MIDUS
Midlife in the United States
A National Study of Health & Well-Being

Conventions for Creating MIDUS Refresher 1 & MIDUS 3 Datasets

The MIDUS Administrative Core has developed conventions for creating MIDUS datasets for the MIDUS Refresher 1 and MIDUS 3. Such conventions are necessary for efficient and accurate data merges across Projects (e.g., survey, cognitive, daily diary, biomarker, neuroscience, and genetics) and across different samples or waves of data (MIDUS 1 through 3, Refresher1 and 2, and Milwaukee). Further, the introduction of metadata standards such as the Data Documentation Initiative (DDI), which MIDUS adheres to, make the conventions an integral part of accurate documentation of the MIDUS study.

The attached pages provide specific guidelines for naming, labeling and formatting variables in the MIDUS Refresher 1 and M3. We have also included coding conventions for variables, missing value designations, and guidelines for date and time variable formats.

I. File Naming Conventions

File naming conventions help manage and organize MIDUS. These conventions become increasingly more useful as MIDUS becomes more complex. For all file types, MIDUS will use the prefixes MR1 and M3 to designate the Refresher 1 and MIDUS 3, respectively, followed by an underscore and the project number. In addition, file names end with a date stamp, and two examples of which follow:

Examples: Documentation/Instruments

Refresher 1: MR1_P1_PhoneInstrument_20160114
MIDUS 3: M3_P1_FieldReport_20180417

Data files should include additional information on the number of cases (and avoid using special characters such as =, &, %, etc.):

Examples: Datasets

Refresher: MR1_P1_DATA_N2100_20160616
MIDUS 3: M3_P1_DATA_N5000_20120508

II. Variable Naming Conventions

Rationale:

- Metadata best practices support a consistent and simple variable naming scheme. Not only does this reinforce the continuity of longitudinal data, but it makes cleaning and

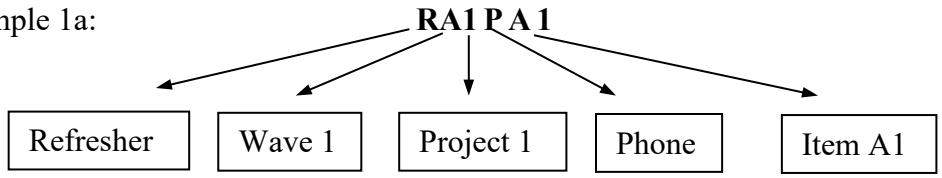
programming new variables more efficient and ensures compatibility across different software platforms.

- The original naming conventions were adopted in 2004 when there were strict character limits on variable names in statistical software. While current software programs are much more lenient in this regard, there are still substantial differences across statistical programs, and some older versions of statistical software still adhere to smaller variable name character limits.
- For these reasons, we will continue to limit variable name size, but because the Refresher cohort requires an "R" as the first character to identify the new sample (see examples 1a and 1b below), a 9-character variable name limit will be used**.**
- For the Refresher sample, the first 4 characters of each variable name will identify the cohort, longitudinal wave, the MIDUS project, and the instrument used to collect the data. The remaining characters identify the specific item or scale score variable that is represented by the measure's name. MIDUS 3 will follow the same conventions but will use the first 3 characters to identify wave, project, and instrument. The exception to these conventions is the Project 1 Milwaukee data. The Milwaukee sample is new at M2 and used a different instrument (a personal interview instead of a phone interview) to collect Project 1 survey data from these individuals. Thus, Project 1 variables for the Milwaukee data include an additional character "A" to designate the project.[1]

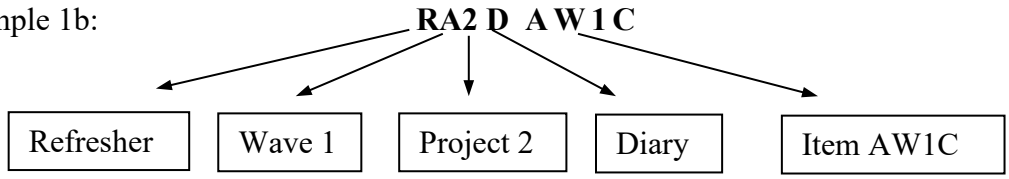
Examples: MIDUS Refresher 1

For the MIDUS Refresher 1 data, the first character of each variables name will be **R**. Otherwise, the extant naming conventions apply, i.e. those developed for M2.

