CSDE 502 2021 Winter **Assignment 6** ***Add Health data; Variable creation*** **Instructor: Phil Hurvitz** **phurvitz@uw.edu**	**My Name: answer key** **My UWNetID:**

Due Date: 2021-02-18 09:00 AM

Instructions:

1. Fill in your name and UWNetID above.
2. Put answers to the questions on this document, using the “00Answers” Word style so your answers are clearly distinguished from the questions.
3. Create a PDF file from this document.
4. Create a **single zip** file including this document as a PDF file, along with the RDS file and R code file.
5. Upload the **single zip file** to Canvas.

Explanation:

For this assignment, you will be perusing some of the documentation for the Add Health Wave 1 data set. You will use the documentation to make some updates to a data frame containing some of the Add Health data, and then save the data frame as an RDS file. You will update a metadata table that partially describes the data set and changes you made to the variable names and variable labels.

To open a Stata version 13 file in R there are two main options:

1. Use `haven::read_dta()`. To access variable labels in R use `labelled::foreign_to_labelled()`. To update variable labels, use the `labelled::var_label()` function.
2. Use `readstata13::read.dta13()`. Variable labels for this format are available, e.g., for a data frame named `dat` as `attributes(dat)$var.labels`. This is a vector of text strings that can be updated by assigning a new value to the specified element, e.g., `attributes(dat)$var.labels[1] <- “foo”`.

To save the RDS file, use the base function `saveRDS()`.

Here is a base R code snippet that will rename a single variable:

```
colnames(data_frame)[grep("^original_variable_name$", colnames(data_frame))]
<- new_variable_name
```

The `grep()` function finds the position of the named variable in the list of variables in the data frame. The characters `^` and `$` are regular expressions to specify the start and end of the string to be matched (assuring that the pattern does not match multiple similar variable names).

It is much simpler with tidyverse and magrittr:

```
data_frame %<>% rename(new_variable_name = old_variable_name)
```

Additional hint for dealing with PDF documentation:

1. Use `pdfgrep` (should be available in a Linux or Mac package manager; for Windows, search for a version or use Cygwin).
2. Use the R `pdftools` package. This could be used in a loop over each PDF file to create a data frame with the name of the PDF file, page number, and text of each page. The str_match() function could be used to identify the file name and page number where specific text strings occur. For a minimal example, this shows that the string "h1gi1m" is found on page 1 of INH01PUB.PDF. Conversion of the PDF file's text to lowercase simplifies the matching:

```
> x <- pdftools::pdf_text(pdf = "INH01PUB.PDF")
> str_match(string = x %>% str_to_lower(), pattern = "h1gi1m")
      [,1]    
 [1,] "h1gi1m"
 [2,] NA      
 [3,] NA      
 [4,] NA      
 [5,] NA      
 [6,] NA      
 [7,] NA      
 [8,] NA      
 [9,] NA      
[10,] NA      
[11,] NA      
[12,] NA      
[13,] NA      
[14,] NA      
[15,] NA      
```

Questions:

1. Explore the Add Health website (http://www.cpc.unc.edu/projects/addhealth) and answer the following questions (making sure to cite as necessary):
   1.1. What was the sampling frame for this study?

The sampling frame was a subset n = 80 of 26,666 high schools from the Quality Education Database.

As reported in "The Add Health Study: Design and Accomplishments" (https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/DesignPaperWave_I-IV.pdf):

> **"Schools as Primary Sampling Units**
>
> **Add Health used a school-based design. The primary sampling frame was derived from the Quality Education Database (QED). From this frame we selected a stratified sample of 80 high schools (defined as schools with an 11th grade and more than 30 students) with**

> **probability proportional to size. Schools were stratified by region, urbanicity, school type (public, private, parochial), ethnic mix, and size."**

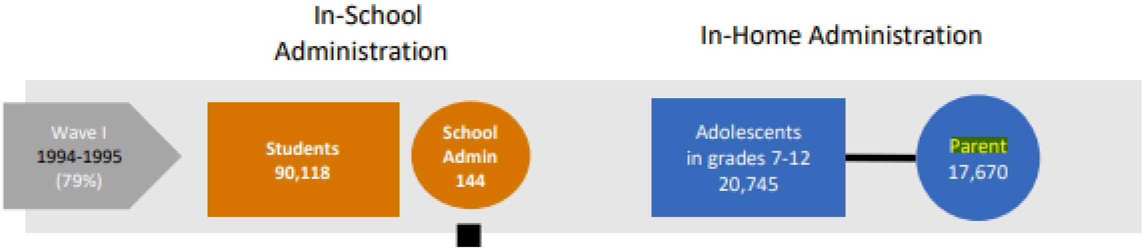
Also in "Guidelines for Analyzing Add Health Data" (https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/GuidelinesforAnalysisofAddHealthData_202004.pdf),

