The Long Format Data of the RAND HRS Longitudinal File 2020

Released May 2024 (V2)

Jinkyung Ha, Mohammed Kabeto, and Andrzej Galecki

September 2024

The Design, Data, and Biostatistics Core (DDBC)
Claude D. Pepper Older Americans Independence Center
Division of Geriatric and Palliative Medicine
University of Michigan

This document describes researchers' contribution in converting the wide format of the RAND HRS Longitudinal File (RANDHRS1992_2020V2 dataset) into a small and more manageable long-format dataset. All the data are constructed solely using the publicly available data that RAND created.

As more HRS waves are available in the future, the number of variables that are required in the wide format data will increase significantly, creating difficulty in managing. For example, the RAND HRS longitudinal 2020 data contains data for N=42,405 respondents with 17,653 variables, and each row is uniquely identified by HHID and PN variables. By using the long format, the number of variables can be reduced significantly . Additionally, the data are easily manageable, including the extraction of wave-specific data.

The main advantage of the proposed data organization is that one can quickly extract any specific wave data, including data for longitudinal analysis. In addition, this data organization, which draws on ideas borrowed from relational databases, gives researchers a better insight into data structure. In using the HRS data, having a structured relational database is useful to efficiently work with the dataset, focusing on a specific dataset and a limited number of variables. For example, transforming the data into a long format helped reduce the number of variables to 1,713, yielding a 90.1% reduction. In addition, separate data files, such as demographics with static variables, long format datasets with variant variables, and exit data, were created, thereby providing easily manageable relational-based datasets. We also created long-format data files for spouse and household data. Therefore, having relational-based long-format data is preferable and easily manageable. We kept the structure of the variable names similar to those in the RAND data so that users can easily refer to the RAND document for further information on how each variable was constructed and operationalized.

We converted the RAND dataset into six output data tables (See Table 1). Data tables Rlong_table, Hlong_table, and Slong_table contain wave-specific data for respondent, household, and spouse variables, respectively. Other respondent wave-invariant variables (like demographic characteristics) and variables from Exit Interviews are included in Rwide_table and Rexit_table. We also created a long-formatted RSSI_table file for disability episode variables. All six data tables contain variables that can be used to identify a unique row in the data tables and merge all the datasets (Table 2).Their contents and data dictionary are also provided.

Table 1. Output Data Tables (stored in `data_tables` subfolder).

Output data table	Variables' Description	Variables' conversion	Filter	Observations	Variables
Rlong_table[1]	Respondent R[w] variables	R[w]*var* converted to R_*var*	if INW[w] = 1	280,343	809
Hlong_table[1]	Household H[w] variables	H[w]*var* converted to H_*var*	if INW[w] = 1	280,343	87
Slong_table[1,2]	Spouse S[w] variables	S[w]*var* converted to S_*var*	if INW[w] = 1 & S[w]HHIDPN is not zero[3]	183,965	853
Rwide_table	Wave-invariant `RA` variables	RAND RA*var* are saved		42,405	65
Rexit_table	Exit `RE` variables	RAND RE*var* from Exit Interview are saved	if REXITWV is not missing[4]	16,512	105
RSSI_table	Respondent SSI/SSDI episode variables	Disability episode variables converted to long format (one row per episode)	if RADAPPD[n] is not missing[5]	11,152	20

[1] The output data tables stored in long format contain wave-specific variables from waves where an individual responded (INW[w] = 1). Some variables are not available for certain waves. For example, R[w]ALZHE (R reports Alzheimer this wave) is not available for wave 1 through wave 9. For the corresponding visits, we coded these variables as missing using "_" symbol to distinguish them from other missing responses (Table 2. Missing codes in RAND HRS Longitudinal File 2020 (V2) Documentation).

[2] Three spouse social security wave-invariant variables SASSAGEB, SASSAGEM, SASSRECV are also included in Slong_table.

[3] If there is no spouse in a given wave, S[w]HHIDPN is set to zero. We don't include them in Slong_table. However, when S[w]HHIDPN is unknown, and the marital status is either missing or married, we still retain them.

