Overview of the National Longitudinal Study of Adolescent Health (Add Health) Data Files

Hsueh-Sheng Wu

Center for Family and Demographic Research

November 28, 2011

Outline

- Introduction
- What is special about Add Health?
- Survey design
- Subject areas
- Data files
- Unit of analysis
- Analytic tips
- Studies using Add Health data
- Help with Add Health analyses
- Conclusions

Introduction

- National Longitudinal Study of Adolescent Health (Add Health) is a study that the Carolina Population Center of University of North Carolina (UNC) has conducted to follow a nationally representative sample of adolescents in grades 7-12 since1994.

- These adolescents were first interviewed in 1994-1995 (Wave I), followed up in 1996 (Wave II), interviewed again in 2001-2002 (Wave III), and last interviewed in 2007-2008 (Wave IV). All of these four waves of data have been released for research analysis.

- In Fall 2001, Population Research Center of University of Texas at Austin collected data that supplement Add Health. These supplementary data focus on (1) educational achievement, (2) course taking patterns, (3) curricular exposure, and (4) educational contexts of Add Health respondents at Wave III.

What Is Special about Add Health?

- Has a large sample that represent adolescents in grades 7 through 12 of the United States in 1994.
- Collect comprehensive information on biological and psychological developments of adolescents and the social contexts, such as home, friends, intimate relations, schools and neighborhood.
- Provide important data on a life stage where adolescents transit into adulthood.

Survey Design

Add Health:

- Add Health used stratified two-stage sampling methods:
  - The sampling frames is stratified by region, urbanization, school size, school type, and race composition
  - 80 high schools and 52 middle schools were selected with unequal probability at the first stage
  - 90,000 students were selected to fill out in-school Add Health questionnaire, and 27,000 of them fill out in-home questionnaire
- Add Health oversampled twins and siblings of twins; non-related adolescents residing together; disabled minority students; blacks from well-educated families; and minority students who are Chinese, Cubans, and Puerto Ricans.
- Data were collected with Computer-assisted personal interviewing (CAPI) and questionnaire

Survey Design (Cont.)

National Longitudinal Study of Adolescent Health (Add Health): Education Data

- The study sample consists of all respondents at Wave III of Add Health
- The study collected high school transcripts and other data from high schools that Add Health respondents last attended
- The data were collected from 130 Add Health high schools and 1,400 additional high schools
- Education data were collected for approximately 12,000 respondents, which is abut 80% of Add Health respondents at Wave III.

Subject Areas

- Add Health covers many interesting subject areas.
- Some of the areas have been covered at all every wave, while others were covered at only certain wave or waves.
- The excel file (i.e., subject areas.xls) summarize the subject areas covered by the in-home interview at each wave of Add Health survey.

Data Files

- CFDR stores a copy of public data in the public folder (R:\CFDR\Public\Data\AddHealth). In addition, public data can be downloaded from ICPSR website (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600).

- CFDR stores a copy of restricted data on the secured server (R:\AddHealth). Only people who have obtained permission from Carolina Population Center of University of North Carolina at Chapel Hill can access the data.

- The difference between the public data and the restricted data is that public data contain about only one third of observations, while restricted data have all of observations.

Data Files (Cont.)

- Add Health: Education Data
  - All education data are restricted data
  - CFDR stores a copy of restricted data on the secured server (R:\AddHealth). Only people who have obtained permission from Carolina Population Center of University of North Carolina at Chapel Hill can access the data

- the public data files are in the folder (R:\CFDR\Public\Data\AddHealth).