> **"The primary sampling frame was derived from the Quality Education Database (QED) comprised of 26,666 U. S. High Schools. From this frame we selected a stratified sample of 80 high schools (defined as schools with an 11th grade and more than 30 students) with probability of selection proportional to school size."**

1.2. What were the three kinds of respondents at Wave I?

The Wave I respondents consisted of students, school administrators, and parents.

See the main documentation, "The Add Health Study: Design and Accomplishments" (http://www.cpc.unc.edu/projects/addhealth/design/researchdesign.pdf) and the overview "Add Health Research Design" slides (https://addhealth.cpc.unc.edu/wp-content/uploads/docs/documentations/2020_Study_Design.pdf)



1.3. What was the instrument with the largest sample size?

There were 90,118 adolescent In-School Questionnaires.

1.4. Is it possible for a respondent to be in Wave III without being in Wave II?

Yes, see https://addhealth.cpc.unc.edu/about/#studies-satellite:

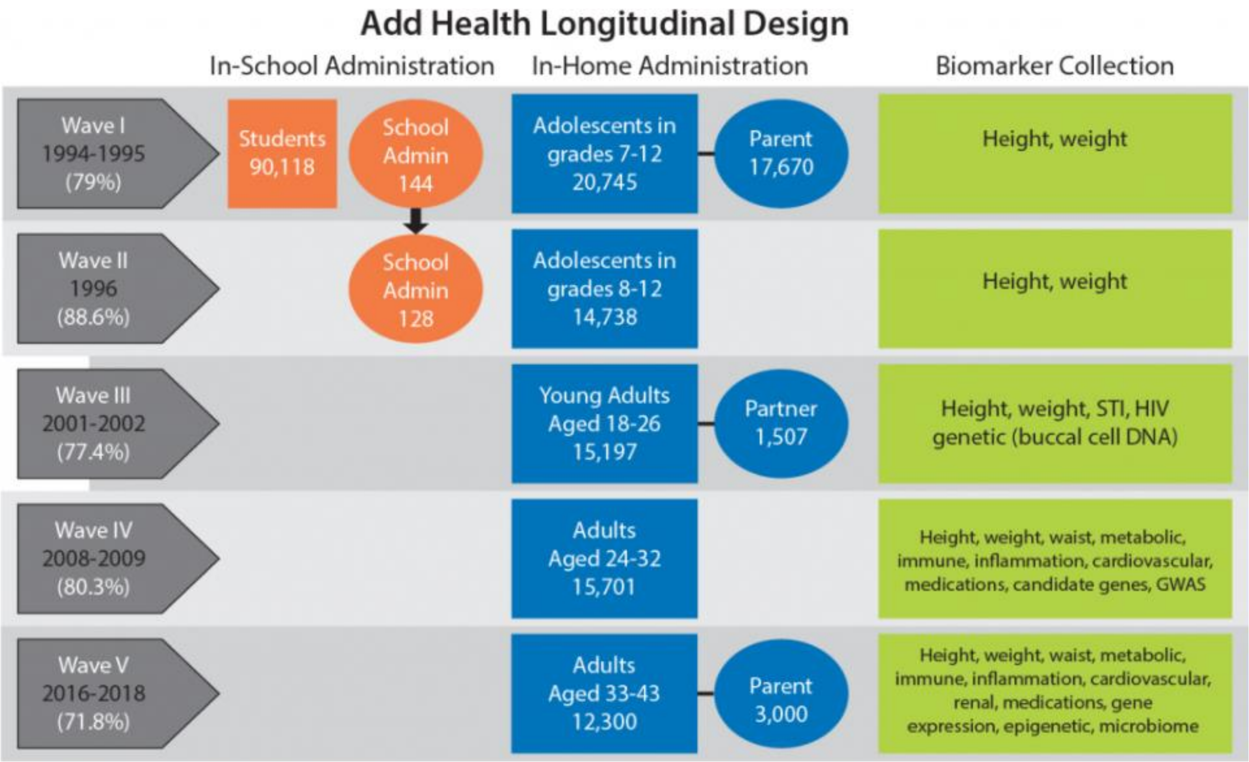
> **Wave I**
>
> **Wave I took place between 1994 and 1995, during which 90,118 students from 145 middle, junior, and high schools completed a 45-minute questionnaire administrated in the school.**
>
> **Wave II**
>
> **During Wave II nearly 15,000 of the Wave I respondents were interviewed from April to August 1996, one year after the Wave I interview.**
>
> **Wave III**

Wave III was conducted from 2001 to 2002, interviewing 15,170 of the Wave I respondents.

