WHI Observational Study Limited Access Data Release
Data Preparation Guide

Introduction

This release of the WHI Observational Study baseline and follow-up data includes data collected on study forms, outcomes, results from blood analyses, and computed variables that have been commonly used in data analyses.

Data File Setup

Each data set is provided as a separate fixed length space-delimited ASCII file. The code needed to create a SAS data set from the ASCII file is also provided. Data sets can be found in the *data\whi\ascii* directory. Each data set is zipped up into a .ZIP file that includes the .DAT raw data file, and the .SAS code file to create the SAS data set. To read the ASCII file 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.

To use the data you will first need to unzip the .ZIP file for the version of the data set you wish to use.

All data files do not have the same number of records since not every form was completed by each participant. Data files for forms that were used at baseline and follow-up will include both baseline and follow-up data. When multiple baseline forms were submitted for a participant, we have included the baseline form with the latest date. The first variable in each file, called ID, is the unique participant identifier that replaces the WHI participant 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. In general, 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. For confidentiality reasons, individual clinical centers are not identifiable.

Each variable has a unique name ranging from two 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

Data Conventions

Dates

No actual dates are included in the data files. All dates have been converted to the number of days since enrollment. 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 enrollment. Likewise, a positive number indicates occurrence after enrollment.

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

Data Edits

The built-in features of the data entry system 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. For some variables, like height, weight, and waist circumference the values outside the 1^{st} and 99^{th} percentiles are truncated to the 1^{st} and 99^{th} percentiles. Where this truncation occurs, it is noted in the variable documentation. 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”.

Specific Data Set Information

Thoughts and Feelings (Form 37) Data
The Thoughts and Feelings form was only collected at baseline for OS, but was collected at baseline and close-out for CT participants. Additional items were added to the form used for close-out. The additional data items added to the close-out form appear in the OS Form 37 data file, but those data items have missing values, as OS participants never completed a close-out Form 37.

Current Medications (Form 44) Data

Included with the Current Medications zipped up data set 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

Data from Form 45 include daily nutrient intake from multivitamins and single supplements and types of supplements taken. 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. 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.

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 dataset that use different units. Investigators using the dataset 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.

Observational Study Follow-up Questionnaires

The OS follow-up data sets include all data items from OS Follow-up Questionnaires for year 1 (Form 48), years 3 through 8 (Forms 143 through 148), and Form 149 (Supplement to OS Follow-Up Questionnaire). No OS Follow-up Questionnaire was collected at year 2.

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.

The OS Follow-up Questionnaire data files contain one record per participant responding. The participant record included in the Form 48 and Forms 143 through 148 data files is the one closest to the associated form's visit target date. For example, a participant record in the F144_AV4 file is the Form 144 returned closest to that participant's AV4 target date. For Form 149, the latest Form 149 returned by the participant is included.

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 for baseline. 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 in 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.

Blood Results: CBC

The CBC data file includes the results from serum collected and analyzed at each CC's local laboratory. All observational study participants were to have serum collected at baseline and year 3. 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 files include results from the DXA scans performed at the Clinical Centers participating in the WHI Osteoporosis substudy. 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 DXA Quality Assurance 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

Only the corrected values have been included in the BMD data files.

It was also recommended that "all statistical models with BMD as a dependent variable include scanner (identified by Scanner ID) as a covariate to account for the slight calibration differences between scanners." Variables for the Scanner Ids have been included in the data file, and can be identified by the SAS variable names HIPSID, SPNSID, and WHLSID.

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 Scanner 3 and Scanner 4, and between Scanner 2 and Scanner 6.

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

Outcomes

The outcomes data includes centrally verified, locally verified and self-reported outcomes. The outcomes data has been split into two data files: a) self-reported outcomes for OS; and b) adjudicated outcomes for OS.

