WHI Data Preparation for Investigator Data Sets

Updated November 30, 2006

1. Introduction

The WHI Investigator data sets available on the study website include most baseline data for all Observational Study (OS) and Clinical Trial (CT) participants. Outcomes and other selected data sets collected during the follow-up period are also included for OS and CT participants. Additions to the follow-up and outcomes data sets will be made periodically.

Data in the new and updated files in this release are current as of September 12, 2005. Data in these files may differ slightly from previously released data due to edits made between the time of the previous release and September 12, 2005.

New data sets in this release

Forms 10 & 50 - HRT Management and Safety Interview, Report of Vaginal Bleeding
Form 17 - CaD Management and Safety Interview
Form 35 - Personal Habits Update
Form 83 - Transvaginal Uterine Ultrasound

Updated data sets in this release

Demographics (five variables added; see section 4)
BMD (CT follow-up added; reorganized into three files; see section 4)
Form 33 - Medical History Update (CT added)
Form 37 - Thoughts and Feelings (CT+OS follow-up added)
Form 38 - Daily Life (CT follow-up added)
Form 39 - Cognitive Assessment (follow-up added)
Form 44 - Current Medications (CT follow-up added; combined into one file)
Form 60 - FFQ (CT follow-up added; reorganized into four files)
Form 80 - Physical Measurements (CT follow-up added; combined into one file)
Form 81 - Pelvic Exam (follow-up added)
Form 82 - Endometrial Aspiration (follow-up added)
Form 84 - Breast Exam (follow-up added)
Form 85 - Mammogram (follow-up added)
Form 90 - Functional Status (follow-up added)

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 (sometimes referred to as the "common 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. For form-based data sets, the order of the variables after ID matches the order of the questions on the most recent version of the form. Computed variables based on form responses have been added at the end of the appropriate form data sets. 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 fifteen 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 the study database application 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

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. Specific Data Set Information

Demographics

Five new variables have been added to the demographics data set for this release: "AGESTRAT" is the age stratum to which the participant was randomized or enrolled; “DMARM” and “CADARM” are the study arms to which DM and CaD participants were randomized; “CADDAYS” is the number of days since CT randomization to CAD randomization; “BMDFLAG” indicates whether a participant is in the BMD subsample (i.e. was randomized at a bone density site).

Computed Variables

Many computed variables that have been commonly used in data analyses are included in various data sets. In general a computed variable resides in the data set which contain the variable(s) from which it was computed. The description of each of these variables in the data dictionary starts with the words “Computed Variable”.

Current Medications (Form 44)

Included with the Current Medications data file are a number of reference files, including a PDF called F44_ReadMe.pdf. The F44_ReadMe document provides further details about the collection and analyses of Current Medications data.

Current Supplements (Form 45)

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 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.

The raw FFQ data (e.g., adjustment question responses, frequencies of consumption, and portion sizes) are not included in this data set.

The nutrient data has been split into four data files, grouped as follows: a) energy, macronutrients, cholesterol, caffeine, fiber, fruits, vegetables, glycemic load; b) vitamins, minerals and carotenoids; c) individual starches, sugars and amino acids, oxalic and phytic acid, and ash; d) individual fatty acids and isoflavones. Consider excluding all nutrient measures for participants with total energy (kcal) less than 600 or greater than 5000 as these energy intake estimates suggest that participants did not complete the FFQ in a reasonable manner.

There are a number of vitamin A related variables in the WHI nutrient data set that use different units. Investigators using the data set are advised to refer to the usage notes included in the variable description report to decide which vitamin A variable(s) to use in manuscript analyses.

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 BMD data have been reorganized into three files by scan type: Hip, Spine and Wholebody. Each file contains baseline and follow-up scans for both CT and OS. DXA scans were performed 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 a hip, spine or whole body scan are included in the data files. 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 for the following values:

Total hip BMD
Total spine BMD
Whole body BMD
Whole body BMC
Whole body total mass
Whole body total fat
Whole body total percent fat
Whole body total lean
Whole body total fat free mass
Whole body total area

In addition, a new computed variable called "Total spine BMD (L2,L3,L4 BMD values are known)" has been included. This value is equal to the *corrected* total spine BMD value if the BMD values of L2, L3 and L4 are all known and is set to missing if any of L2, L3 or L4 are missing.

We have included both the uncorrected and corrected values in the BMD data files.
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.

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.