Because the count of respondents in Wave III was greater than in Wave II, the Wave III sample logically had to contain some respondents who were not interviewed in Wave II.

1.5. What is the time span of the Add Health data collection (all waves)?

Waves 1 thru 4 were collected between September 1994 to February 2009
Wave 5: 2016-2018



Add Health Longitudinal Design

	In-School Administration		In-Home Administration		Biomarker Collection
Wave I 1994-1995 (79%)	Students 90,118	School Admin 144	Adolescents in grades 7-12 20,745	Parent 17,670	Height, weight
Wave II 1996 (88.6%)		School Admin 128	Adolescents in grades 8-12 14,738		Height, weight
Wave III 2001-2002 (77.4%)			Young Adults Aged 18-26 15,197	Partner 1,507	Height, weight, STI, HIV genetic (buccal cell DNA)
Wave IV 2008-2009 (80.3%)			Adults Aged 24-32 15,701		Height, weight, waist, metabolic, immune, inflammation, cardiovascular, medications, candidate genes, GWAS
Wave V 2016-2018 (71.8%)			Adults Aged 33-43 12,300	Parent 3,000	Height, weight, waist, metabolic, immune, inflammation, cardiovascular, renal, medications, gene expression, epigenetic, microbiome

(from https://addhealth.cpc.unc.edu/documentation/study-design/)

1.6. What is the difference between the public and the restricted-use Add Health data?

The public data set is a subset of cases and variables that is free, and has no data security restrictions; the restricted-use data set is more extensive, containing data for all study participants, but requires a contract between the researcher and UNC/CPC. See https://addhealth.cpc.unc.edu/data/#public-use and https://addhealth.cpc.unc.edu/data/#restricted-use.

1.7. Describe a research question that you might be able to answer using the Add Health dataset.

Any relatively well thought out answer will be acceptable. Better answers include named variables as well as hypothesized relationships between or among variables.

2. Download the public-use Add Health documentation at https://canvas.uw.edu/courses/1434040/files. Answer the following questions:

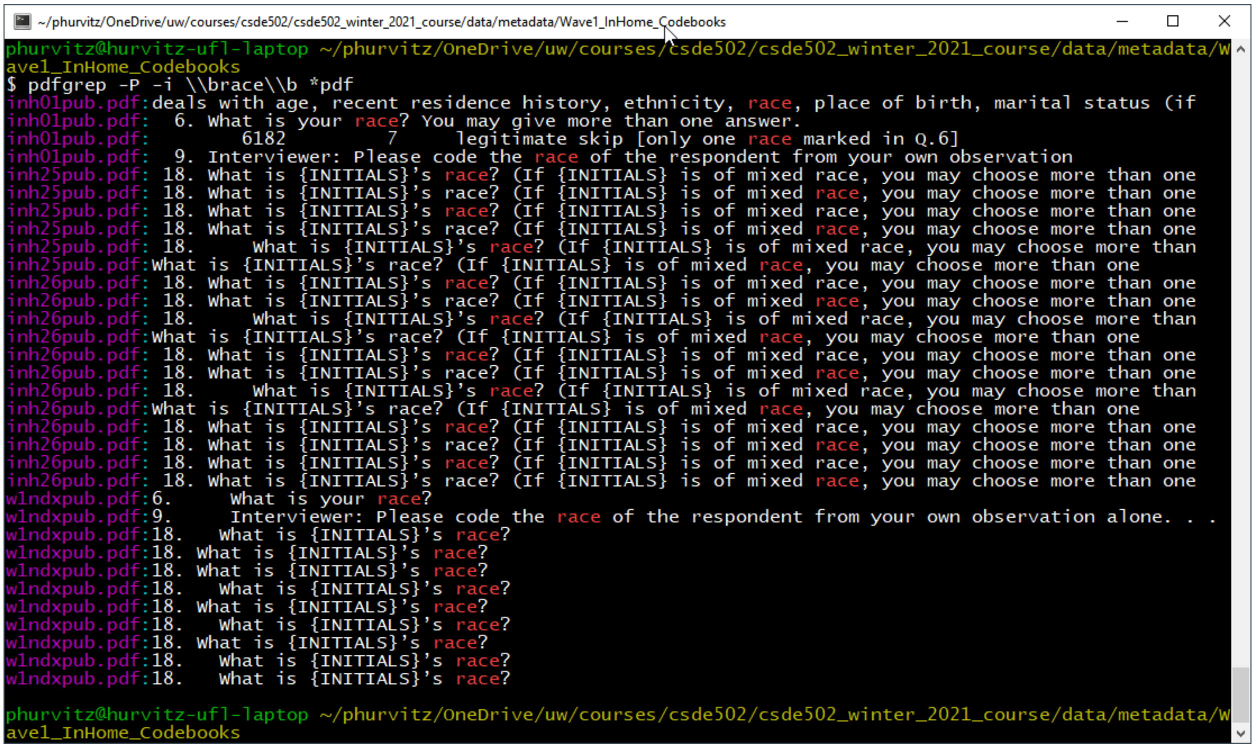
2.1. In what pdf document is the documentation for the race items for the Wave I In-Home questionnaire?