[4] Not all deceased had exit interviews. The output file contains only those who completed exit interviews (REXITWV is not missing).

[5] RSSI_table only contains the cases where the imputed SAS date of application for episode n (n = 1, 2, …, 11) is not missing.

Table 2. Selected variables in the output data tables.

			Output tables					
RAND variable(s)	**Output variables**	**Variable label**	**Rlong**	**Hlong**	**Slong**	**Rwide**	**Rexit**	**RSSI**
HHID	HHID	HHold ID /6-Char	√	√	√	√	√	√
PN	PN	Person Number /3-Char	√	√	√	√	√	√
HHIDPN	HHIDPN	HHIDPN: Hhold ID + Person Number /Num	√	√	√	√	√	√
RAHHIDPN	RAHHIDPN	Hhold ID + Person Number /9-Char	√	√	√	√	√	√
HACOHORT	HACOHORT	Sample cohort	√	√	√	√	√	
HAOAHDHH	HAOAHDHH	Overlap/AltID-Ahead(orig) HHID/Num		√				
	WAVE_NUMBER	HRS wave number	√	√	√			
	WAVE_NAME	Wave label/12-Char	√	√	√			
	STUDYYR[6]	Study year /Num	√	√	√			
	CYEAR	Study year /10-Char	√	√	√			
H[w]HHID	H_HHID	Hhold ID + SubHHold /Num	√	√	√			
H[w]HHIDC	H_HHIDC	Hhold ID + SubHHold /7-Char	√	√	√			
S[w]HHIDPN	S_HHIDPN	Spouse HHIDPN /Num	√		√			
INW[w]	INW_ALLW[7]	INW variables summary/15-Char				√		
R[w]IWSTAT	R_IWSTAT_ALLW[7]	Interview status summary/15-Char				√		
S[w]IWSTAT	S_IWSTAT_ALLW[7]	Spouse interview status summary/15-Char				√		
R[w]PEXIT	R_PEXIT_ALLW[7]	Exit or Post-Exit interview summary/15-Char				√		
REXITWV	REXITWV	EXIT Wave Exit Interview was Administered				√	√	
	RSSI_episode	Respondent SSI/SSDI episode number /Num						√

[6] The study year is identified using wave number (WAVE_NUMBER) and sample cohort variable (HACOHORT) (Table 1. Source of Data for Entry Cohorts in RAND HRS Longitudinal File by Wave codes in RAND HRS Longitudinal File 2020 (V2) Documentation).

[7] INW_ALLW, R_IWSTAT_ALLW, S_IWSTAT_ALLW and R_PEXIT_ALLW are 15-character string summary variables. The character at wth position indicates INW[w], R[w]IWSTAT, S[w]IWSTAT and R[w]PEXIT responses at wave w, respectively.

Rlong, Hlong and Slong data tables: Wave-specific variables R[w]*var*, H[w]*var* and S[w]*var*

Out of 17,653 variables in RANDHRS1992_2020V2 file, 98% (N=17,331) are wave-specific: 7,875 R[w]var for respondent-level variables, 976 H[w]var household variables, 8,465 S[w]var for spouse-level variables and 15 INW[w] for wave indicators, where [w] indicates the wave from 1 through 15 for which the variable was collected. We reshaped these variables into the long format with variable names assigned as R_var, H_var, and S_var. We also created wave number (WAVE_NUMBER) and study year (STUDYYR) to indicate the wave and HRS interview year, respectively. An example is R[w]BACK (R had back problems). Instead of having 15 variables R1BACK – R15BACK that represent back problem, the long-formatted Rlong_table contains a single variable R_BACK with multiple wave-specific values, and the `WAVE_NUMBER` variable identifies each wave-specific values. This approach yields a 90.1% reduction of variables (N = 1,714), reducing the number of variables that are needed for the respondent, household, and spouse data.