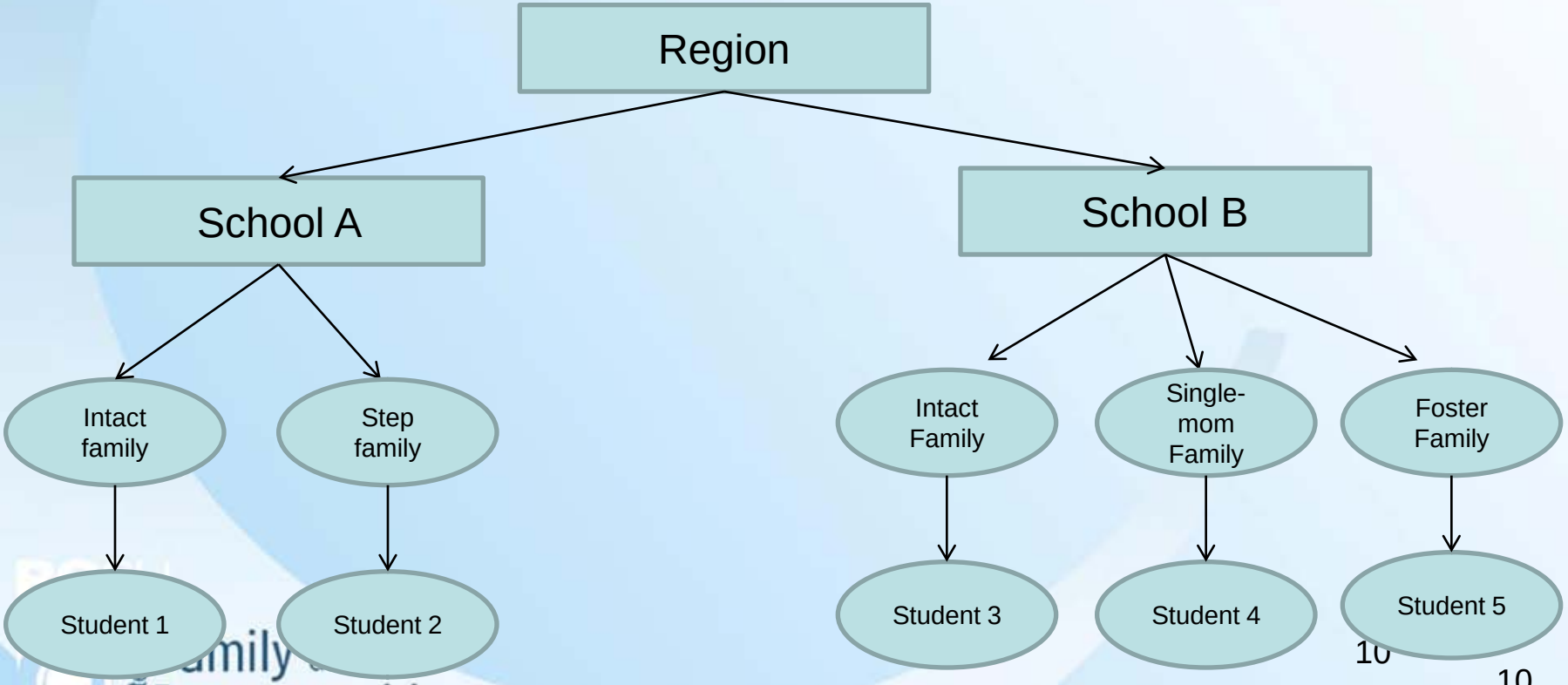
- The locations of the restricted data files are shown in the excel file (R:\AddHealth\ADD Health\Summary of Add Health and AHAA files.xls).

- CFDR has constructed some SAS data sets from the restricted Add Health data, including the in-home interview data, weighted data, and family structure measures from Wave I through IV. The locations of these constructed data are shown in the excel file (R:\AddHealth\ADD Health\Add Health study\CFDR SAS data\CFDR SAS data sets.xls).

- If you need to use restricted Add Health data, please contact Krista Payne (kristaw@bgsu.edu) for the form and procedure of getting permission.