INH01PUB.PDF contains the race questions (see W1NDXPUB.PDF for guide to different sections)

This answer can be helped by using pdfgrep. Under Cygwin, Linux, or Mac as

```
pdfgrep -P -i "\\brace\\b" *.pdf
```

The flag "-P" uses Perl-compatible regular expressions. The "\\b" is used to demark word boundaries, so the *word* "race" is searched (rather than the *string* "race", which would also match the word "contraception" and "braces").


~/phurvitz/OneDrive/uw/courses/csde502/csde502_winter_2021_course/data/metadata/Wave1_InHome_Codebooks

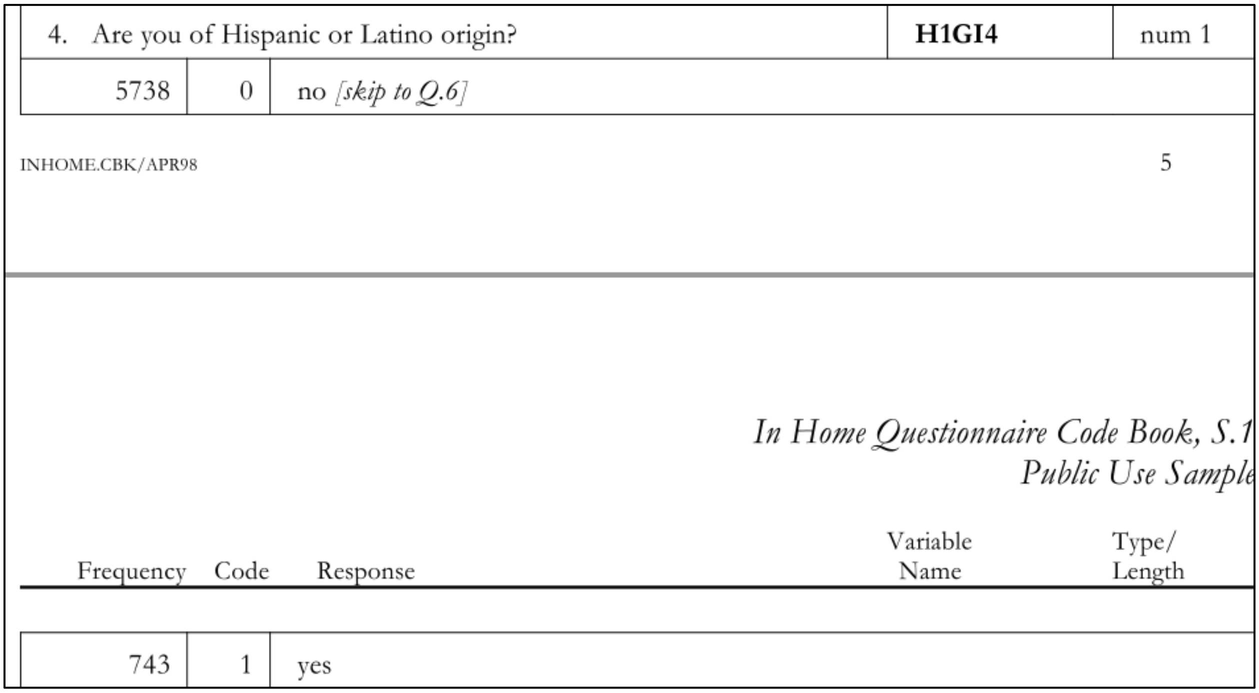
```
phurvitz@hurvitz-ufl-laptop ~/phurvitz/OneDrive/uw/courses/csde502/csde502_winter_2021_course/data/metadata/W
avel_InHome_Codebooks
$ pdfgrep -P -i \\brace\\b *pdf
inh01pub.pdf:deals with age, recent residence history, ethnicity, race, place of birth, marital status (if
inh01pub.pdf:  6. What is your race? You may give more than one answer.
inh01pub.pdf:        6182        7     legitimate skip [only one race marked in Q.6]
inh01pub.pdf:  9. Interviewer: Please code the race of the respondent from your own observation
inh25pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh25pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh25pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh25pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh25pub.pdf: 18.    What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than
inh25pub.pdf:What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18.    What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than
inh26pub.pdf:What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18.    What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than
inh26pub.pdf:What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
inh26pub.pdf: 18. What is {INITIALS}'s race? (If {INITIALS} is of mixed race, you may choose more than one
w1ndxpub.pdf:6.     What is your race?
w1ndxpub.pdf:9.     Interviewer: Please code the race of the respondent from your own observation alone. . .
w1ndxpub.pdf:18.    What is {INITIALS}'s race?
w1ndxpub.pdf:18. What is {INITIALS}'s race?
w1ndxpub.pdf:18. What is {INITIALS}'s race?
w1ndxpub.pdf:18.    What is {INITIALS}'s race?
w1ndxpub.pdf:18. What is {INITIALS}'s race?
w1ndxpub.pdf:18.    What is {INITIALS}'s race?
w1ndxpub.pdf:18. What is {INITIALS}'s race?
w1ndxpub.pdf:18.    What is {INITIALS}'s race?
w1ndxpub.pdf:18.    What is {INITIALS}'s race?

phurvitz@hurvitz-ufl-laptop ~/phurvitz/OneDrive/uw/courses/csde502/csde502_winter_2021_course/data/metadata/W
avel_InHome_Codebooks
```