In most cases, R_var, H_var, and S_var names are derived from R[w]var, H[w]var, and S[w]var names using the convention as illustrated above for the R_BACK variable. However, there are a few exceptions. Specifically, some RAND variables are associated with different SAS formats (different values) across all waves, even though they have the same var names. For example, R1BATH and R2BATH have different formats compared to the remaining ones, R3BATH – R15BATH. We did not combine them into a single R_BATH variable because it could lead to an incorrect interpretation of the data. Instead, we created variables R1_BATH for wave 1, R2_BATH for wave 2, and R_BATH for the remaining waves (Table 3). As we noted, the created variables R1_BATH, R2_BATH, and R_BATH pertain to different study waves, and this facet is captured in the `WAVE_SUMMARY` column in the dictionaries.txt file. More specifically, positions of the 'x' character in a ` WAVE_SUMMARY ` string represent a wave for which a given variable was asked, allowing researchers to identify the data collection pattern quickly (Table 3). For the 1992 – 2020 interview, the number of waves was represented by 15 strings in 'WAVE_SUMMARY' column.

Table 3. Converting R[w]BATH wave-specific variables

Wave number	Description	RAND variables	RLONG variables	SAS format	WAVE_SUMMARY column in dictionaries.txt
wave 1	1.Not at all difficult 2.A little difficult 3.Somewhat difficult 4.Very difficult/can't do 5.Don't do	R1BATH	R1_BATH	ADL_1C.	xoooOooooOooooO
wave 2	0.No 1.Yes,a little 2.Yes,a lot 4.Yes,RF how much 9.Don't do	R2BATH	R2_BATH	ADL_2C.	oxooOooooOooooO
waves 3 - 15	0.No 1.Yes 2.Can't do 9.Don't do	R3BATH - R15BATH	R_BATH	ADLA.	ooxxXxxxxXxxxxX

To preserve household information contained in the RAND data, we created Hlong_table at the respondent level same way as it was done in the RAND data. Users can create a household-level file by selecting records where H[w]PICKHH = 1. For further information, we suggest users need to refer to the RAND document.

Rwide and Rexit tables: wave-invariant variables RA*var* and RE*var*

There are 55 respondent-level variables not specific to any single wave, and the variable names are organized as RAvar, which are identical to the variable names in the RAND. An example is RABDATE, which is the birth date of the respondent. Also, there are 100 exit interview variables, REvar. An example is REMSTAT, which is the respondent's marital status at the time of death. These variables are saved 'as is' in the output data tables using the same format as RAND.

In creating the output tables of the longitudinal data, we applied some filters to remove unnecessary missing observations (see Table 1). The longitudinal data includes only those who participated in a specific wave interview, requiring only those with INW[w] =1. In addition, we created variables that indicate respondents' status at each wave, including non-responders, those deceased in a specific wave, those deceased in previous waves, and those who dropped out of the study. To provide comprehensive details of respondents' status at each wave, we created summary variables in the Rwide table. These summary variables are INW_ALLW, R_IWSTAT_ALLW, S_IWSTAT_ALLW, and R_PEXIT_ALLW, each containing a 15-character string to represent the respondent's or spouse's status at each wave. The character at a wth position indicates the value in INW[w], R[w]IWSTAT, S[w]IWSTAT, and R[w]PEXIT for the wth wave as presented in the original RAND data. We suggest that user refer to the RAND document for the value labels. For example, the summary variable 'INW_ALLW' contains strings of '1' and '0', representing participated and not participated respondents, respectively (Please see Table 4).

Table 4. An example of summary variables in Rexit table

Respondent ID	INW_ALLW	R_IWSTAT_ALLW[8]	S_IWSTAT_ALLW[9]	R_PEXIT_ALLW	REXITWV[10]
A	111110000000000	111114444477777	111U4..........	.00000000000000	Y
B	111110000000000	111115666666666	UUUUU..........	.00001000000000	6

[8] 1.Resp,alive; 4.NR,alive; 5.NR,died this wave; 6.NR,died previous wave; 7.NR,dropped from sample

