# Introduction to Add Health

Luyin Zhang

Princeton University

April 16, 2025

# Outline


1. Study Design and Sampling
2. Datasets and Variables
3. Data Access
4. Special Features: Genetic Data
5. Resources

# Outline


1. Study Design and Sampling
2. Datasets and Variables
3. Data Access
4. Special Features: Genetic Data
5. Resources

# What is Add Health?

- The National Longitudinal Study of Adolescent to Adult Health (Add Health)
- One of the most comprehensive longitudinal studies of adolescents in the United States
- Follows a nationally representative sample from adolescence into adulthood
- Began in 1994-95 with adolescents in grades 7-12
- Now includes five waves of data spanning over 25 years
- Interdisciplinary design: social, behavioral, and biological data

# Sampling Design

- School-based, cluster sampling design
- Implicit stratification by size, school type, census region, level of urbanicity, and percent white
- Schools selected with probability proportional to size
- Replacement schools if original school refused to participate or was not eligible
- A sample of 80 high schools and 52 associated feeder schools
- In-school questionnaires: all students in grade 7 through 12 within the 132 sample schools – over 90,000 students

# Sampling Design

- In-home interviews in Wave I
  - A core sample of 16,044 students
  - All of the students (N = 3,350) at two high schools
  - Supplementary samples
- The core sample: roughly equal-sized samples drawn from 12 student-level strata
- Two groups of supplementary samples (i.e., oversamples)
  - the non-genetic supplements: ethnic minorities & students with disabilities
  - the genetic supplements: individual students and pairs of students in various types of sibling relationships
- Wave I: 20,745 adolescents
- Sampling weights, stratum, and PSU variables are provided
- See details here

# Sample Weights

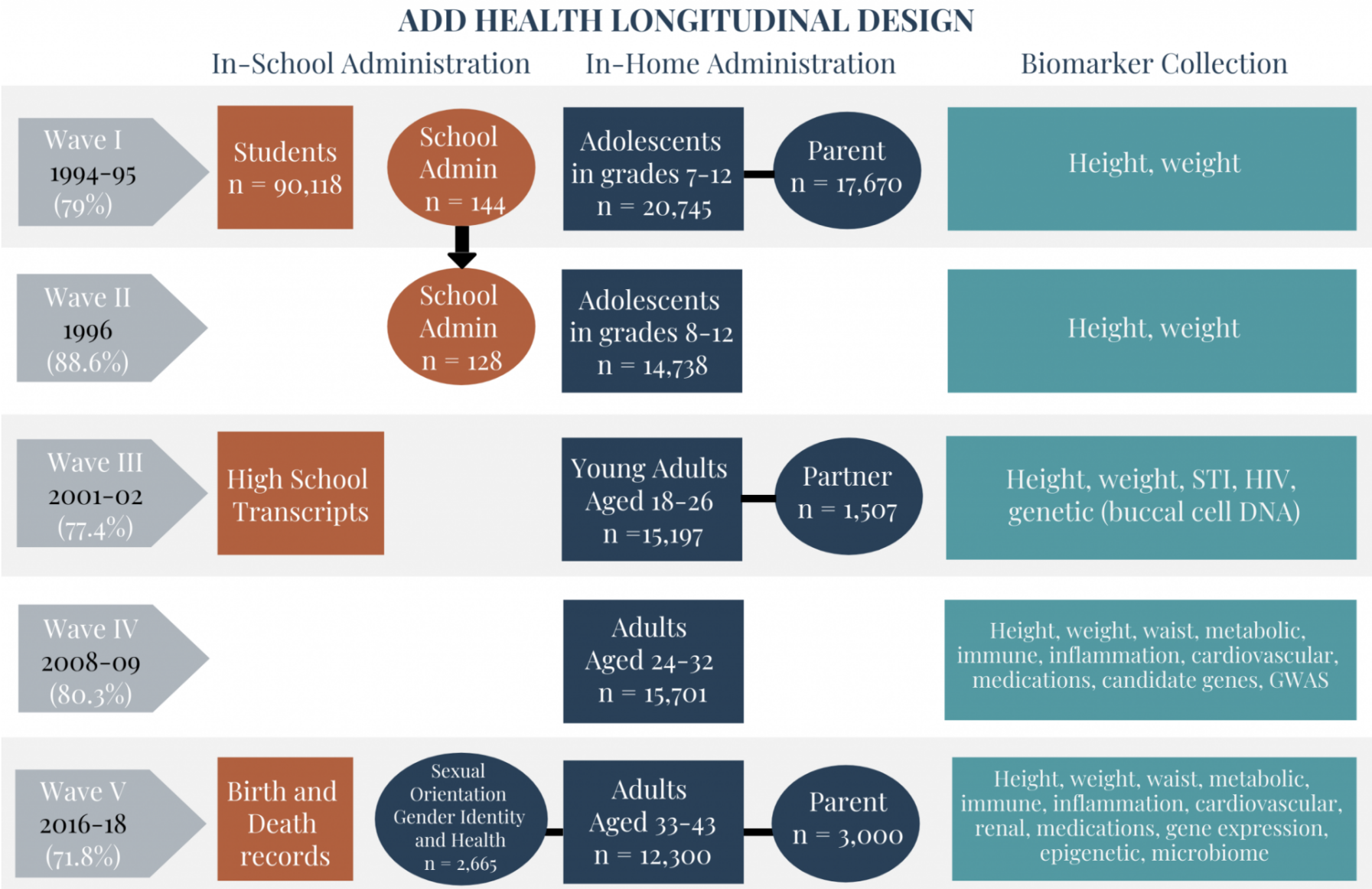
- Weights adjust for
  - Unequal probability of selection
  - Non-response bias
  - Post-stratification to population totals
- Different weights for different waves and combinations
- Cross-sectional weights vs longitudinal weights
- See details here