In certain situations, the change in BMD or other DEXA variables between two time points is invalid. Do not compute change if:

1. The two scans were done on different machines, except for calibrated scanner upgrades. Changes are okay between QDR 2392 and QDR 47606, and between QDR 2412 and QDR 49454.

2. The two hip scans were done on different sides of the hip (HIPSDSCN).

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.

Observational Study Follow-up Questionnaires

The OS follow-up data sets include all data items from OS Follow-up Questionnaires for years 3 through 8 (Forms 143 through 148) and Form 149, "Supplement to OS Follow-Up Questionnaire". These data sets are based on data collected through September 12, 2005.

In the April 12, 2006 release, the variable "Alcohol servings per week" was moved to the Form 60 data sets. The Form 60 data sets are a more appropriate location for this variable because it is derived entirely from Form 60 data. The variable "Contact Type" was removed for consistency with the other OS follow-up data sets.

Please note that Form 149 was not necessarily collected at the participants' year 9 anniversary as the name might imply; rather it was collected from participants who did not reach year 7 by the close-out contact. Form 149 was collected during the close-out year only.

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 physical activity variables computed at baseline from Form 34 (Personal Habits) have been computed again based on the 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 variables 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 third release of OS outcomes data includes centrally verified, locally verified and self-reported outcomes collected through September 12, 2005. 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 September 12, 2005, 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 5.5 of the September 12, 2005 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 September 12, 2005 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 death (or the last contact if no death occurred) 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=623) 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 37 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=660). Participants with no Form 33, 33D or other outcomes forms (Form 121, 122, 123, etc.) will have missing values for all adjudicated outcomes.

6. Choosing forms for analysis when there are multiple forms per participant

In most of the data files there are multiple rows of data for a single participant. When using these files you will need to be careful when selecting rows to use in your analyses. We recommend that you consult "whi_data_collection_frequency .pdf" (included in your data set download) before analyzing data collected at multiple visits.

First it is important to understand the definition of a few variables included in most of the follow-up data files: days since randomization/enrollment, visit type, visit number, closest to visit within visit type and number, and expected for visit.

Days since randomization/enrollment

Days since randomization/enrollment is calculated by subtracting the date of CT randomization or OS enrollment from the date on the front of the form. For example, on Annual Visit 3 forms you would expect this variable to be somewhat close to 1095 (3 years * 365 days/year).

Visit Type

On the front of all forms there is a place for the Clinical Center to enter the Visit Type for which the form corresponds.

1 - Screening
2 - Semi-Annual
3 - Annual
4 - Non-Routine
5 - 6 Week HRT/4 Week CaD Call

6 - Diet Intervention (used for Diet Intervention sessions)
7 - Interim (briefly used on Form 33)
8 - Amendment (briefly used on Form 33)

For Annual Visit 3 forms you would expect this variable to be “3”.

Visit Number
On all forms there is a field for the Clinical Center to enter the number of the visit at which the form was collected. For Non-Routine and 6 Week HRT/4 Week CaD Call Visit Types, a visit number is not required and is set to missing. Except for Form 44 – Current Medications data, the visit number for a screening visit type is set to zero. In the Form 44 data file it is left as entered by the Clinical Center, because data from more than one screening visit can exist in the file.

The Visit Number for Semi-Annual contacts should be coded as follows:
1 - for semi-annual contacts 6 months following randomization,
2 - for semi-annual contacts 18 months following randomization,
3 - for semi-annual contacts 30 months following randomization,
etc

The Visit Number for Annual contacts should be coded as follows:
1 - for annual contacts 12 months following randomization,
2 - for annual contacts 24 months following randomization,
3 - for annual contacts 36 months following randomization,
etc.

For Annual Visit 3 forms you would expect this variable to be “3”.

Visit Numbers greater than 12 were considered out of range and have been set to missing in the data files.

Closest to Visit within Visit Type and Number
This variable is useful for Visit Types “2-Semi-Annual” and “3-Annual”. There are instances where a Clinical Center entered the same form with the same visit type and number for the same participant. To handle these cases this variable (or “flag”) is included in many of the data files. The flag indicates the form that is closest to the target visit date for the Visit Type and Number entered on the form (the target visit date for a participant’s Annual Visit 1 form would be their randomization/enrollment date + 365 days). The flag is only included in data files where it is deemed useful. It is not included where the data file is limited to one form per visit, or where looking at all rows makes the most sense (e.g. Form 33). With the exception of Form 44 data, screening visits have a value of “1” for the Closest to Visit within Visit Type and Number variable.