[9] 1.Resp,alive; 4.NR,alive; .U Reference person is not married

[10] .Y No Exit Interview, No Death Recorded

In Table 4, we illustrate the summary variables using scenarios of respondents' status of two respondents. Both A and B respondents participated in 1st through 5th waves. However, respondent A did not participate in the next five waves but known to be alive (R_IWSTAT_ALLW = 4 from 6th through 10th waves), and this respondent was completely dropped from the study at 11th wave (R_IWSTAT_ALLW=7 from 11th through 15th waves). On the other hand, respondent B is known to be dead at the 6th wave (R_IWSTAT_ALLW = 5) and the exit interview was administered, which was identified by the variable R_PEXIT_ALLW for the exit interview (R_PEXIT_ALLW = 1 at 6th wave). In addition, examining marital stats, respondent A had a spouse who participated in the first three waves (S_IWSTAT_ALLW = 1 from 1st through 3rd waves), but the marital status changed thereafter. Respondent A was unmarried or unpartnered at the fourth wave (S_IWSTAT_ALLW = .U) and remarried at the fifth wave, but the spouse didn't participate. These seeing variables, containing participating pattern, can help users easily determine a respondent's status at each wave.

RSSI table: Disability Episode Variables

The disability episode variables are not wave specific. However, the respondent can have multiple episodes (up to 11 episodes in RANDHRS1992_2020V2 file). For users' convenience, we reshaped 154 RAvar[n] (where n is an episode number 1, 2, ..., 11) into 14 RAvar_E variables and created episode number variable (RSSI_episode).

Table 5. Disability Episode Variables

Measure	Description	RAND variables	RSSI variables
Number of episodes	Count, at most 11 as of wave 15	RADNEPI	RADNEPI
Episode number	Respondent SSI/SSDI episode number		RSSI_EPISODE
Application date	Month	RADAPPM1 to RADAPPM11	RADAPPM_E
	Year	RADAPPY1 to RADAPPY11	RADAPPY_E
	"Best-guess" SAS date	RADAPPD1 to RADAPPD11	RADAPPD_E
Appeal or Reapply date	Month	RADREAM1 to RADREAM11	RADREAM_E
	Year	RADREAY1 to RADREAY11	RADREAY_E
	"Best-guess" SAS date	RADREAD1 to RADREAD11	RADREAD_E
Receive date	Month	RADRECM1 to RADRECM11	RADRECM_E
	Year	RADRECY1 to RADRECY11	RADRECY_E
	"Best-guess" SAS date	RADRECD1 to RADRECD11	RADRECD_E
Stop date	Month	RADENDM1 to RADENDM11	RADENDM_E
	Year	RADENDY1 to RADENDY11	RADENDY_E
	"Best-guess" SAS date	RADENDD1 to RADENDD11	RADENDD_E
Type	1=SSDI 2=SSI	RADTYPE1 to RADTYPE11	RADTYPE_E

	3=DK which 4=SSDI/SSI at different waves		
Current Status	Indicates if applied, receiving, stopped receiving, or illogical ends	RADSTAT1 to RADSTAT11	RADSTAT_E

Appendix

Merging multiple data tables

The output data tables can be merged by key variables in Table 2. Rlong_table, Hlong_table and Slong_table are uniquely identified by a respondent ID (HHID and PN) and study year (STUDYYR) while Rwide_table and Rexit_table are not wave-specific, only respondent-level files. To merge these five data from all waves, simply run the following code:

```
libname lib "&output"; /* &output is the name of the directory where the output tables are stored. */

/* Create `work.formats` catalog */
proc format lib = WORK cntlin = lib._RANDfmts_long;
run;

/* One-to-one merge */
data rsh;
 merge lib.rlong_table lib.slong_table lib.hlong_table ;
    by hhid pn studyyr;
run;

/* One-to-many merge
   mrg5_tables is uniquely identified by HHID, PN and STUDYYR */
data mrg5_tables;
 merge rsh lib.rexit_table lib.rwide_table;
  by hhid pn;
run;

proc contents data = mrg5_tables position;
run;
```

RSSI_table is not wave specific and can be merged by a respondent ID (HHID and PN). However, the respondent can have multiple episodes. By merging it with Rwide/Rexit table, a new dataset is produced and uniquely identified by HHID, PN, and RSSI_EPISODE variables.