# Wave Structure

| **Wave** | **Year(s)** | **Ages** |
|---|---|---|
| Wave I | 1994-95 | 12-18 |
| Wave II | 1996 | 13-19 |
| Wave III | 2001-02 | 18-26 |
| Wave IV | 2008 | 24-32 |
| Wave V | 2016-18 | 33-43 |

# Wave Structure

- Sample size decreases across waves due to attrition



# Outline


1. Study Design and Sampling
2. Datasets and Variables
3. Data Access
4. Special Features: Genetic Data
5. Resources

# Multi-level and Multi-source Data

- **Individual level**: Survey responses
- **Family level**: Parent interviews (primarily Wave I)
- **School level**: Administrator surveys, classroom data
- **Contextual data**: State, county, tract, & block group level information, pseudo-IDs for coordinates & geographic identifiers
- **Social network and relational data**: Friendship nominations, romantic partnerships
- **Biomarker data**: Blood, saliva, physical measurements
- **Genetic data**: Genotype data, imputed data (via dbGaP), and polygenic scores
- DNA methylation data to come soon – epigenetic clock or biological ageing
- **Constructued variables**: E.g., family structure, SES indices, mover distance

# Datasets

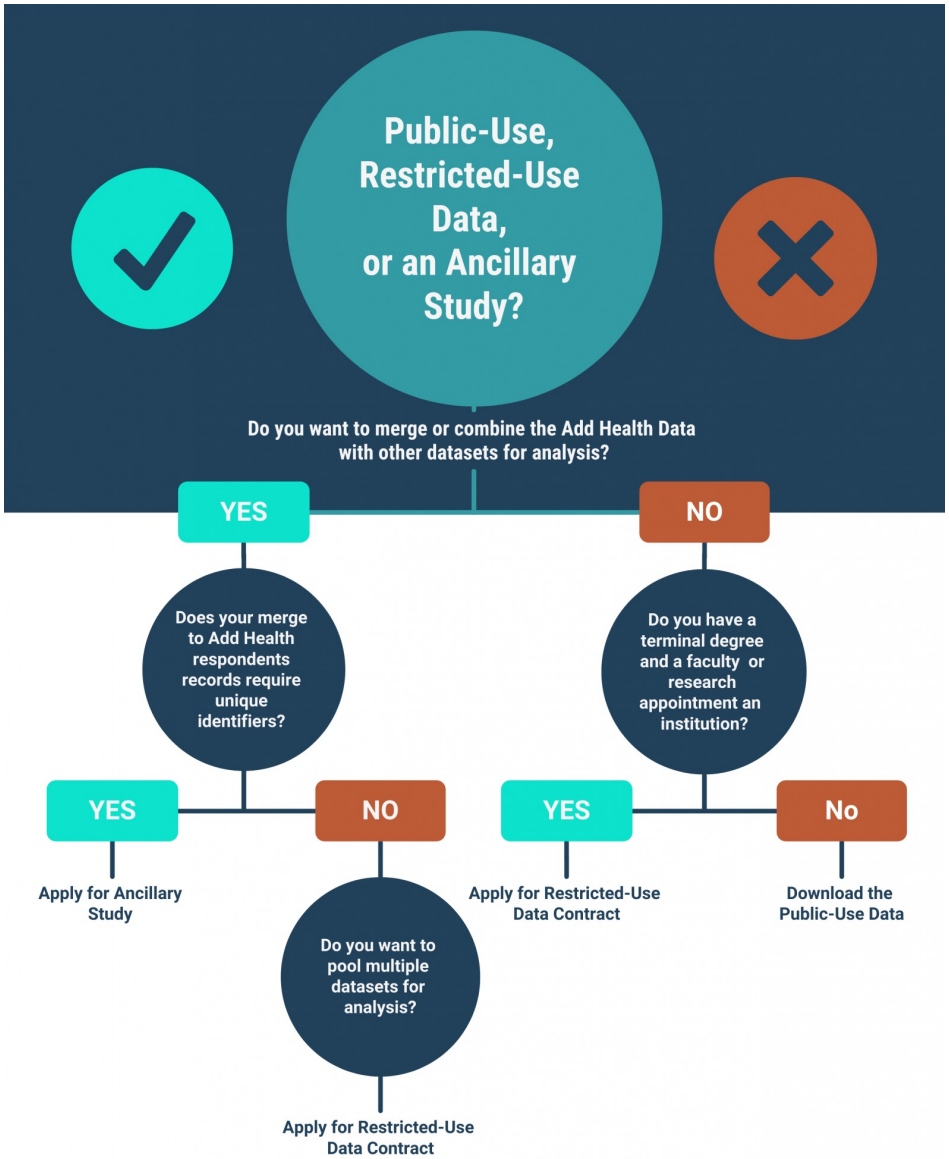
- Don't know where to start with? This website here can be helpful
- The online codebook explorer is even more helpful
- **Public-use data**
  - Smaller ($\sim$ 6,000 cases)
  - Fewer variables
  - Less detailed geographic information
  - Freely available to researchers
- **Restricted-use data**
  - Full sample
  - Complete set of variables
  - Detailed contextual and geographic data
  - Requires data security plan and contract

