MIDUS 2 BIOMARKER PROJECT DATA FILE NOTES
March 2024

This document highlights aspects of the MIDUS 2 Biomarker data that analysts should be aware of prior to working with the data. In particular, Biomarker specific variable naming conventions and administrative variables are described along with issues that arose across multiple sections of the data (e.g. complex items that are represented using multiple variables, accommodating instances when respondents offer more information than our instruments provide room for). Information also includes coding of some text data. General issues and information that apply across multiple types of data are listed first, followed by issues that are unique to specific instruments.

This document provides general information. For more detailed information about specific sections of the data, see the documentation files for the following data types. The files are described in more detail in the README FIRST file:

- Blood, Urine and Saliva Assay Data
- Musculoskeletal Health Data
- Medication Data
- Physical Exam Data
- Psychophysiology Protocol
- Sleep Data (includes Actigraphy data)

Also included among the documentation files are:

- Biomarker Project Summary – an overview of the Biomarker Project
- Documentation for PsychoSocial Constructs and Composite Variables
- Documentation for Coded Text Responses
- Self-Administered Questionnaire (SAQ) – PDF of the SAQ booklet
- Medical History Interview – PDF of the Medical History booklet

Details about the data and links to the related documentations are also available via the MIDUS Colectica Portal (http://midus.colectica.org/) which houses interactive codebooks for all the publicly available MIDUS projects. The Portal includes search and explore functions, links to documentations, and a custom download function. A link to the portal is also available on the MIDUS website (http://midus.wisc.edu/) under QuickLinks.

Variable Names

MIDUS variable naming and coding conventions (See "Naming and Coding Conventions" documentation for the MIDUS 2 Survey project) specify that the first two characters of the variable name indicate the wave and project number, thus, the MIDUS 2 Biomarker variables begin with 'B4'. The third character of the variable name is a letter that identifies the data type, collection method or name of the instrument used to collect the data. The Biomarker (P4) data types, instruments, and data collection methods are designated by the indicated letters below:

Z = Biomarker (P4) Administrative Variables

Q = Self-Administered Questionnaire (SAQ)
S = Pittsburgh Sleep Questionnaire (PSQ)
H = Medical History
P = Physical Exam (Short & Long versions)
X = Medication Chart
O = Bone Questionnaire
B = Biomarker Assays (Blood, Urine, and Saliva)
A = Daily Sleep Diary (used with the Actiwatch)
W = Actiwatch Watch data
D = Bone Densitometry, Body Composition
V = Psychophysiology - HRV, Beat-to-Beat Blood Pressure & Respiration

Administrative Variables

The data file includes the following administrative variables from the MIDUS 2 Survey:

- M2ID – MIDUS 2 ID included in all public MIDUS files to facilitate linkage to data from other MIDUS longitudinal projects (i.e. all MIDUS 1, MIDUS 2, and MIDUS 3 projects)
- M2FAMNUM – MIDUS 2 family ID, facilitates identification of MIDUS participants who are related to each other (i.e. Twin pairs)
- SAMPLMAJ – the major sample identifier (i.e. Main sample, Twins, or Milwaukee)
- B1PAGE_M2 – participants age at the time of the MIDUS 2 Survey Phone Interview
- B1PRSEX – participants' gender

The file also includes the following administrative variables specific to the MIDUS Biomarker data:

- B4ZSITE – Biomarker data collection site
- B4ZCOMPM, B4ZCOMPY – Month and Year the General Clinical Research Center (GCRC) visit was completed
- Time Zone variables: created to provide a means of controlling for variation in saliva cortisol level that could be due to changes in time zone for data collection
  - B4ZRZONE – indicates the Time Zone the Respondent lived in at the time of the clinic visit
  - B4ZSZONE – time zone where the data collection Site is located
- Lag Variables: indicating the time elapsed between the Biomarker (P4) Completion date and completion of other MIDUS 2 projects
  - B4ZB1PLG – Project 1 (Survey) Phone interview
  - B4ZB1SLG – Project 1 (Survey) SAQ
  - B4ZB3CLG – Project 3 (Cognition) Cognitive assessments
- Respondent age when the Biomarker data was collected computed in two ways (see Documentation of Scales and Composite Variables for computation details) one uses rounding the other does not.
  - B4ZAGE – computed age included in original MIDUS 2 Biomarker data release
  - B4ZAGE_D – new age variable (rounded *down*) computed to be consistent with the age variable in the Refresher Biomarker and the MIDUS Survey data sets

