

**Report prepared by**

Kathleen Mullan Harris
Carolyn T. Halpern
John M. Hussey
Eric A. Whitsel
Robert A. Hummer
John R. Knapp

Add Health Sample Member Birth Records Database User Guide



CAROLINA POPULATION CENTER | CAROLINA SQUARE - SUITE 210 | 123 WEST FRANKLIN STREET | CHAPEL HILL, NC 27516

Waves I-V of Add Health were funded by grant P01 HD31921 (Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health is currently directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and Hummer) at the University of North Carolina at Chapel Hill. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill.

<https://doi.org/10.17615/gp33-z087>

Acknowledgments

The Add Health data and material use agreement requires that the following be included in each written report or other publication based on analysis of this data:

The Office of Vital Statistics from the Ohio Department of Health (ODH) provided some data used in this study. Use of these data does not imply ODH agrees or disagrees with any presentations, analyses, interpretations, or conclusions. Information on how to obtain the Add Health data files is available on the Add Health website (<http://www.cpc.unc.edu/addhealth>).

Suggested Citation

Citations of this Add Health User Guide should use the following format:

Harris, K.M., Halpern, C.T., Hussey, J.M., Whitse, E.A., Hummer, R.A., and Knapp, J.R. 2024. Add Health Sample Member Birth Records Database. The National Longitudinal Study of Adolescent to Adult Health. Carolina Population Center, University of North Carolina at Chapel Hill.
<https://doi.org/10.17615/gp33-z087>

Introduction

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a longitudinal study of a nationally representative sample of over 20,000 adolescents who were in grades 7-12 during the 1994-95 school year, and have been followed for five waves to date, most recently in 2016-18. Data collection for Wave VI is currently in progress. See Harris et al. (2019) for more information about the study.

The Add Health Sample Member Birth Records Database provides data from the birth records of a subset (N = 2,750) of Add Health sample members (AHSMs). The AHSMs were born between 1974 and 1983. The Birth Records Database is composed of AHSMs who consented to Add Health's use of their birth record data and who were born in states that have agreed to make birth record data available to Add Health.

Data Structure and Form

Birth record data was collected from participating states for AHSM birth years, 1974-83. When these states provided birth data for all recorded births occurring during that time interval, an AHSM-specific subset was created using Link Plus, a statistical linkage software developed by the U.S. Centers for Disease Control and Prevention (CDC), Cancer Division. One participating state performed its own AHSM linkages and provided Add Health with the linked subset of births. Add Health then performed transformations on all of the original data from the participating states to create the categorical variables present in this release.

There were challenges with transforming the original source data into categorical variables. First, each state's method for collecting and reporting birth data not only varied from one another, but also at times

varied within states across years. As a result of these reporting differences, constructed variables in this database may be either direct transformations of source data single variables or composite variables constructed using two or more unique distinct source variables.

While the challenges described above affected all the variables in this data set, best examples are RIWA002 and RIWA003 (calculated and obstetric estimates of gestational age, respectively). Over one-half of the records in this data set originate from states where one or both gestational age measurements were not reported or were unable to be calculated. And the subjective and variable nature of certain measurements, such as the mother's reported date of last normal menses (DLNM) used in calculated gestational age, as well as the varying nature of obstetric estimations of newborns' gestational age during that time period, likely affects the overall quality of the data in these variables.

It should also be noted that the data in the original vital records data used to calculate mother's parity represented parity *prior to the birth event of the AHSM*. In the released Add Health data, the decision was made to report the mother's parity *including* the AHSM's birth in the total (i.e. by adding +1 across the board to all the original parity data).

Second, states differed in what they would allow for data release by Add Health. The state-specific regulations drove the decision to constrain *all* available data to the strictest reporting requirements of one specific state.

These guidelines served as the basis for overall reporting requirements and ultimately informed the Add Health variable construction and reporting formats. In some instances, birth record data from one state may have been available, but if the most restrictive state guidelines limited its version of the data, those cases were assigned a reserve code (described in the Reserve Codes section of this guide).

Also, if another state's data could not be transformed into the most restrictive state's reporting format, a reserve code was assigned. For example, although the most restrictive state reported a specific data measurement, another state did not report the same measurement, or their method of collecting and reporting a similar measurement was in a format untranslatable to the most restrictive state's coding requirement. This is further detailed in the descriptions for reserve codes 95 and 96 below.

Before release, Add Health staff performed a deductive disclosure mitigation review to protect participant confidentiality. While every effort was made to minimize the impact on the analytical utility of the data, in a limited number of cases, a small selection of the original vital records data may have been manipulated by Add Health to protect confidentiality.