pdftools::pdf_text() can also be used

2.2. How many respondents were of Hispanic/Latino origin?

There were 743 Hispanic/Latino respondents.
See p 4-5 in inh01pub.pdf"



4. Are you of Hispanic or Latino origin?			H1GI4	num 1
5738	0	no *[skip to Q.6]*		

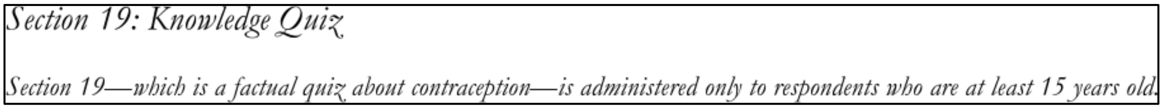
INHOME.CBK/APR98 5

In Home Questionnaire Code Book, S.1
Public Use Sample

Frequency	Code	Response	Variable Name	Type/ Length
743	1	yes		

2.3. What is the "Knowledge Quiz" in the Wave I In-Home questionnaire?

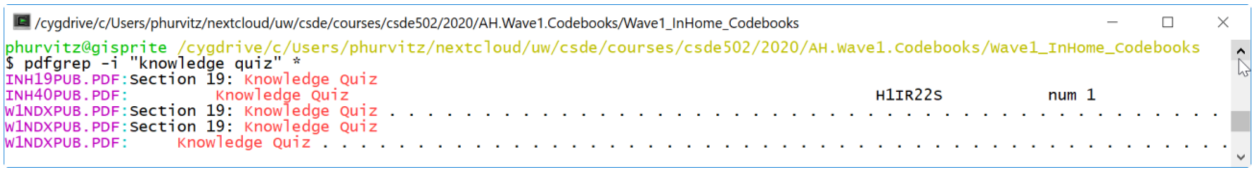
The "Knowledge Quiz" is a factual quiz about contraception, as shown in inh19pub.pdf.



Section 19: Knowledge Quiz

Section 19—which is a factual quiz about contraception—is administered only to respondents who are at least 15 years old.