- Tissue Sample Collection: variables indicating whether complete tissue samples were obtained. These variables appear with the assay data (B4B variables) immediately in front of the data related to a given tissue sample type
    - B4ZBLOOD – Blood sample completeness
    - B4ZURINE – Urine sample completeness
    - B4ZSALIVA – Saliva sample completeness

Filter Variables

Filter variables can be used to identify subsets of cases to include or exclude from analysis. Some of the administrative variables above can be used as filter variables. In addition, one or more filter variables were created for use with individual data types (e.g. assay data, bone and whole body composition, psychophysiology, etc.). Details about filter variables are provided in the documentation for a given data type.

Date and Time Variables

Date and time data often have proprietary or conflicting formatting characteristics that can create problems when moving between data file types (i.e. SPSS to Excel). Thus, the MIDUS Naming and Coding Conventions require that dates and times be converted to formats that allow them to be read by a wider array of software programs. In response to this requirement, dates and times in the updated MIDUS 2 Biomarker are included as follows:

1. Dates – dates continue to be presented as separate month and year variables.
2. Times – in the original MIDUS 2 Biomarker data times were reported in a 12-hour clock SPSS time format. *That convention was modified at the Refresher* such that all MIDUS 2 time variables have now been converted to a 24-hour clock (i.e. day runs from midnight to midnight and is divided into 24 hours, indicating the number of hours since midnight, from 0 to 23). The time variables are also formatted as 4-digit SPSS Restricted Numeric, which allows leading zero(s) to be displayed. Thus, 1:00 a.m. appears in the data as '0100', while 10:00 p.m. as '2200'.
    a. Time variables in this format are found in:
        i. The Pittsburgh Sleep Questionnaire (PSQ) data
        ii. Urine Collection Times (included with the Urine Assay data)
        iii. Daily Sleep Diary and Actiwatch data
        iv. Psychophysiology Flowsheet data

Added Variables

The following describes instances where multiple variables may be used to report items that are more complex.

1) One Item – Multiple Variables:

The format of the following questions in the indicated instruments requires that multiple variables (as noted) be created in order to adequately capture the complexity of the data.

a. Medical History: Current Health Practices: Diet & Exercise (p.12)
    i. "Please estimate your daily calcium intake" (Q17).

The interviewer asks the respondent how many servings of Milk (Q17a), Yogurt (Q17b), and Cheese (Q17c) s/he consumes on a daily basis.

ii. "On an average DAY, how many 8 ounce cups or glasses do you drink of …. Coffee with caffeine, Tea with caffeine, Other beverages with caffeine (e.g. Coke)? (Q18a-c).

Many respondents consume the above on a daily basis, but others consume them on a weekly, monthly, and even yearly basis. Thus, responses to these questions are entered as two variables. The first variable is the number of times in a given time frame that a person consumes the indicated item. The second variable is a categorical variable reflecting the commonly mentioned time frames (day, week, month, year). These variables have the following format:

B4H17_F or B4H18_F – frequency
B4H17_T or B4H17_T – time frame categories

b. Medical History – Current Exercise (p.14). "….do you engage in regular exercise, or activity, of any type for 20 minutes or more at least 3 times/week?" Q25a-c. If the respondent says "Yes", the interviewer obtains information about the type of activity (recorded as open-ended text) along with other details including the frequency with which the activity is performed.

i. Respondents often report more than 3 types of exercise or activity. Thus, a supplemental "Additional Activities" form was created allowing staff to record information up to 7 forms of regular exercise/activities (B4H25a-g). Data about additional exercise are entered in a format paralleling that of the items appearing in the booklet.