Example 1a:

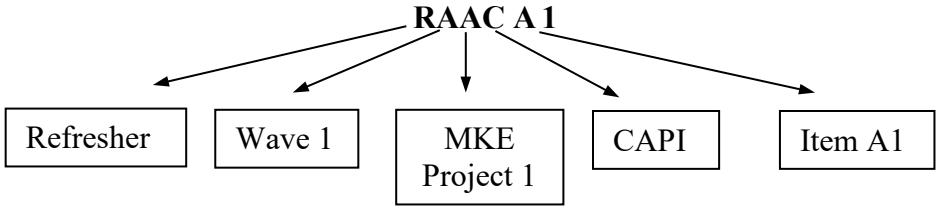


Example 1b:



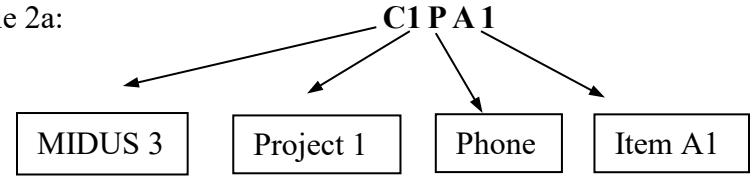
[1] Moving forward, the MKE2 variable names will continue the convention adopted at wave 2; MKE1 baseline was begun in 2005 during M2 data collection and so adopted the "B" character to indicate Wave2. Likewise, other projects like biomarkers and neuroscience who began baseline data collection at Wave2 also adopted variable names that began with "B". For MKE2, variable names will mimic those of M3 with "C" as the first character.

Example for Milwaukee Project 1:

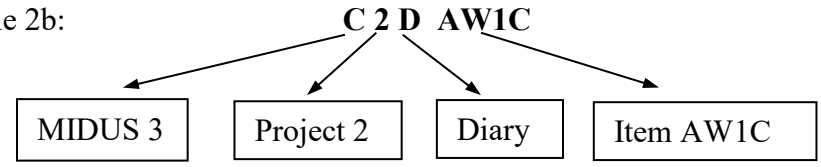


Examples: MIDUS 3

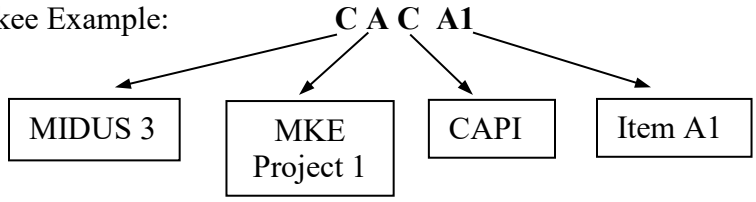
Example 2a:



Example 2b:



Milwaukee Example:



III. Variable Labeling Conventions

Additional information about variables can (and should!) be included in the variable label. The label is the appropriate metadata field to more fully and clearly describe a variable. New technological metadata standards can use the rich information contained in a label to harvest, search, and identify specific variables. For variable labels, we encourage the use of mixed case text for more sentence-like descriptions of variables.
See examples below.