Unit of Analysis

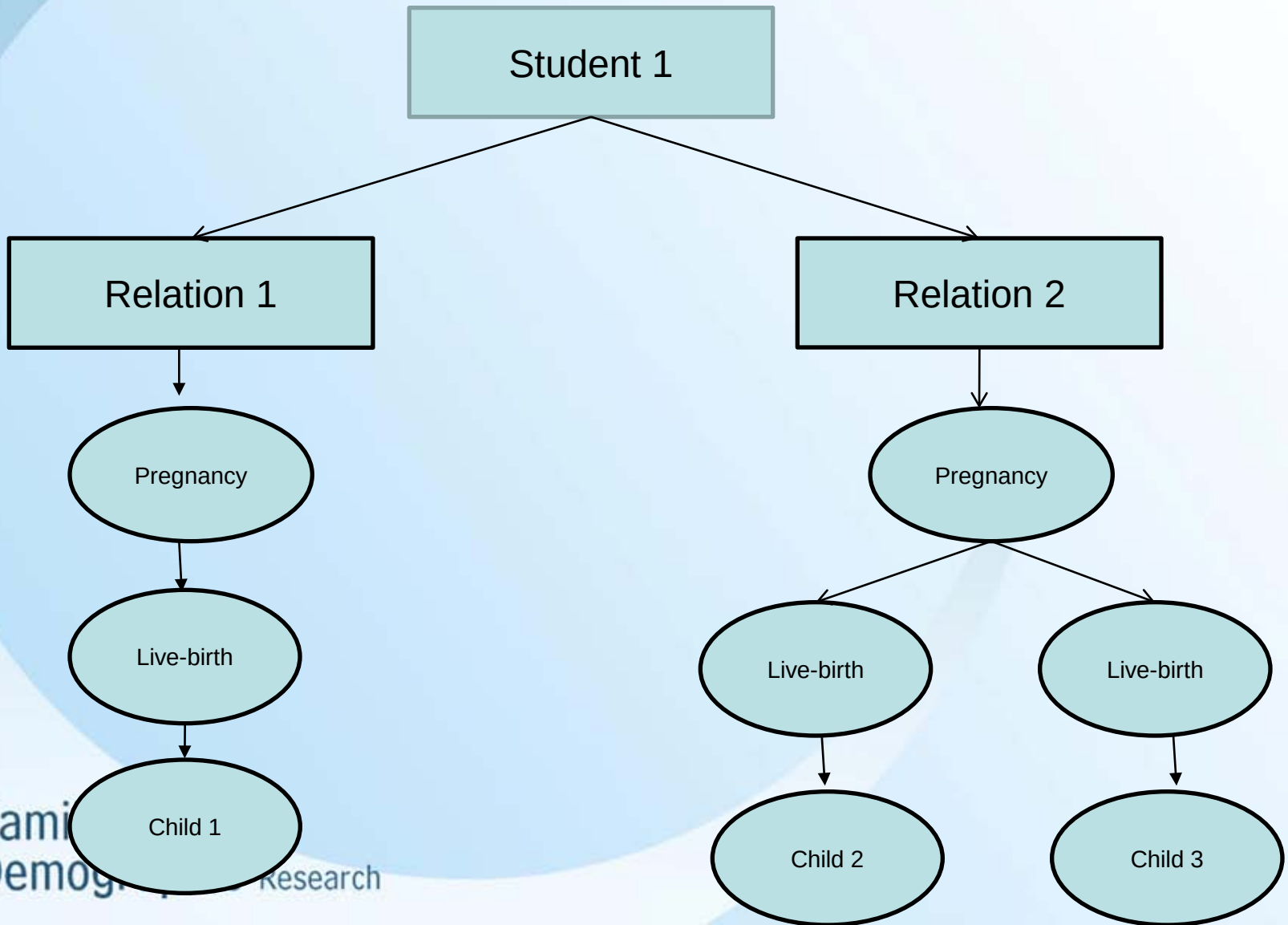
- Add Health data collect information on individual adolescents, their social environment (e.g., neighborhood, school, family) and various aspects of social relations and experiences (e.g., intimate relation, pregnancy, live births, and parent-children relation).
- An example of the nested structure of Neighborhood, School, Family, and Individual Adolescents



10

Unit of Analysis (Cont.)

- An example of the nested structure of Individual, relation, pregnancy, live births, and parent-children relation.



Analytic Tips

- How to find the variables you need?
- How to read the Add Health data?
- How to merge data?
- How to weight Add Health data?
- How to change the unit of analysis?

How to Find the Variables You Need?

- You need to use codebooks to locate the variables of interest.
- Add Health provides codebook that list all of the name and wordings of the variables at each wave. Thus, if you are interested in the in-home interview data, you should start finding your variables by search with the following codebooks:
  - WAVE1NDX.PDF
  - WAVE2NDX.PDF
  - wave3ndx.pdf
  - wave4ndx.pdf
  - Each subject area usually has its own codebook, and you can only find value labels in each codebook.