Note: Text responses describing the type of exercise/activity reported have been coded. See details in the section on Recodes/Additional Codes below.

ii. Sometime respondents report performing activities more than once in a given day several times per week (e.g. walking in the morning and evening 5 days per week). This information is entered using 2 variables for each type of exercise, one reflecting the number of times per day and the other reflecting the number of days per week. These variables have the following format:

B4H25_FD – how many times per day
B4H25_FW – how many days per week

iii. Sometimes respondents indicate that certain activities are seasonal (e.g. gardening, snow skiing, bicycling). To capture this information, we also assess seasonality of an activity. The options include "Not Seasonal" for year round activities, to codes for individual season or combinations of seasons (e.g. Spring, Summer) for each activity reported. These variables have the format B4H25_S where the 6th character is a-g as appropriate.

iv. Metabolic Equivalent of Task (METs) - Quantitative details about frequency and intensity of activities has been used to compute a new variable (B4HMETW) indicating the 'total number of Metabolic Equivalent of Task (MET) minutes per

week'. Details about this variable can be found in the Documentation of Psychosocial Constructs and Composite Variables.

c. Medication Chart: The following columns require two variables to capture the relevant information.

Drug Dosage – the drug dosage has two components, the quantity and the unit (e.g. 200 mg or 3 tablets)

Frequency – the frequency with which a respondent takes a given medication has two components, "how often" and a time frame (e.g. 2 times per day or every other day)

Taken for how long? – the period of time in which the respondent has been taking a given medication has two components, quantity and time frame (e.g. 6 months or 10 years)

These three pieces of information are entered using 2 variables each. The first variable is the amount or frequency, and the second variable is categorical and specifies the valid responses for the unit (e.g. mg, tablet, puff, etc.) or time frame (e.g. day, week, etc.) respectively. These variables have the following format, where the fourth character indicates the medication type and the final character indicates the medication number:

B4X_DD_ – dosage
B4X_DDU – dosage unit

B4X_F_ – frequency
B4X_FU_ – time frame category

B4X_T_ – how long
B4X_TU_ – time frame category

2) Additional Others, Events, and Medications:

a) There are several sections in the Medical History in which respondents have the opportunity to report "Other" symptoms or conditions or are asked about "events" of a certain type. There are a few instances where respondents regularly report more conditions or events than space allows for. Thus, at the end of the following sections of data additional variables are included to facilitate inclusion of this information in analyses. Specifically, a variable that reflects the total number of 'other' or 'events' was created. If this number is greater than the number of spaces provided in the medical history, data staff records all the information about these additional occurrences in open-ended string variables.

Count variables, and related Yes/No variables are created for the following sections:

- Symptoms & Conditions
- Family Medical History
- Other Major Events

Details about creation of new variables can be found in:

- Documentation for Psychosocial Constructs and Composite Variables
- Documentation for Coded Text Responses

b) The number of medications that a respondent takes or the number of medication allergies reported may exceed the space provided on the Medication Chart. Procedures have been developed to maximize inclusion of relevant data. Details can be found in the Medication Data documentation.

3) Respondent Cannot Recall Date:

Respondents are not always able to recall when an event occurred, but they are able to provide verbal references (e.g. summer of 89 or in the mid 90's). In most instances, the appropriate missing value code was entered and text about the event was entered into marginal comments. At data cleaning, the reference text was reviewed, and the "missing" data was recoded to a valid response as appropriate.

Recodes/Additional Codes

During data collection and/or data cleaning, we found instances where additional codes were needed to be added to existing variables or new variables were need to be created to accurately represent respondent reports. The following describes these instances and how they were addressed.

