cpc.unc.edu/projects/addhealth/documentation/guides

… especially: Chen, Ping, and Kathleen Mullan Harris. 2020. Guidelines for Analyzing Add Health Data. Carolina Population Center at the University of North Carolina at Chapel Hill. https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/GuidelinesforAnalysisofAddHealthData_020422.pdf




# Add Health Sample Design

# Effects of Add Health Survey Design

- **Stratification by region**
- **Clustering of individuals within schools**
  - Adolescents within the same school are not independent of one another
  - Adolescent outcomes will be more similar within schools than across schools
- **Unequal probability of sample selection**

# Sampling Design Adjustments

| Design Attribute | Usual impact | Adjustment variable |
|---|---|---|
| Stratification | Reduce variance | Poststratification variable: census region |
| Clustering of students | Increase variance | PSU variable: School Identification |
| Unequal probability of selection | Increase variance;<br>Biased parameter estimate | Sampling Weights:<br>• Cross-sectional weights for schools<br>• Cross-sectional weights for analyzing each wave of data<br>• Cross-sectional weights for analyzing sub-samples from WIII<br>• Longitudinal weights for conducting analysis combining data from multiple waves<br>• Multilevel weights for two-level analysis where schools and individuals are levels of interest |

# Selection of Sampling Weights for Analysis

# Cross-Sectional Grand Sample Weights Single-Level (Population Average) Models

| Data Set (Year collected) | Sampling Weight Variable (N) | Sample | Target Population |
|---|---|---|---|
| Wave I (1995) | GSWGT1 (N=18,924) | Adolescents chosen with a known probability of being selected from 1994-1995 enrollment rosters of US schools | Adolescents in 1995 enrolled in grades 7-12 during the 1994-1995 academic year |
| Wave II (1996) | GSWGT2 (N=13,570) | Wave I respondents who were interviewed at Wave II | Adolescents in 1996 enrolled in grades 7-12 during the 1994-1995 academic year |
| Wave III (2001) | GSWGT3_2 (N=14,322) | Wave I respondents who were interviewed at Wave III | Adults in 2001 enrolled in grades 7-12 during the 1994-1995 academic year |
| Wave IV (2008) | GSWGT4_2 (N=14,800) | Wave I respondents who were interviewed at Wave IV | Adults in 2008 enrolled in grades 7-12 during the 1994-1995 academic year |
| Wave V (2016-2018) | GSWGT (N=12,300) | Wave I respondents who were interviewed at Wave V | Adults in 2016-18 enrolled in grades 7-12 during the 1994-1995 academic year |

# Using Cross-Sectional Weights

| X Variable | Y Variable | Weight |
|---|---|---|
| Same wave as Y | One Wave | Cross-sectional (population-average model) |
| Multiple Waves | One Wave | Cross-sectional (population-average model) |

# Longitudinal Weights: Single-Level (Population Average) Models

| Data Set (Year collected) | Sampling Weight Variable (N) | Sample | Target Population |
|---|---|---|---|
| Wave III (2001) | GSWGT3 (N=10,828) | Eligible Wave I Respondents who were interviewed at both Wave II & Wave III | Adolescents enrolled in grades 7-12 during the 1994-1995 academic year interviewed in 1994, 1996 & 2001 |
| Wave IV (2008) | GSWGT4 (N=9,421) | Eligible Wave I respondents who were interviewed at Wave II, III & IV | Adolescents enrolled in grades 7-12 during 1994-1995 interviewed in 1995, 1996, 2001 & 2008 |
| Wave IV (2008) | GSWGT134 (N=12,288) | Eligible Wave I respondents who were interviewed at Wave III & IV | Adolescents enrolled in grades 7-12 during 1994-1995 interviewed in 1995, 2001 & 2008 |
| Wave V (2016-2018) | GSWGT12345 (N=7,295) | Eligible Wave I respondents who were interviewed at II, III, Wave IV & V | Adolescents enrolled in grades 7-12 during 1994-1995 interviewed in 1995, 1996, 2001, 2008 & 2016-18 |
| Wave V (2016-2018) | GSWGT1345 (N=9,349) | Eligible Wave I respondents who were interviewed at Wave III, IV & V | Adolescents enrolled in grades 7-12 during 1994-1995 interviewed in 1995, 2001, 2008 & 2016-18 |
| Wave V (2016-2018) | GSWGT145 (N=10,914) | Eligible Wave I respondents who were interviewed at Wave IV & V | Adolescents enrolled in grades 7-12 during 1994-1995 interviewed in 1995, 2008 & 2016-18 |