# Datasets

- **Ancillary studies**

| Study ID | Study Title | Data Released File Name | Release Date |
|---|---|---|---|
| 200600-01 | Wave III Contextual Data | Wave III Contextual Data | 2008 |
| 200600-03 | Alcohol outlet density, alcohol use, and intimate partner violence | Wave III Alcohol Outlet Density Data | 2010 |
| 200600-04 | Additional genotypes - Wave III full sibs and twins | Adolescent Pairs Data | 2006 |
| 200600-05 | Waves I, II, and III contextual and build environment data | ONE: Obesity and Neighborhood Environment Files | 2009 |
| 200600-06 | Wave III education data - design and implementation of the adolescent health and academic achievement study | Wave III Education | 2007 |
| 200600-09 | Molecular genetics and behavior: alcohol and tobacco use | Wave III DNA | 2013 |
| 200704-04 | Whole Genome Association of Alcohol, Tobacco, and BMI | Wave IV Education Genetic Risk Score | 2015 |
| 200704-05 | Exploration of SNPs associated with BMI | Wave IV BMI Genetic Risk Score | 2015 |
| 200809-11 | Gene-environment interactions with the political context | Wave I, II, III Political Context Database | 2010 |
| 201105-19 | Healthy people and healthy neighborhoods: an empirical model of weight status for young asults in the U.S. | Wave III & IV Supplementary Tract-Level Database | 2013 |
| 201108-20 | Ambient air pollutant exposures and cardiovascular health effects among the National Longitudinal Study for Adolescent Health cohort | Wave IV Ambient Air Pollutants Data | 2018 |
| 201112-22 | Creating and utilizing a wave IV contextual database | Wave IV Contextual Data | 2012 |
| 201309-29 | Study of social studies coursetaking and civic engagement, using the National Longitudinal Study of Adolescent Health | Wave III Academic Transcript Social Studies and Civic Coursework (ATRCVC) Data | 2018 |
| 201312-30 | Racial inequalities in marriage outcomes | Wave III Sex Ratio | 2015 |
| 201502-36 | Creating a sexual minority policy contextual database | Wave III & IV Sexual Minority Policy Data | 2019 |
| 201603-42 | Understanding the short and long-term effects of sleep on BMI in adolescents and young adults using an instrumental variables approach | Wave I, III & IV Sunset Data | 2019 |

# What Data to Use?



# Outline


1. Study Design and Sampling
2. Datasets and Variables
3. Data Access
4. Special Features: Genetic Data
5. Resources

# Access Pathways

- **Public-Use Data**
  - Available through ICPSR (Inter-university Consortium for Political and Social Research) or ARDA (Association of Religion Data Archives)
  - Requires registration and agreement to terms of use
  - No charge
- **Restricted-Use Data**
  - Contract required with data security plan
  - Principal Investigator must have a PhD or equivalent and hold a faculty appointment or research position
  - Annual reports
  - Institutional IRB approval required

# Application Process: Restricted-Use Data

1. Identify and select specific files/components needed
2. Fill out and submit contract application (in pieces or as a whole) via the CPC Data Portal
3. Obtain institutional signatures
4. Fill out the form for data security plan
5. Access data via the UNC Secure Research Workspace (SRW)

- See details here and here

# The UNC SRW

- Alternative to traditional restricted data contracts
- Access through secure, virtual environment
- Data never leaves secure server
- Pre-installed statistical software
- Input and output review by Add Health staff
- No need for physical secure room

# Outline


1. Study Design and Sampling
2. Datasets and Variables
3. Data Access
4. Special Features: Genetic Data
5. Resources

# Genetic Data Collection

- Wave IV
- Saliva-based DNA samples from consented participants
- Genetic data
  - Genome-wide genotyping and quality controls at the individual and genetic variant levels (9,974 individuals and 609130 genetic variants)
  - Imputation to $\sim$ 18 million genetic variants
- Sample composition
  - Twin pairs, full siblings, half siblings
  - Unrelated individuals

# Polygenic Scores (PGSs)

- Weighted sums of genetic variants associated with traits
- PGS $= \sum_{i=1}^{N} w_i \cdot g_i$
- $w_i$: effect size, and $g_i$: the count of the reference allele (0, 1, or 2)
- $w_i$ from publicly available genome-wide association studies (GWASs)
- Aggregate measure of genetic predisposition

# Available PGSs in Add Health

- You don't need to construct them by yourself
- Available from the Social Science Genetic Association Consortium (SSGAC)
- Educational attainment, cognitive performance, height, BMI, obesity, depression, ADHD, schizophrenia, and many others

# Resources for Researchers

- Add Health has several virtual workshops and videos available on YouTube

# Thank you!

Contact Information:
luyin.zhang@princeton.edu