1. There are a few questions for which enough respondents gave responses that were sufficiently ambiguous that a valid code could not be assigned or a specific number entered. As it would be inappropriate to assign the "DON'T KNOW" Missing Value code, the code "96" is utilized. The meaning of this code changes according to the item or set of items where it is specified.

a) Medical History: Immune Function: Immunizations (Q11a-d) – "Have you had….?, If Yes, how old were you?" While many respondents report the age at which they had an illness or were immunized, many are only able to say something like 'when I was a child' or 'before started school' or some other phrase indicating it occurred during childhood. In such instances, the age is coded as "96".

b) Medical History: Other Events (p.24) (Q57) – respondents report about long and short-term ongoing events/experiences (family member going through major ongoing health event, ongoing preparations for a wedding, etc.) as well as acute events. If the event is "ongoing", a "96" is entered for the month, along with the year the medical history was completed for the year.

c) Medication Chart: Some medications have two or more active ingredients having different dosages. Others contain a single active ingredient but are taken in different quantities throughout the day or on alternate days. Details about how these instances are handled can be found in the Medication Data documentation.

2. Medical History: Immune Function: Allergies (p. 8): "Do you have any allergies that have been diagnosed by a doctor or allergist?" (Q10). Many respondents report allergies that have not been diagnosed by a physician. Staff, thus, record information about all allergies reported and also note whether or not the allergy had been diagnosed by a physician. The codes for Q10 (B4BH10) were modified from Yes/No to 4 codes reflecting whether the allergies reported were diagnosed and/or undiagnosed.

3. Medical History Current Health Practices: Diet and Exercise: "On an average Day how many…"

- glasses of water do you drink? (Q19)
- sugared beverages do you drink? (Q20)
- servings of fruits and vegetables do you eat? (Q21)
- servings of whole grain do you eat? (Q22)

The response options listed in the booklet (p.12-13) for these questions do not allow for respondents who consume the indicated item less than once per day. To accommodate this omission, a fifth category "Other – Less than once a day" is added to the response options at data entry.

4. Bone Questionnaire:

a. Page 4, Q22 "When did you first notice irregularity in your menstrual cycle length (cycle length variability 7 days or more)? Women are not always able to recall exactly when they first noticed irregularity in their menstrual cycle and often provide an open-ended response that cannot be collapsed into the Months and Years format. The text responses were reviewed in combination with reported Month and Years data to create 2 variables:

- B4O22 – categorical variable indicating if the participant ever noticed any irregularity as well as other information related to irregularity (i.e. Never Irregular, Always Irregular, Surgery – No Irregularity)
- B4O22Y – numeric variable indicating the number of years ago that the participant noticed irregularity

b. Page 5, Q 26: "Do you have any rods, plates, or screws, or pins in your bones or joints?" Data cleaning revealed that the response "Elsewhere" had been excluded from Item 26. Since responses to this item appear in the data as a series of dichotomous Y/N variables, 2 existing variables were modified and a new variable was added to this series as follows:

B4O26D: relabeled to "Rod/plate location: elsewhere"

B4O26E: relabeled to "Pin location: hands or feet"

B4O26F: NEW variable labeled "Pin location: elsewhere"

5. Sometimes lab values are reported as ">" or "<" some value. When this occurs, the reported value is replaced with a value 'one unit' below the minimal or maximal detectable score. For example:

If the lower limit was <1.0, then we could change all of those scores to 0.9.

If the highest possible score was 120 for a particular test, and was not normally reported with decimal values, then all the >120 would be converted to 121.

Extremely high and low values, therefore, are curtailed. Consequently, the variance on the tails of the distribution is truncated. This is only problematic if a high percent of values fall in this category.

Missing Values

1. Throughout all the data files Missing Values appear as follows (see MIDUS Naming and Coding Conventions for details):
   7, 97, 997, etc. = DON'T KNOW
   8, 98, 998, etc. = MISSING
   9, 99, 999, etc. = INAPP

2. Physical Exam (Short Version): "What is the tallest you've been measured in your life?" (Q2d). This question was added to this protocol in October/November 2004, thus, this item is INAPP for respondents who completed P4 before the question was added.

3. Bone Questionnaire: Questions (Q23-27d) regarding metal in the body were added to the questionnaire in June 2007 when a Whole Body scan was added to the Bone Density protocol. Due to site-specific differences, this change was implemented in June 2007 at Site 2, in August 2007 at Site 1, and in February 2008 at Site 3. These items are missing or INAPP for respondents who completed P4 before this change was implemented.