Example 1a:

Variable name: **RA1PA4**
Variable label: **Days unable to work because of health (30 days)**

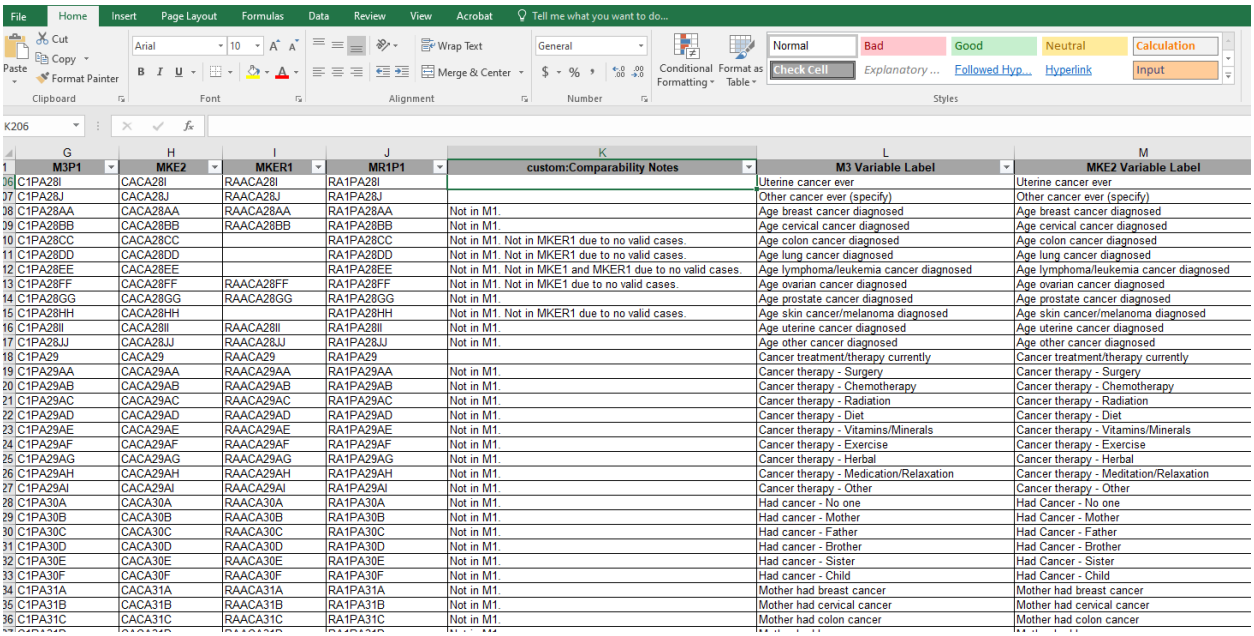
Example 1b:

Variable name: **RA4QCESD**
Variable label: **CESD: Center for Epidemiologic Studies Depression Scale**

IV. Variable concordance tables

The increasing number of waves and samples in MIDUS can make navigating among the datasets a challenge. We strongly suggest that each MIDUS project create variable concordance or crosswalk tables similar to the Excel spreadsheet created by Project 1 (see Figure 1 below). These tables not only help researchers find related variables, but also can be used by DDI codebook applications to facilitate variable searches across different MIDUS datasets. See the "Explore" and "Concordance Variables" views in the online MIDUS Portal (http://midus.colectica.org/Explore) to see how DDI applications make use of such concordance tables. Contact midus_help@aging.wisc.edu for more details if needed.

