WHI Data Preparation for Investigator Datasets

Updated January 2006

1. Introduction

The WHI Investigator data website includes baseline data for all observational study and clinical trial participants. Outcomes and other information collected during the follow-up period are also included for the observational study participants. Additions to the follow-up information and outcomes updates will be made periodically.

2. Data File Setup

Each data set is provided in a separate fixed length space-delimited ASCII file. The code needed to create SAS data sets from the ASCII files is provided in the files with the .SAS extension. To read the ASCII files into any other statistical program, refer to the INFILE statement in the SAS code file for the order of the variables and to the PROC FORMAT section for the values of all categorical variables.

All data files do not have the same number of records since not every form was completed by each participant. When multiple screening forms were submitted for a participant, we have included the form with the latest date. The first variable in each file, called ID, is the unique participant identifier that replaces the WHI Member ID. All files are linked by this identifier which MUST be used to merge the data files. The order of the variables after ID matches the order of the questions on the most recent version of the form. Computed variables have been added at the end of the appropriate form. The form questions used in the computation of the computed variables have been noted in the variable descriptions; if you would like a copy of the SAS code used to create a variable contact statinfo@whi.org. For confidentiality reasons, individual clinical centers are not identifiable.

Each variable has a unique name ranging from three to eight characters long. In general, the following extensions were used:

AG	= age
DAYS or DY	= days
EVR	= ever
LST	= last
NUM	= number
NW	= now
OTH	= other
REL	= relative
Y	= year

3. Data Conventions

Dates

No actual dates are included in the data files. All dates have been converted to the number of days since randomization for clinical trial participants or since enrollment for observational study participants. When only the month and year were recorded, the first day of the month was used to convert the date. A negative number of days indicates the date occurred before randomization or enrollment. Likewise, a positive number indicates occurrence after randomization or enrollment.

A small number of screening forms for required tasks have encounter dates after the date of randomization or enrollment. We assume these dates reflect edits to the data after the actual randomization or enrollment occurred.

Data Edits

At data entry, the built-in features of WHILMA prevented entry of most invalid or impossible data values for all categorical variables. Broad range checks applied to continuous variables have set out-of-range responses to missing. There still may be values that appear extreme; **it is up to the user to examine all data before proceeding with data analysis.**

Consistency checks between data items on different forms were not done. Therefore, discrepancies do exist. For example, history of breast cancer was collected on both Form 2 and Form 30 and the two data items do not agree exactly. Again it is up to the user to carefully examine the data and determine which values are most appropriate for the specific analyses.

Form Versions

The versions of the data collection forms have changed over time and questions on the forms have been added, deleted, re-ordered and/or modified. To prepare the data for analysis, all questions on each form version were compared to determine if they could be combined into one variable for analysis. In some cases, versions have not been included in the final variables because of incompatibility or because a question was not asked on an early version of a form. This is noted in the data dictionary under usage notes. The text of the question in the data dictionary refers to the latest version of the form. The latest version is assumed to be the final version at the time of this data release.

Missing Data

Missing data can result from a form not being required, a required form not being completed, a particular question on a form not being answered or not required because it was part of a skip pattern, or a question not being asked on all versions of a form. If an entire form is missing for a participant, that participant does NOT have a record in the data file. Missing values in the data files are represented by a single period (“.”). The data dictionary gives the number with missing values for all categorical variables. The

frequency of missing values could be due to any of the reasons listed above. These frequencies should be confirmed before using the data.

Skip Patterns

In general, the same skip pattern coding rule has been applied to all data items. If a sub-question is answered inappropriately based on the main question response, it is set to missing. For example, if a sub-question should be answered only if the main question is answered YES, but the main question is answered "No" or "Don't know" or "missing", the sub-question has been set to "missing". If a question is a sub-question, it has been noted as such in the data dictionary. Referring back to the current form should also clarify the question flow. A few exceptions have been made when a large percentage of participants answered the sub-question even though their response to the main question indicates they should have skipped the main sub-question. In these instances, the data in the sub-question was left as is. These exceptions are noted in the usage notes.

Mark-All-That-Apply

Questions involving "mark all that apply" responses have been recoded. Each possible response has been turned into a yes/no variable with a "yes" coded if the response was marked and "no" otherwise. If all possible responses for the question were missing, all possible responses are set to missing. For example, question 16 on Form 20 (medical insurance information) has seven possible responses (codes 1-6 and 8). Seven "yes/no" variables have been created for each participant. If a participant marked 3=Medicare and 8=Other, the variables for the "Medicare" category and "Other" category are coded as "yes", and the variables for the remaining categories are coded as "no".

4. Baseline Data

The release of WHI baseline data includes all data with the exception of character (write-in) data from Forms 2/3, 20, 30-32, 34, 36/37, 39, 42, 43, 60, 80, 81, 82, 84, 85, 90 and 92 for all participants randomized to one of the clinical trials or enrolled in the observational study. Also included is a set of computed variables that have been commonly used in data analyses of the baseline data. The description of each of these variables in the data dictionary starts with the words "Computed Variable". The baseline data is taken from the August 31, 1999 database.