# Using Longitudinal Weights

| X Variable | Y Variable | Weight |
|---|---|---|
| Wave I | Multiple Waves<br>for a sample of respondents who have data at every wave | Longitudinal |

# Sampling Weights for Time-to-Event (Survival/Hazard) Analysis: One Time Point

| Data Source<br><br>(Y from One Wave) | Weight for Population Average Models | Weights for Multilevel Models | Sample | Target Population |
|---|---|---|---|---|
| Wave I only (1995) | GSWGT1<br>(N=18,924) | SCHWT1 (N=132)<br>W1_WC (N=18,924) | Adolescents chosen with a known probability of being selected from 1994-1995 enrollment rosters of US schools. | Adolescents in 1995 enrolled in grades 7-12 during 1994-1995 |
| Wave II only (1996) | GSWGT2<br>(N=13,570) | W2_WC (N=13,568) | Wave I respondents who were interviewed at Wave II. | Adolescents in 1996 enrolled in grades 7-12 during 1994-1995 |
| Wave III only (2001) | GSWGT3_2<br>(N=14,322) | W3_2_WC (N=14,322) | Wave I respondents who were interviewed at Wave III. | Adults in 2001 enrolled in grades 7-12 during<br>1994-1995 |
| Wave IV only (2008) | GSWGT4_2<br>(N=14,800) | W4_2_WC (N=14,800) | Wave I respondents who were interviewed at Wave IV. | Adults in 2008 enrolled in grades 7-12 during<br>1994-1995 |
| Wave V only<br>(2016-18) | GSWGT<br>(N=12,300) | W5_2_WC<br>(N=12,300) | Wave I respondents who were interviewed at Wave V. | Adults in 2016-18 enrolled in grades 7-12 during<br>1994-1995 |

# Sampling Weights for Time-to-Event (Survival/Hazard) Analysis: Multiple Time Points

| Data Source<br><br>(Y from Multiple Waves) | Weight for Population Average Models | Weights for Multilevel Models | Target Population |
|---|---|---|---|
| Wave I & II | GSWGT1<br>(N=18,924) | SCHWT1 (N=132)<br>W1_WC (N=18,924) | Adolescents in 1995 enrolled in grades 7-12 during 1994-1995 |
| Wave II & III | GSWGT2<br>(N=13,570) | SCHWT1 (N=132)<br>W2_WC (N=13,568) | Adolescents in 1996 enrolled in grades 7-12 during 1994-1995 |
| Wave I, II, & III | GSWGT1<br>(N=18,924) | SCHWT1 (N=132)<br>W1_WC (N=18,924) | Adolescents in 1995 enrolled in grades 7-12 during 1994-1995 |
| Wave I, II, III, & IV | GSWGT1<br>(N=18,924) | SCHWT1 (N=132)<br>W1_WC (N=18,924) | Adolescents in 1995 enrolled in grades 7-12 during 1994-1995 |
| Wave I, II, III, IV & V | GSWGT1<br>(N=18,924) | SCHWT1 (N=132)<br>W1_WC (N=18,924) | Adolescents in 1995 enrolled in grades 7-12 during 1994-1995 |

# Sampling Weights for Wave V

| Data Set (Year collected) | Sampling Weight Variable (N) | Type | Sample | Target Population |
|---|---|---|---|---|
| Wave V (2016-18) | GSW5 (N=12,300) | Cross-sectional weight | Eligible Wave I respondents interviewed at Wave V | Grades 7-12 in 1994-95 |
| Wave V (2016-18) | GSW12345 (N=7,295) | Longitudinal weight | Eligible Wave I respondents interviewed at Waves II, III, IV & V | Grades 7-12 in 1994-95 |
| Wave V (2016-18) | GSW1345 (N=9,349) | Longitudinal weight | Eligible Wave I respondents interviewed at Waves III, IV & V | Grades 7-12 in 1994-95 |
| Wave V (2016-18) | GSW145 (N=10,914) | Longitudinal weight | Eligible Wave I respondents interviewed at Waves IV & V | Grades 7-12 in 1994-95 |
| Wave V (2016-18) | W5BIOWGT (N=5,377) | Cross-sectional weight | Wave I respondents who were interviewed at Wave V | Adolescents in grades 7-12 in 1994-95 interviewed at Wave V who participated in the biomarker data collection. |

# Longitudinal Analysis

Research questions that investigate changes in measures taken on same respondents over time

Outcome variable is measured multiple times

Data organization:

1) "Stacked records": One record per respondent (AID) per time point (i.e., wave)
   - use cross-sectional weight specific to each wave of measurement
2) Multiple waves of measures combined to create new record that computes differences (i.e., change) in values of variables across waves
   - use cross-sectional weight of most recent wave of measurement