Summary variables

Although the summary variables are string formatted, representing all the 15 waves, users can create informative variables at a specific wave. For example, among respondents who participated in wave 10 (R_IWSTAT10), we can identify who withdrew from the study within the next 2 waves (DROP) and how many interviews they participated in last 5 waves (N_LOOKBACK). These functions are useful in determining respondents' status and pattern for cross-sectional and longitudinal analyses.

```
data summary;
 set lib.rwide_table;

 /*An example is wave 10*/
 %let w = 10;

 /*a respondent's status at wave 10*/
 R_IWSTAT&w = input(substr(R_IWSTAT_ALLW, &w, 1), 3.);
 if R_IWSTAT&w = 1 then do;
    /*the respondent was dropped out of the study within the next 2 waves*/
      DROP_WAVE = find(substr(R_IWSTAT_ALLW, 1, &w.+ 2), "7");
      DROP = 1*(DROP_WAVE > 0);
    /*the total number of time a respondent participated in last 5 interviews*/
      N_LOOKBACK = count(substr(R_IWSTAT_ALLW, &w.-6, 5), "1");
 end;
run;
```

Furthermore, users can easily recreate the original RAND variables (15 variables, each representing a wave) using the macro SAS codes below.

```
/* Create `work.formats` catalog */
proc format lib = WORK cntlin = lib._RANDfmts_long;
run;

/* To convert summary variables to the original RAND variables, INW[w], R[w]IWSTAT, S[w]IWSTAT, R[w]PEXIT
*/
%let INWw = INW1 - INW15;
%let RwIWSTAT = R1IWSTAT R2IWSTAT R3IWSTAT R4IWSTAT R5IWSTAT R6IWSTAT R7IWSTAT R8IWSTAT
                R9IWSTAT R10IWSTAT R11IWSTAT R12IWSTAT R13IWSTAT R14IWSTAT R15IWSTAT;
```

```
%let SwIWSTAT = S1IWSTAT S2IWSTAT S3IWSTAT S4IWSTAT S5IWSTAT S6IWSTAT S7IWSTAT S8IWSTAT
                S9IWSTAT S10IWSTAT S11IWSTAT S12IWSTAT S13IWSTAT S14IWSTAT S15IWSTAT;
%let RwPEXIT = R1PEXIT R2PEXIT R3PEXIT R4PEXIT R5PEXIT R6PEXIT R7PEXIT R8PEXIT
               R9PEXIT R10PEXIT R11PEXIT R12PEXIT R13PEXIT R14PEXIT R15PEXIT;
data org_RAND;
 set lib.rwide_table(keep=RAHHIDPN INW_ALLW R_IWSTAT_ALLW S_IWSTAT_ALLW R_PEXIT_ALLW);

 missing U V;
 array INWw{*} &INWw;
 array RwIWSTAT{*} &RwIWSTAT.;
 array SwIWSTAT{*} &SwIWSTAT.;
 array RwPEXIT{*} &RwPEXIT.;

      do i = 1 to dim(INWw);
            INWw{i}     = input(substr(INW_ALLW, i, 1), 3.);      /*original RAND variables INW[w]*/
            RwIWSTAT{i} = input(substr(R_IWSTAT_ALLW, i, 1), 3.); /*original RAND variables R[w]IWSTAT*/
            SwIWSTAT{i} = input(substr(S_IWSTAT_ALLW, i, 1), 3.); /*original RAND variables S[w]IWSTAT*/
            RwPEXIT{i}  = input(substr(R_PEXIT_ALLW, i, 1), 3.);  /*original RAND variables R[w]PEXIT*/
      end;
 drop i;

 format &INWw. INWv. &RwIWSTAT. &SwIWSTAT. IWSTAT. &RwPEXIT. PEXIT.;
run;
```