Verified outcomes include all hip fractures, cancers, and cardiovascular outcomes, except DVT and PE. The outcomes for which central adjudication is required for all OS participants are hip fracture, in situ breast, invasive breast, colon, endometrium, ovary, rectosigmoid junction and rectum cancers.

Where central adjudication is required, if the central adjudication was closed as of September 12, 2005, the central adjudication result was used; otherwise the local adjudication was used. In addition, the outcomes occurring after the study close-out date are censored. For OS participants, the close-out date is September 12, 2005.

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. In rare instances, an outcome is reported to have occurred, but the diagnosis date is missing. If this happens, the indicator variable will be coded as 'Yes', but the corresponding 'number of days' variable will have a missing value. Additionally, for each OS centrally verified outcome, a third variable was added that indicates if the outcome was verified centrally or locally.

The variables that indicate the first occurrence of an outcome do not include instances where the first occurrence is determined solely by the cause of death. The number of such outcomes is small since, if at the time of death enough documentation has been collected, a regular adjudication is conducted and the relevant forms are completed. The deaths with insufficient documentation tend to be "out of hospital" deaths. The context of the analysis should determine whether to combine the adjudicated outcomes and those obtained from the death information only. Keep in mind that outcomes defined by the

cause of death only will not have the detailed information as collected on the adjudicated outcomes. For example, estrogen receptor status would not be available for a breast cancer, or whether a stroke was ischemic versus hemorrhagic would be unknown.

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 these data files.

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 before close-out date 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 variable with the days from enrollment to last contact is included in the self-reported outcomes files.

A small number of participants 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. A small number of participants have a Form 33D but no Form 33 after enrollment. These participants have missing values for the outcomes collected from Form 33. Participants with no Form 33, 33D or other outcomes forms (Form 121, 122, 123, etc.) will have missing values for all adjudicated outcomes.

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.

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

Days since randomization/enrollment
Days since randomization/enrollment is calculated by subtracting the date of 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 was 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
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 Year
On all forms there was a field for the Clinical Center to enter the number/year of the visit at which the form was collected. For Non-Routine Visit Types, a visit year was not required and is set to missing. Except for Form 44 – Current Medications data, the visit year 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 Year for Semi-Annual contacts should be coded as follows:
1 - for semi-annual contacts 6 months following randomization/enrollment,
2 - for semi-annual contacts 18 months following randomization/enrollment,
3 - for semi-annual contacts 30 months following randomization/enrollment,
etc

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

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

Visit Years 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 Year

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 year 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 Year 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 Year variable.

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

Example A:

Participant Id (ID)	Days since Randomization/ Enrollment (F80DAYS)	Visit Type (F80VTYP)	Visit Year (F80VY)	Closest to visit within Visit Type and Year 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 Year (F80VY)	Closest to visit within Visit Type and Year 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 year, each one is flagged with a 1 for F80VCLO.

Example C:

Participant Id (ID)	Days since Randomization/ Enrollment (F80DAYS)	Visit Type (F80VTYP)	Visit Year F80VY)	Closest to visit within Visit Type and Year 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 Year entered on the form. According to protocol, forms were to be collected at specific visits. For example Form 80 – Physical Measures was to be collected for all OS at Baseline and Annual Visit 3. 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 year (technique used most often by the WHI Clinical Coordinating Center)
You can choose to select rows for analyses by using visit type and visit year; and breaking duplicates using the Closest to Visit within Visit Type and Year flag.

To pick all Annual Visit 3 Form 80s from a Form 80 data file, you could restrict the rows in the file to the following:

```
F80VTYP = 3 and F80VY = 3 and F80VCLO = 1
```

Note that this will miss all the Semi-Annual Visit 3s and 4s. These could possibly be an Annual Visit 3 where an Annual Visit 3 is missing for a participant. If a participant's Annual Visit 3 is missing, but they have a Semi-Annual Visit 3 or 4, 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 Year, and within that by Closest to Visit within Visit Type and Year.

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

```
F80VTYP = 3 and F80VY = 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.