How to Read Add Health Data?

- The public data of Add Health may be in SAS, Stata, or SPSS format. You can use Stat/transfer to change the data from one format to another.

- The restricted data of Add Health are initially in SAS export format. The following commands provides instruction to use SAS and Stata to read in the SAS export file.

How to Read Add Health Data? (Cont.)

SAS command:

```
LIBNAME wave1 xport "T:\ADD Health\Add Health study\data and related
documents\in-home interview at Waves I, II, and III\in-home interview
at Wave I\data\allwave1.exp";

LIBNAME out "T:\Temp";

DATA out.wave1;
SET wave1.allwave1;
RUN;

PROC CONTENTS DATA = out.wave1;
RUN;
```

Stata command:

```
fdause "T:\ADD Health\Add Health study\data and related
documents\in-home interview at Waves I, II, and III\in-home
interview at Wave I\data\allwave1.exp“
describe
```

How to Merge Data?

- When do data need to be merged?
  - If you want to combine data from different waves of Add Health
  - If you want to combine data with different unit of measurements
  - If you want to use both Add Health data and Education data

- SAS and Stata sample commands to merge Waves I and II data

How to Merge Data? (Cont.)

- An example of SAS commands

```
Libname in "R:\AddHealth";
************************;
PROC SORT DATA=in.wave1;
BY aid;
RUN;
************************;
PROC SORT DATA=in.wave2;
BY aid;
RUN;
************************;
DATA  in.wave12;
MERGE in.wave1 (IN=in_wave1) in.wave2 (IN=in_wave2);
BY aid;
RUN;
```

How to Merge Data? (Cont.)

- An example of Stata Command

```
use "R:\AddHealth\wave1.dta"
sort aid
save "R:\AddHealth\wave1_2.dta", replace
********************************
use "R:\AddHealth\wave2.dta"
sort aid
save "R:\AddHealth\wave2_2.dta", replace
********************************
use "R:\AddHealth\wave1_2.dta", clear
sort aid
merge aid using "R:\AddHealth\wave2_2.dta
tab1 _merge
rename _merge wave12
label variable wave12 "indicator for merging waves 1 and 2"
sort aid
save "R:\AddHealth\wave12.dta", replace
```

How to Adjust for Design Effect?

- Add Health data were collected with a complex survey design. Therefore, each respondent does not have the same probability of being selected into the sample and needs to be reweighted.
- Clustering of students from the same regions and schools.
- The analysis of Add Health data always need to be weighted in order to adjust for the effects of its complex survey design
- SAS and Stata can differ in their abilities of performing statistical analyses, while controlling for the effects of the complex survey design.

How to Weight Add Health Data? (Cont.)

Table 3. Select Stata and SAS procedures for Analyzing Survey Data

Analysis	Stata command	SAS command
Estimate means for survey data	svy: mean	Proc Surveymeans
Estimate proportions for survey data	svy: tab	Proc Surveyfreq
Linear regression for survey data	svy: regress	Proc Surveyreg
Logistic regression for survey data, reporting odds ratios	svy: logistic	Proc Surveylogistic
Cox proportional hazards model for survey data	svy: stcox	Proc Surveyphreg
Ordered logistic regression for survey data	svy: ologit	
Ordered probit regression for survey data	svy: oprobit	
Multinomial (polytomous) logistic regression for survey data	svy: mlogit	
Multinomial probit regression for survey data	svy: mprobit	
Parametric survival models for survey data	svy: streg	
Generalized linear models for survey data	svy: glm	
Generalized negative binomial regression for survey data	svy: gnbreg	
Poisson regression for survey data	svy: poisson	
Zero-inflated negative binomial regression for survey data	svy: zinb	
Zero-inflated Poisson regression for survey data	svy: zip	

How to Weight Add Health Data? (Cont.)