Current Supplements (Form 45) Data

Data from Form 45 include daily nutrient intake from multivitamins and single supplements and types of supplements taken. The supplement data has been split into two data files as follows: a) nutrients from all supplements, b) types of supplements.

The average intake per day from combination and/or single supplements for 25 nutrients has been calculated. The units of measure for these nutrients match those of the dietary nutrients calculated from the FFQ so that the variables can be summed to yield current

nutrient intake from diet and supplements. In calculating these nutrients, the sum has been taken across all types of supplements which can result in extraneous values. After examining the distribution of the nutrient, it may be necessary to truncate extreme values before analysis.

The second file consists of a set of yes/no variables that provide information on the types of supplements taken. For each of the 25 nutrients, a variable was created that indicates if the participant was taking a single supplement containing that nutrient. In addition, variables indicating use of any type of supplement, multivitamins with or without minerals, stress tabs or other combination supplements are included.

FFQ (Form 60) Data

Data from Form 60 include over 100 nutrients that are calculated from participant responses to the FFQ. These nutrient measures are estimates of average daily intake from foods and beverages. Nutrient intake from vitamin and mineral supplements are not included in these totals. Although we provide all nutrients available from the University of Minnesota Nutrition Coding Center nutrient database, there are substantial differences in the reliability of these measures as estimated from an FFQ, where some measures are considered fairly reliable (e.g., percent energy from fat) and others are clearly unreliable (e.g., selenium). For additional information on the WHI FFQ, see: Patterson RE, Kristal AR, Carter RA, Fels-Tinker L, Bolton MP, Agurs-Collins T. Measurement characteristics of the Women's Health Initiative food frequency questionnaire. Annals Epidemiol 1999;9:178-97.

In addition to the nutrient estimates, there are three summary food variables: total servings per day of fruits, vegetables, and grains. The raw FFQ data (e.g., adjustment question responses, frequencies of consumption, and portion sizes) are not included in this database.

The nutrient data has been split into four data files, grouped as follows: a) energy, macronutrients, cholesterol, caffeine, fiber, fruits, vegetables and grains; b) vitamins, minerals and carotenoids; c) individual starches, sugars and amino acids, oxalic and phytic acid, and ash; d) individual fatty acids. There is one record in each file for each participant who completed a screening FFQ. All nutrient measures for participants with total energy (kcal) less than 600 or greater than 5000 have been set to missing because these energy intake estimates suggest that participants did not complete the FFQ in a reasonable manner.

Blood Results: CBC

The data file named “CBC” includes the results from serum collected at a screening visit and analyzed at each CC’s local laboratory. All clinical trial and observational study participants were to have serum collected. Data is missing if the lab was unable to process the sample. Values were reported for the following tests: white blood cell count (Kcell/ml), platelet count (Kcell/ml), hematocrit (%) and hemoglobin (gm/dl).

Broad range checks have been applied to the CBC results to exclude biologically implausible values. Extreme values and inconsistencies between results (i.e. hemoglobin and hematocrit) may still exist. **Careful inspection of the data is recommended before using these results in analyses.**

Bone Densitometry Results: BMD

The data file named "BMD" includes results from the DXA scans performed during screening at the three Clinical Centers participating in the WHI Osteoporosis substudy. The participating centers are located in Birmingham, Pittsburgh, Tucson and Phoenix, the Tucson satellite site. Participants with valid results from either a hip, spine or whole body scan are included in the data file. These data have been analyzed and monitored by the UCSF Bone Density Center before being transferred to the CCC.

In the most recent UCSF DXA QA Report (November 2005), several recommendations were made regarding the data to be used for analysis. They recommended longitudinal and scanner upgrade corrections and provided the necessary correction factors. These corrections apply to the TOTAL HIP BMD and TOTAL SPINE BMD only. No correction factors were computed or provided for the corresponding BMC and area values. In the BMD data file, we have included both the uncorrected and corrected BMD results for these two locations.

It was also recommended that "all statistical models with BMD as a dependent variable include scanner (identified by serial number) as a covariate to account for the slight calibration differences between scanners." Variables for the scanner serial numbers have been included in the data file, and can be identified by the SAS variable names HIPQDR, SPNQDR, and WHLQDR.

Previous releases of the BMD data included corrections to TROCHANTER BMD and INTERTROCHANTER BMD. These values are no longer corrected in the current data set, per the recommedations from UCSF.