To demonstrate how the Closest to Visit within Visit Type and Number is calculated, some Form 80 examples are presented below:

Example A:

Participant	Days since	Visit	Visit	Closest to visit within Visit Type

Id (ID)	Randomization/ Enrollment (F80DAYS)	Type (F80VTYP)	Number (F80VNUM)	and Number 0 = No, 1 = Yes (F80VCLO)
100000	365	3	1	1
100000	730	3	1	0
100000	1095	3	3	1

In Example A above, the Clinical Center coded two Form 80s as an Annual Visit 1. The one closest to the Annual Visit 1 target date is coded with a 1 while the other one (which is closest to an Annual Visit 2) is coded with a 0.

Example B:

Participant Id (ID)	Days since Randomization/ Enrollment (F80DAYS)	Visit Type (F80VTYP)	Visit Number (F80VNUM)	Closest to visit within Visit Type and Number 0 = No, 1 = Yes (F80VCLO)
100001	365	3	1	1
100001	365	3	2	1
100001	1095	3	3	1

In Example B above, the Clinical Center coded two Form 80s with the same date, but with different visits. Because there is only one form per visit type and number, each one is flagged with a 1 for F80VCLO.

Example C:

Participant Id (ID)	Days since Randomization/ Enrollment (F80DAYS)	Visit Type (F80VTYP)	Visit Number F80VNUM)	Closest to visit within Visit Type and Number 0 = No, 1 = Yes (F80VCLO)
100001	365	3	1	1
100001	365	3	1	0
100001	700	4	2	0
100001	800	4	3	0

In Example C above, the Clinical Center coded two Form 80s with the same date and visit. One of these is flagged with a 1 and the other with a 0 for F80VCLO. In this case, the flag is based on a timestamp in the database which indicates the form most recently entered (the timestamp is not available in the data file). The form entered most recently is flagged with a 1 while the other is flagged with a 0.

Also notice that the Non-Routine visits are flagged with a 0. This is true of all Non-Routines, because the flag is only valid for Semi-Annual and Annual visits, where a target date can actually be calculated.

Expected for Visit
This variable indicates if the form/data was expected for the Visit Type and Visit Number entered on the form. According to protocol, forms were to be collected at specific visits. For example Form 35 – Personal Habits Update was to be collected for all CT at Annual Visits 1, 3, 6, and 9. It is possible that the Clinical Center collected the form at an Annual Visit 4, but it was not expected at that visit.

Putting it all together to select data rows for analyses

There are two basic ways in which to select rows of data for analyses:

1. By visit type and number (technique used most often by CCC)
You can choose to select rows for analyses by using visit type and visit number; and breaking duplicates using the Closest to Visit within Visit Type and Number flag.

To pick all Annual Visit 1 Form 80s from a Form 80 data file you could restrict the rows in the file to the following:

```
F80VTYP =  3 and F80VNUM = 1 and F80VCLO = 1
```

Note that this will miss all the Semi-Annual Visit 1s and 2s. These could possibly be an Annual Visit 1 where an Annual Visit 1 is missing for a participant. If a participant's Annual Visit 1 is missing, but they have a Semi-Annual Visit 1 or 2, you could choose to use data from one of those visits instead.

To pick all Form 80s expected for a visit from a Form 80 data file you could restrict the rows in the file to the following:

```
F80VCLO = 1 and F80EXPC = 1
```

2. By days since randomization/enrollment
You can choose to select rows for analyses using days since randomization/enrollment. In this case you will have to pick a range in which you consider a visit to be valid, for example you may say I will consider any form done within 180 to 545 days of randomization/enrollment to be an AV1. This range will probably change depending on the interval in which the form is collected. If there is more than one form that falls into the range, you will have to come up with an algorithm to pick the one to use. You can limit by picking the one closest to the target visit for which you are selecting. You can limit based on Visit Type and Number, and within that by Closest to Visit within Visit Type and Number.

You can use the two techniques above in combination as well. You may decide to use the By Visit Type and Number mechanism, but throw out rows which seem to be out of the date range. For example:

```
F80VTYP = 3 and F80VNUM = 1 and F80VCLO = 1 and F80DAYS < 520
```

Basically, how you choose data rows within a data file needs to be based on your analysis objectives.

Before starting any data analyses, it is imperative that you check to make sure you have the desired number of records per participant, and per visit if applicable.