4. Biomarker assay result data (variable name begins B4B…) may be INAPP or MISSING for the following reasons:
   a) The assay could not be performed because blood, urine or saliva sample was partial or the sample was not collected for some reason. This can be determined by looking at the B4ZBLOOD, B4ZURINE and/or B4ZSALIV.
   b) If the final evening void before bed, or the first void on rising are missed or a significant portion of the urine sample were missed for some other reason the 12-hour catecholamine values (Norepinephrine, Epinephrine, Dopamine) were not computed.
   c) Saliva Duplicates: Initially, each saliva sample was assayed in duplicate. In January 2007 we stopped running assays in duplicate after it was determined that that the cortisol assay provided high quality, reliable results. Thus, the first 621 cases have saliva cortisol data for 2 samples, while the remaining cases have data for a single sample. An "average" value was created for all cases to provide a single consistent variable for each sampling time that could be used in analysis.

Data Inconsistencies

1. Pittsburgh Sleep Questionnaire: "If you have a roommate or bed partner, ask him or her how often during the past month you have had… Other restlessness while you sleep?" (Q10e). If a response other than "Not during the past month" was circled, then a response should have been recorded on the "Please describe:" line. Sometimes it appears as if the respondent,

rather than the respondent's Bed Partner/Roommate, provided this information. For example, the following responses raised this question for data staff.

"unable to remain sleeping for long periods of time"
"Sore shoulders - can't seem to find comfortable position"
"overactive bladder, tossing and turning - having to sleep w/husband"

Staff preparing data for entry flag responses to this item that appear problematic by instructing data entry staff to enter '7 – Don't Know' at B4S10E. During data cleaning, decisions about these cases are made on a case-by-case basis. In some instances, B4S10E is recoded from "7" back to the original response circled by the respondent, in others, B4S10E will be recoded to "1 – Not during the past month'.

2. Physical Exam Vision Assessment (Item 3b): Visual acuity (corrected and uncorrected) is measured from a single distance. Consequently, if an individual wears glasses to correct for distance vision, their corrected vision could be worse than their uncorrected vision. If an individual has no vision or light perception these variables are coded as "INAPP". For details about protocol for measuring visual acuity, see the Physical Exam documentation.

Other Specify & Open-Ended Items – Text Coding

All of the paper and pencil instruments include places to record responses to open-ended questions (i.e. have you ever had surgery, do you engage in regular exercise) or requests to 'Please Specify' or 'Please Describe' when the respondent gives a response in the category "Other". This information was entered but the text responses are **not** included in the public data. Coding has been completed for most of these items. Coded text data are included in the public dataset in one of the following ways.

1. New categorical variables are created: Participants are often asked to report on specific events or describe situations in an open-ended format. When this text data is coded, a new categorical variable is created so that the data can be used in analyses. To distinguish the new variable from the source variable a single character is added to the end of the source variable name. For example, in the Medical History:
    a. Simple Falls section includes text variables describing 'Which' bone was broken (B4O1A1) and the 'Circumstances' of the fall in which the bone was broken (B4O1A2). The new categorical variables corresponding to these variables are named B4O1A1W and B4O1A2C, respectively.
    b. Current Health Screenings: in this section participant are asked if they have ever had cholesterol checks etc. and, if yes, asked to report the results. The results have been coded and the new categorical variables corresponding to the source variables have a 'C' at the end. For example, coded cholesterol results are reported as B4H45ARC.
2. Extant Codes for a related variable are expanded: This is done when specify "Other" options were included in a predefined set of categories. The text responses specified were reviewed and new categorical codes created if appropriate.
    a. For example, in the Medical History participants are asked how often they go to the dentist and several options, including an "Other Specify". Review of the specify responses revealed other categories of frequency (i.e. 3 times a year, 4 times a year, etc.) that were incorporated into the existing categories. See B4H43A variable.

Details about coded text variables are provided in the documentation file ‘M2_P4 Documentation for Coded Text Responses’.