# Examples of Sampling Weights for Wave III Special Sub-Samples: Estimating Single-Level (population average) Models

| Data Set (Year collected) | Sampling Weight Variable (N) | Sample | Represented Population of Interest |
|---|---|---|---|
| Wave III (2001) | W3PTNR (N=1,317) | Wave III Romantic Partner Sample: Eligible Wave I respondents and romantic partners interviewed at Wave III. | Romantic Partners of Adolescents enrolled in Grades 7-12 in 1994-1995 |
| | TWGT3_2 (N=11,637) | Wave III Education Sample: Eligible Wave I respondents interviewed at Wave III. | Adolescents enrolled in Grades 7-12 in 1994-1995 who participated in high school transcript study |
| | TWGT3 (N=8,847) | Wave III Education Sample: Eligible Wave II respondents interviewed at Wave III. | Same as above |

# Preparing Data for Analysis

- Determine the wave(s) of data
- Exclude/delete cases with missing sampling weights
- Design type: Specify With Replacement as the Design Type
- Stratum Variable: **REGION**
- Cluster Variable or Primary Sampling Unit (PSU): **PSUSCID**

# Example 1. Descriptive Statistics

*Research Question*: What is the mean number of hours of TV watched during a week among adolescents (data from Wave I in-Home Questionnaire)?

Notes: Each program specifies the stratification variable (region), sampling weight variable (gswgt1), and cluster (primary sampling unit) variable (psuscid). Stata and SAS default to a "With Replacement" sample.

***SAS syntax:***

```
proc surveymeans data=ahw1;
var hr_tv;
cluster psuscid;
strata region;
weight gswgt1;
run;
```

***STATA syntax:***

```
use ahw1.dta, clear
svyset psuscid [pweight=gswgt1], strata(region)
svy: mean hr_tv
```

# Example 1. Descriptive Statistics

Parameter estimates and standard errors to predict the average number of hours TV watched during a week by adolescents

| Variable | SAS<br>Estimate (Std Err) | Stata<br>Estimate (Std Err) |
|---|---|---|
| hr_tv | 15.57 (.36) | 15.57 (.36) |

# Example 2. Regression Example for Single-Level Model

*Research Question:* Is the performance on the Add Health vocabulary test (PVT_PT1C) influenced by an adolescent's age (AGE_W1), gender (BOY), and time spent watching TV (HR_WATCH)?

*STATA syntax:*

```
use ah2006.dta, clear
svyset psuscid [pweight=gswgt1], strata(region)
svy: regress pvtpct1c agew1 boy hr_watch
```

*SAS syntax:*

```
proc surveyreg data=from_w1;
cluster psuscid;
strata region;
weight gswgt1;
model pvtpct1c=agew1 boy hr_watch;
run;
```

# Example 2. Regression Example for Population-Average Models

Parameter estimates and standard errors to predict the percentile score on the Add Health PVT test

| Parameter | SAS Estimate (Std Err) | Stata Estimate (Std Err) |
|---|---|---|
| $\beta_0$ (INTERCEPT) | 69.946 (7.855) | 69.946 (7.854) |
| $\beta_1$ (AGE_W1) | -1.085 (0.489) | -1.085 (0.489) |
| $\beta_2$ (BOY) | 3.395 (0.673) | 3.395 (0.673) |
| $\beta_3$ (HR_WATCH) | -0.150 (0.020) | -0.150 (0.020) |

# Add Health Contact Info

Many other analytic questions possible:

- Multi-level modeling
- Sub-population analyses
- Multiple imputation
- Etc.

Every user tends to have unique questions

Please consult User Guides. If still unsure, send questions to:
addhealth@unc.edu

# Acknowledgements

Wave VI of Add Health is supported by two grants from the National Institute on Aging (1U01AG071448, principal investigator Robert A. Hummer, and 1U01AG071450, principal investigators Allison E. Aiello and Robert A. Hummer) to the University of North Carolina at Chapel Hill. Co-funding for Wave VI is being provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Institute on Minority Health and Health Disparities, the National Institute on Drug Abuse, the NIH Office of Behavioral and Social Science Research, and the NIH Office of Disease Prevention. Waves I-V data are from the Add Health Program Project, grant P01 HD31921 (Kathleen Mullan Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. The content of this presentation is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or the University of North Carolina at Chapel Hill.

Add Health was originally designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill. Add Health is currently directed by Robert A. Hummer; it was previously directed by Kathleen Mullan Harris (2004-2021) and J. Richard Udry (1994-2004).

Information on obtaining Add Health data is available on the project website (https://addhealth.cpc.unc.edu).