Data Limitations

The most significant limitation of these data is variation across states in the years the source data were available. In the AHSM population, the range of all possible birth years was from 1974 to 1983, and all source data focused specifically on that range of years. However, each state varied in the number of overall years of data provided within this range of birth years. Specifically, some states provided data that started later than 1974. Because Add Health data security policy does not disclose information related to the specific birth dates of individual respondents or their states of birth, the distributions of birth years affected by non-reported or unavailable data cannot be inferred from these data.

Second, it is important to note that the AHSM birth records were only available from a subset of

states and thus only a subset of the Add Health respondents. If a variable from the birth records data is used as an outcome variable, we recommend a careful analysis of missing data and use of the grand sample weight at Wave V (GSW5) in analysis.

In addition to the gestational age variable challenges described above, calculated gestational age for the two gestational age variables often relied on incomplete data. Thus, we defaulted to a day value of 15 when month and year values were reported but day values were missing or coded as unknown. Also, the conversion of time intervals and use of the “Continuous Method” of the SAS INTCK function may have contributed to variability between gestational age values across the two measures of gestational age. Cross-tabulation showed that in 90% of cases, response codes either matched or differed by only one to two codes. Overall, we recommend careful use of the gestational age variables; it's possible that combining them together may result in the best estimate of gestational age for each particular birth.

Data Dictionary

Variable naming schema is summarized in Table 1 below. The first character “R” is constant throughout all the data set variables. Characters 2–3 are used to group thematically similar variables together. The four groupings are shown in the table below. Character 4 is used to differentiate versions of the same variable. In this release only one version of each variable exists so character 4 has a value A in all instances. Characters 5-7 indicate a variable number.

Table 1. Variable Naming Schema		
Character Position	Character	Description
1	R	Respondent
2 - 3	IW	Infant Wellbeing
	ID	Infant Demographics
	PD	Parental demographics
	MM	Maternal Medical Risk Factors
4	A	Version 1
	B	Version 2
	C	Version 3
5 - 7	001-008	Number

A general outline of all the variables appearing in this data set along with their definitions are listed in Table 2 below. Detailed information including frequencies of individual variable codes appear in the accompanying codebook.

Table 2. Variable Definitions		
#	Name	Description
Infant Wellbeing–IW		
1	RIWA001	Birth weight, categorical, in grams
2	RIWA002	Gestational age at birth based on last menstrual period, completed weeks

3	RIWA003	Gestational age at birth based on obstetric estimate of gestation at delivery, completed weeks
4	RIWA004	Method of delivery
Infant Demographics–ID		
5	RIDA001	Parity
6	RIDA002	Plural Birth
Parental demographics–PD		
7	RPDA001	Mother’s Race
8	RPDA002	Father’s Race
9	RPDA004	Mother’s nativity
10	RPDA005	Mother’s age at birth
11	RPDA006	Father’s age at birth
12	RPDA007	Mother’s level of education at time of birth
13	RPDA008	Father’s level of education at time of birth
Maternal Medical Risk Factors–MM		
14	RMMA001	Previous Live Births Now Deceased

Reserve Codes

Codes for missing values always begin with “9” and end with either “5”, “7”, or “8.” The interior is padded with enough additional “9”s to make the length exceed by one the maximum value of the variable.

The following table illustrates the convention. Detailed breakdowns of missing codes and their occurrence frequencies can be found in the accompanying codebook.

Code	Definition
95, 995, 9995	No Basis for Calculation
96, 996, 9996	Missing in Source Data
98, 998, 9998	Unknown

Reserve code 95, "No Basis for Calculation," was used when there was some form of available relevant data present in the source data sets, *however*, the data were reported in a format that could not be accurately transformed into the accepted Add Health categories. Code 95 was also used in instances where a source data set had otherwise valid data, but some piece of relevant data required was missing from the record.

Reserve code 96, "Missing in the Source Data," was used to indicate instances in which data were missing from a source data set. Note, "Missing data" in this instance can represent two possible scenarios:

- 1) Source data had a relevant variable that could be transformed into the final variable categories, *but* the individual record in the source data was missing.
- 2) Source data contained no variables relevant to Add Health variable in question. (e.g., "Method of Delivery," RIWA004, if a particular state's source data simply did not report delivery method, code 96 would be used for that collection of records).

Reserve code 98, "Unknown," was used in instances where a valid variable existed in the source data, but an individual record contained some form of special missing or reserve code indicating 'unknown' (e.g., '9999', '-', etc.) for that record's response.

References

Harris, Kathleen Mullan; Halpern, Carolyn Tucker; Whitset, Eric A.; Hussey, Jon M.; Killeya-Jones, Ley A.; Tabor, Joyce; & Dean, Sarah C. (2019). Cohort profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *International Journal of Epidemiology*. <https://doi.org/10.1093/ije/dvz115>