Blood Results: Core Analytes

The "CORE" data file contains the baseline results from the subsample of participants selected at random for blood specimen analysis. The analytes examined include micronutrients, clotting factors, hormones and lipoproteins. The subsample includes approximately 8.6% of the HRT and 4.3% of the DM participants. **Because the subsampling incorporated oversampling of minorities, it is recommended that all analyses using these data either weight the reporting of means by the overall OS race/ethnicity distribution, or include race/ethnicity as a covariate in any modeling.** Also included in the data file are the baseline results from the participants in the Observational Study Measurement Precision Study (OS-MPS). This is approximately 1% of the OS.

ECG Results

The data file named "ECG" includes baseline results for all clinical trial participants with a baseline ECG. ECGs were not performed on the observational study participants.

5. Observational Study Follow-up and Outcomes Data

Follow-up Data

The observational study (OS) participants have now completed annual visits 1 and 3 (AV1 and AV3). All data items from the OS Follow-up Questionnaires from these two visits are now added to the WHI Investigator Data website. Two separate files have been created. F48_AV1 includes the data from AV1 and F143_AV3 the data from AV3. These two data files include all data collected through February 29, 2004. In addition to the data items from the forms, additional computed variables are included for each form. The set of variables includes constructs or summary variables that are comparable to those included with the baseline data release. For example, the same baseline physical activity variables computed from Form 34 (Personal Habits), have been computed again based on Form 143 data to provide the same physical activity information at AV3.

A set of questions on hormone use are included on each OS follow-up form. These questions on Form 48 (AV1) changed between version 1 and 2 of the form in a way that prevents mapping the variables between the two versions. As an example, questions on estrogen use on version 1 do not distinguish between a combined pill and a pill that includes estrogen only. For this reason, only the questions from version 2 of Form 48 are included in the file F48_AV1. These questions are compatible with the hormone use questions on all subsequent OS follow-up forms. It was possible, though, to compute overall summary variables from both versions of Form 48, reporting any estrogen use, any progesterone use and any hormone use. These variable are on the file F48_AV1.

To be consistent with the baseline hormone use variables computed from the Form 43 data (Hormone Use), only hormone use from pills and patches are considered in the OS follow-up hormone use summary variables.

Outcomes

The second release of OS outcomes data includes centrally verified, locally verified and self-reported outcomes collected through February 29, 2004. The new data file and documentation replaces the existing "OUTOS" files. Verified outcomes include all cancers, hip fracture and all cardiovascular outcomes except DVT and PE. The outcomes for which central adjudication is required for all OS participants are hip fracture, and in situ breast, invasive breast, colon, endometrium, ovary, rectosigmoid junction and rectum cancers. If the central adjudication was closed as of February 29, 2004, the central

adjudication result was used; otherwise the local adjudication was used. Self-reported outcomes included are all non-hip fractures, and those routinely reported in Table 5.8 of the February 29, 2004 Semi-Annual Progress Report. For each outcome, two variables are provided: one indicates the occurrence of the outcome since enrollment, and the second variable provides the number of days from enrollment to the **first occurrence** of the outcome. Additionally, for each centrally verified outcome, a third variable was added that indicates if the outcome was verified centrally or locally.

A few of the self-reported outcomes were not included on early versions of Form 33. In addition, when Form 33D was initiated, information on fractures was moved from Form 33 to Form 33D, and the list of fractures was expanded. Specifically, leg was split into lower leg, knee and upper leg, and new categories for pelvis, tailbone and elbow were added. There were also additions to the list of locally verified cancers on later versions of Form 122. Outcomes affected by these form changes have been noted in the data dictionary for the OUTOS data file. Over 98% of the OS participants reported outcomes on version 3 or higher of Form 33, so these form changes affect few outcomes.

Four verified outcomes have a "subsequent condition" rule (angina, TIA, carotid artery disease, and in situ breast cancer). This rule means that an angina occurring on the same date or after an MI is not counted as an outcome. The same rule applies to a TIA or carotid artery disease occurring on the same date or after a stroke. In addition, we do not count an in situ breast cancer that occurs on the same date or after an invasive breast cancer.

Information on death and last contact is also provided. All deaths occurring through February 29, 2004 have been included even if they have not yet been adjudicated. Those deaths not yet adjudicated do not have a cause of death. The date of a participant's last Form 33 or 33D is considered their date of last contact for outcomes collection. **When performing time-to-event analyses, the days from enrollment to the minimum of death or last contact should be used as the censoring time for those participants without the event. If death is the event of interest, the censoring time would be just the days from enrollment to last contact.**

A small number of participants (n=669) have no Form 33 or 33D in the study database. These participants have missing values for the outcomes reported on Form 33D and last contact date. An additional 36 participants have a Form 33D but no Form 33 after enrollment. These participants have missing values for the outcomes collected from Form 33 only (n=705). Participants with nor Form 33, 33D or other outcomes forms (Form 121, 122, 123, etc.) will have missing values for all adjudicated outcomes.

Other Data

A set of baseline medication use variables from Form 44 (Current Medications) is planned for a future release, as well as additional data collected at the OS Annual Visit 3.