This question can also be helped by the use of pdfgrep. pdfgrep will find the pattern "knowledge quiz", identifying the PDF containing the text.



```
phurvitz@gisprite /cygdrive/c/Users/phurvitz/nextcloud/uw/csde/courses/csde502/2020/AH.Wave1.Codebooks/Wave1_InHome_Codebooks
$ pdfgrep -i "knowledge quiz" *
INH19PUB.PDF:Section 19: Knowledge Quiz
INH40PUB.PDF:           Knowledge Quiz                                        H1IR22S          num 1
W1NDXPUB.PDF:Section 19: Knowledge Quiz . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
W1NDXPUB.PDF:Section 19: Knowledge Quiz
W1NDXPUB.PDF:      Knowledge Quiz . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

2.4. What is the unique identifier for the In-home data?

"AID" is the unique identifier.

3. Download the Stata 13 format file AHwave1_v1.dta (http://staff.washington.edu/phurvitz/csde502_winter_2021/data/AHwave1_v1.dta).

For answers, see the Rmd and HTML on the Canvas site for assignment 6.

3.1. Fill in the grey missing cells in Table 1 below based on the data and/or documentation. Optimally, use the documentation to familiarize yourself with the structure of the code books.

3.2. Using questions 6 and 8 in INH01PUB.PDF, create a new variable named "race" that uses recoded values (white = 1; black/African American = 2; American Indian = 3; Asian/Pacific Islander = 4; other = 5; unknown/missing = 9).

3.3. Rename the variables, and update variable labels using Table 1 as a guide and save the data frame as the file as **AHwave1_v2.rds**. Use a single R code file for your edits to the data file.

3.4. Update the status in Table 1 as needed.

Table 1: Codebook for variables from Add Health Wave 1 data

Depending on how the variables/columns were treated, the data type could be as in the original data set (integer for most), or as factors, with value labels as indicated.

variable name	original variable name	status*	data type	values	variable label	codebook file name
aid	aid	unchanged	text	8 digit string	unique case (student) identifier	SECTAPUB.PDF
imonth	imonth	unchanged	integer	integer 1, 4 to 12	month interview completed	SECTAPUB.PDF
iday	iday	unchanged	integer or factor†	1-31	day interview completed	SECTAPUB.PDF
iyear	iyear	unchanged	integer or factor	94, 95	year interview completed	SECTAPUB.PDF
bio_sex	bio_sex	unchanged	integer or factor	1 = male 2 = female 6 = refused	interviewer confirmed sex	SECTAPUB.PDF
bmonth	h1gi1m	renamed	integer or factor	1-12 96=refused	birth month	INH01PUB.PDF
byear	h1gi1y	renamed	integer or factor	74 to 83 96=refused	birth year	INH01PUB.PDF
hispanic	h1gi4	renamed	integer or factor	0 = no 1 = yes 6 = refused 8 = don’t know	Hispanic/Latino	INH01PUB.PDF
white	h1gi6a	renamed	integer or factor	0 = not marked 1 = marked 6 = refused 8 = don't know	race white	INH01PUB.PDF
Black	h1gi6b	renamed	integer or factor	0 = not marked 1 = marked 6 = refused 8 = don't know	race black or African American	INH01PUB.PDF
AI	h1gi6c	renamed	integer or factor	0 = not marked 1 = marked 6 = refused 8 = don't know	race American Indian or Native American	INH01PUB.PDF
asian	h1gi6d	renamed	integer or factor	0 = not marked 1 = marked 6 = refused 8 = don't know	race Asian or Pacific Islander	INH01PUB.PDF

variable name	original variable name	status*	data type	values	variable label	codebook file name
raceother	h1gi6e	renamed	integer or factor	0 = not marked 1 = marked 6 = refused 8 = don't know	race other	INH01PUB.PDF
onerace	h1gi8	renamed	integer, factor, character[‡]	1 = white 2 = black 3 = AI 4 = asian 5 = other 6 = refused 7 = legitimate skip 8 = don’t know 9 = not applicable	one category best describes racial background	INH01PUB.PDF
observedrace	h1gi9	renamed	integer, factor, character	1 = white 2 = black 3 = AI 4 = asian 5 = other 6 = refused 8 = don’t know	interviewer observed race	INH01PUB.PDF
health	h1gh1	renamed	integer, factor, character	1 = excellent 2 = very good 3 = good 4 = fair 5 = poor 6 = refused 8 = don’t know	how is your health	INH03PUB.PDF
race	not applicable	derived	integer, factor, character	1 = white 2 = black or African American 3 = American Indian 4 = Asian or Pacific Islander 5 = other 9 = unknown/missing	race recoded as white; black/African American; American Indian; Asian/Pacific Islander; other; unknown/missing	INH01PUB.PDF

*status categories: unchanged, renamed, missing defined, derived

[†] all variables other than AID are stored internally as integers where the values act as codes that correspond to specific text responses. These could also be formatted as “factors”.

[‡] nominal categorical variables could be stored as character strings.