Typically, you will have some missing cases in your data sets and need to analyze part of the sample. It is hard to use SAS to control design effects and conduct sub-population at the same time. Thus, we recommend using Stata for such analyses

Sample Stata commands for analyzing the full sample and the partial sample:

```
use "R:\AddHealth\TEMP\logit3.dta", clear
svyset psuscid3 [pweight =gswgt3]
strata(region3)
svy: logit h3ed3 bio_sex3 calcage3
```

```
use "R:\AddHealth\TEMP\logit3.dta", clear
svyset psuscid3 [pweight =gswgt3], strata(region3)
svy, subpop(marker): logit h3ed3 bio_sex3 calcage3
```

Center for Family and Demographic Research

How to Change the Unit of Analysis?

- Changing the unit of analysis means changing the unit of observations in the data set. Because of the nested structure of Add Health data, you can change the unit of observations from one level to another.
- Provide link to the words document on the linking

- When the unit of analysis change, the number of valid observation changes, too.

Table 2. The number of Units at Different Levels of Analysis for Section 25 of the Wave III of Add Health

Unit of Analysis	Number of Analysis Units
CHILD	4,181
BIRTH	4,181
PREGNANCY	4,055
RELATION	3,293
RESPONDENT	2,960

- SAS and Stata examples of changing unit from birth to pregnancy

How to Change the Unit of Analysis? (Cont.)

- SAS commands

```
PROC SORT DATA = out.sect25 OUT=out.preg;
BY aid rrelno rpregno;
RUN;

PROC FREQ DATA = out.preg;
TABLES birthno;
RUN;

PROC TRANSPOSE DATA =out.preg
OUT=out.f_preg PREFIX =birth;
BY aid rrelno rpregno;
ID birthno;
VAR c_age;
RUN;

PROC CONTENTS DATA = out.sect25;
RUN;

PROC CONTENTS DATA = out.f_preg;
RUN;
```

How to Change the Unit of Analysis? (Cont.)

- Stata commands

```
use “R:\AddHealth\TEMP\sect25.dta”, clear

des  aid rrelno rpregno

sum rrelno rpregno

tostring  rrelno, generate(srrelno)
tostring  rpregno,generate(srpregno)

gen said_rp3 =  aid + srrelno + srpregno
replace said_rp3 =  aid + "0" + srrelno + srpregno if rrelno >=1 & rrelno <=9

label variable said_rp3 "string id for pregnancy record"

sort said_rp3

des

tab1 birthno

reshape wide c_age, i(said_rp3) j(birthno)

des

rename c_age1 birth1
rename c_age2 birth2
save " R:\AddHealth\TEMP\pregnancy.dta", replace
```

Studies Using Add Health Data

- There have been more than 4,000 publications using Add Health data. You can locate them through Add Health ICPSR web sites.
  - Add Health Web site
    - http://www.cpc.unc.edu/projects/addhealth/pubs
  - ICPSR website:
    - After you find Add Health data, click on the "View related literature"

Help with Add Health Analyses

- CFDR Add Health Working Group

  CFDR working group provides a forum for Add Health users at BGSU to present their findings and obtain feedback from group members. This group is organized and supervised by an experienced Add Health user, Dr. Kara Joyner. If you are interested in joining the working group, please contact her at kjoyner@bgsu.edu.

- Official Add Health listserve

  Listserve is a place where Add Health users ask and answer questions about analyzing Add Health data. To subscribe to the official Add Health listserv, send e-mail to: listserv@unc.edu and in the body of the message put: subscribe addhealth2 <firstname lastname> .

- Add Health Users Conference

  Carolina Population Center of University of North Carolina has hosted nine Add Health Users Conference on how to construct and analyze Add Health data. Add Health website (http://www.cpc.unc.edu/projects/addhealth/news) will provide information on the upcoming Add Health User Conference.

  •CFDR Programming Help

  If you have programming problems, please contact Hsueh-Sheng Wu at whu@bgsu.edu.

Conclusions

- Add Health is an excellent data set for studying how adolescents make transition into adulthood.
- The construction of Add Health can be difficult because it may involve using data collected from different measurement units and at different waves
- The analysis of Add Health data always need to be weighted in order to adjust for the effects of its complex survey design.
- Given the excellence of the data set, many interesting studies can be done using Add Health.