Figure 1: MIDUS Concordance Table



M3P1	MKE2	MKER1	MR1P1	custom:Comparability Notes	M3 Variable Label	MKE2 Variable Label
C1PA28I	CACA28I	RAACA28I	RA1PA28I		Uterine cancer ever	Uterine cancer ever
C1PA28J	CACA28J	RAACA28J	RA1PA28J		Other cancer ever (specify)	Other cancer ever (specify)
C1PA28AA	CACA28AA	RAACA28AA	RA1PA28AA	Not in M1.	Age breast cancer diagnosed	Age breast cancer diagnosed
C1PA28BB	CACA28BB	RAACA28BB	RA1PA28BB	Not in M1.	Age cervical cancer diagnosed	Age cervical cancer diagnosed
C1PA28CC	CACA28CC		RA1PA28CC	Not in M1. Not in MKER1 due to no valid cases.	Age colon cancer diagnosed	Age colon cancer diagnosed
C1PA28DD	CACA28DD		RA1PA28DD	Not in M1. Not in MKER1 due to no valid cases.	Age lung cancer diagnosed	Age lung cancer diagnosed
C1PA28EE	CACA28EE		RA1PA28EE	Not in M1. Not in MKE1 and MKER1 due to no valid cases.	Age lymphoma/leukemia cancer diagnosed	Age lymphoma/leukemia cancer diagnosed
C1PA28FF	CACA28FF	RAACA28FF	RA1PA28FF	Not in M1. Not in MKE1 due to no valid cases.	Age ovarian cancer diagnosed	Age ovarian cancer diagnosed
C1PA28GG	CACA28GG	RAACA28GG	RA1PA28GG	Not in M1.	Age prostate cancer diagnosed	Age prostate cancer diagnosed
C1PA28HH	CACA28HH		RA1PA28HH	Not in M1. Not in MKER1 due to no valid cases.	Age skin cancer/melanoma diagnosed	Age skin cancer/melanoma diagnosed
C1PA28II	CACA28II	RAACA28II	RA1PA28II	Not in M1.	Age uterine cancer diagnosed	Age uterine cancer diagnosed
C1PA28JJ	CACA28JJ	RAACA28JJ	RA1PA28JJ	Not in M1.	Age other cancer diagnosed	Age other cancer diagnosed
C1PA29	CACA29	RAACA29	RA1PA29		Cancer treatment/therapy currently	Cancer treatment/therapy currently
C1PA29AA	CACA29AA	RAACA29AA	RA1PA29AA	Not in M1.	Cancer therapy - Surgery	Cancer therapy - Surgery
C1PA29AB	CACA29AB	RAACA29AB	RA1PA29AB	Not in M1.	Cancer therapy - Chemotherapy	Cancer therapy - Chemotherapy
C1PA29AC	CACA29AC	RAACA29AC	RA1PA29AC	Not in M1.	Cancer therapy - Radiation	Cancer therapy - Radiation
C1PA29AD	CACA29AD	RAACA29AD	RA1PA29AD	Not in M1.	Cancer therapy - Diet	Cancer therapy - Diet
C1PA29AE	CACA29AE	RAACA29AE	RA1PA29AE	Not in M1.	Cancer therapy - Vitamins/Minerals	Cancer therapy - Vitamins/Minerals
C1PA29AF	CACA29AF	RAACA29AF	RA1PA29AF	Not in M1.	Cancer therapy - Exercise	Cancer therapy - Exercise
C1PA29AG	CACA29AG	RAACA29AG	RA1PA29AG	Not in M1.	Cancer therapy - Herbal	Cancer therapy - Herbal
C1PA29AH	CACA29AH	RAACA29AH	RA1PA29AH	Not in M1.	Cancer therapy - Medication/Relaxation	Cancer therapy - Meditation/Relaxation
C1PA29AI	CACA29AI	RAACA29AI	RA1PA29AI	Not in M1.	Cancer therapy - Other	Cancer therapy - Other
C1PA30A	CACA30A	RAACA30A	RA1PA30A	Not in M1.	Had cancer - No one	Had Cancer - No one
C1PA30B	CACA30B	RAACA30B	RA1PA30B	Not in M1.	Had cancer - Mother	Had Cancer - Mother
C1PA30C	CACA30C	RAACA30C	RA1PA30C	Not in M1.	Had cancer - Father	Had Cancer - Father
C1PA30D	CACA30D	RAACA30D	RA1PA30D	Not in M1.	Had cancer - Brother	Had Cancer - Brother
C1PA30E	CACA30E	RAACA30E	RA1PA30E	Not in M1.	Had cancer - Sister	Had Cancer - Sister
C1PA30F	CACA30F	RAACA30F	RA1PA30F	Not in M1.	Had cancer - Child	Had Cancer - Child
C1PA31A	CACA31A	RAACA31A	RA1PA31A	Not in M1.	Mother had breast cancer	Mother had breast cancer
C1PA31B	CACA31B	RAACA31B	RA1PA31B	Not in M1.	Mother had cervical cancer	Mother had cervical cancer
C1PA31C	CACA31C	RAACA31C	RA1PA31C	Not in M1.	Mother had colon cancer	Mother had colon cancer

V. Variable Formats

1. Utilize "Numeric" whenever possible; avoid using raw string variables. Open-ended responses, text, and verbatim data should be numerically coded where possible. Raw text or qualitative data can be formatted as a separate text file or spreadsheet.

2. Variable formats should be precise – variable lengths should not exceed the maximum number of digits possible for a response. Thus, if a response code has a maximum of 2 digits (e.g., a scale ranging from 1-10) then the variable length should be formatted as 2 digits.

3. Decimals: Specify up to 3 decimal places (an ICPSR convention). If important details of the data require more decimal places, please contact midus_help@aging.wisc.edu.

4. Date/time formats:

- Because of conflicting or proprietary formatting, date or time data provided **by *respondents*** must be separated into individual components. For example, date information must be recorded in separate month and year variables, and temporal information must be recorded in separate hour, minute and meridian (am/pm, etc.) variables (one alternative for temporal variables is using a 24-hour clock or military time, in which hours and minutes can be represented as a numeric variable HHMM). Treating such variables this way allows them to be read by a wider array of software programs with fewer formatting problems or errors.
- Dates or times recorded **by *project staff for administrative purposes*** (e.g. date & time Medical History was completed) can be submitted to the Core in the date and time formats specified in the established conventions for MIDUS (e.g., mm/dd/yyyy or hh:mm:ss, etc.).
- **Note: because of confidentiality/disclosure precautions, ICPSR suggests that month and year are sufficient for most sensitive date variables.**

5. Derived or constructed variables: If possible, any constructed or derived variables (e.g., scales, scores or summary variables) are placed directly after their components in the dataset. That is, derived variables should follow their constituent variables in the sequence of variables in a dataset. Also, the details of the construction or derivation of such variables are to be explained in accompanying documentation (e.g., a Word document explaining the formula, procedure, source, criteria, etc., used in creating the constructed or derived variables).

VI. Value Labeling Conventions

A. Format

All value labels should be **UPPER CASE**.

Example:

1 = YES
2 = NO
7 = DON'T KNOW
8 = REFUSED/MISSING
9 = INAPP

B. Coding Conventions for Non-response (Don't Know, Missing Data, Inapplicable, Filters)

The following values should be used to indicate different types of non-response to questions or data fields (a series of 9's can be used as place holders):

DON'T KNOW (7's) - 7, 97, 997, 9997, etc.
Used to indicate explicit "Don't know" responses (where a specific response option of "Don't know" was offered).

REFUSED/MISSING (8's) - 8, 98, 998, 9998, etc.
Used to indicate R did not provide a response to a particular question.

INAPPLICABLE (9's) - 9, 99, 999, 9999, etc.
Used to indicate that R was not asked a particular question. This will occur most often because of skip patterns that the R is asked to follow or questions that R determines do not apply to him or her.

INCOMPLETE SAQ (-1)
The Refresher and M3 uses a new convention for coding SAQ non-responders in the Project 1 protocol. SAQ variables for those cases that did not return a completed SAQ will be coded -1. Project 1 SAQ variables from M1 and M2 will eventually be coded in the same way.

C. Additional Coding Situations.

In M2, some projects used additional codes to indicate invalid values or incomplete data. We suggest using the value of 96 (and working backwards to add additional codes). For example, Project 2 assigned additional codes to cortisol variables that did not contain valid data (e.g., 96 = empty vial, 95 = not done, 94 = unreliable).