Data Preparation and Use of WHI Investigator Data Sets
Posted on the Study Operations Website

Updated: September 24, 2010

Table of Contents

1. Introduction

The WHI Investigator data sets available on the Study Operations website include most baseline and follow-up data for all Observational Study (OS) and Clinical Trial (CT) participants, including outcomes, adherence, CT unblindings and core blood analytes. Additions to the follow-up and outcomes data sets will be made periodically.

The WHI study ended March 31, 2005, and the closeout date for data collection was April 8, 2005. Participants consenting to join the WHI Extension Study continue to be followed, primarily for outcomes data collection. This data release includes data sets current as of August 14, 2009. No updates to the previously released baseline or follow-up data from the WHI Study have been made. The outcomes data sets include all WHI participants, and the first occurrence of outcomes since the beginning of WHI. Some outcomes data may differ slightly from previously released data due to edits made between the time of the previous release and August 14, 2009. Section 4 contains additional information on selected data sets. **Substantial changes have been made to the way the outcomes detail information is provided, so please review the Outcomes section carefully.**

1.1 Changes in the current release (September 24, 2010)

- ECG Results – See section 4.11.
- Form 31 – Added the variable "Age at Menopause"

Changes in the previous release (April 30, 2010)

New data sets:

- CaD Breast Cancer Outcomes Detail (Forms 122, 130)
- CaD Cardiovascular Outcomes Detail (Form 121 - not stroke or carotid)
- CaD Cancer Outcomes Detail (Forms 122, 130 - not breast cancer)
- CaD Death Detail (Form 124)
- CaD DVT/PE Outcomes Detail (Form 126)
- CaD Fracture Outcomes Detail (Form 123)
- CaD Stroke/Carotid Outcomes Detail (Forms 121, 132)
- CT+OS Breast Cancer Outcomes Detail (Forms 122, 130)
- CT+OS Cardiovascular Outcomes Detail (Form 121 - not stroke or carotid)
- CT+OS Cancer Outcomes Detail (Forms 122, 130 - not breast cancer)
- CT+OS Death Detail (Form 124)
- CT+OS DVT/PE Outcomes Detail (Form 126)
- CT+OS Fracture Outcomes Detail (Form 123)
- CT+OS Stroke/Carotid Outcomes Detail (Forms 121, 132)
- Extension Study Membership
- Follow-up Status History
- Form 40 - Addendum to Medical History Update (Family History of DVT/PE)
- Form 41 - Addendum to Personal Information (Race)
- Form 134 - Addendum to Medical History Update
- Form 150 - Hormone Use Update (WHI Extension Study)
- Form 151 - Activities of Daily Living (WHI Extension Study)

Updated data sets:

- Form 33X - Medical History Update (WHI and Extension Study)
- Form 85X - Mammogram (WHI and Extension Study)
- CaD Outcomes, Adjudicated
- CaD Outcomes, Self-Reported
- CT+OS Outcomes, Adjudicated
- CT+OS Outcomes, Self-Reported

1.2 Changes to outcomes definitions

The following changes have been made to the outcome definitions from the last data set

- These data now include outcomes collected in the WHI Extension Study. The outcome ascertainment procedures changed for the Extension Study.
    - Some outcomes are no longer adjudicated (e.g. Angina, TIA, CHF, non hip fractures). We censor any occurrences of these outcomes if their diagnosis date is after the WHI closeout.
    - We censor any outcomes whose diagnosis date is after the WHI closeout for participants who did not consent to the Extension Study.
    - All adjudications during the Extension Study are done centrally at the Clinical Coordinating Center (CCC).
    - Central (CCC) adjudication forms are often more detailed than the local adjudication forms. Therefore, the detail data may have changed. These changes are detailed for each variable.
- Coronary Revascularizations –A participant is considered to have a coronary revascularization outcome if they have either a CABG or a PTCA outcome. Previously the definition used the coronary revascularization question on the outcome form and would count the first revascularization that occurred after an MI.
- Strokes – Due to increased interest in classifying strokes, central adjudications will be used whenever available. Previously only HT central adjudications were used.
    - Hemorrhagic and ischemic stroke have been added as separate outcomes
    - Carotid and TIA outcomes are not counted after a stroke. Those outcomes will be affected by changes to the stroke outcome.
- TIA – using central adjudications for outcomes in HT participants

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 helpdesk@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

4.1 Demographics

Five new variables were added to the demographics data set in the 11/30/06 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).

4.2 Extension Study Membership

The variable called "EXTFLAG" indicates if a participant is enrolled in the WHI Extension study. The other variables in the file can be used to determine if an outcome, or study visit occurred during the extension.

4.3 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".

4.4 Addendum to Personal Information - Race (Form 41)

The variables on this form provide race/ethnicity information according to the 2000 U.S. Census. Administration of the form did not begin until 2003, so the variables are not available on all participants. Known discrepancies exist between the Form 41 and Form 2 race/ethnicity questions.

4.5 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.

4.6 Current Supplements (Form 45)

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.

4.7 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.

4.8 Blood Results: CBC

The data file named "CBC" includes the results from blood collected at a screening visit and analyzed at each Clinical Center's (CC) 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.**

4.9 Bone Densitometry Results: BMD

In the 11/30/06 release, the BMD data were 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 computed variable called "Total spine BMD (L2,L3,L4 BMD values are known)" is 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 recommendations 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).

4.10 Blood Results: Core Analytes

The "CORE" data file contains results from the subsample of CT 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 CT race/ethnicity distribution, or include race/ethnicity as a covariate in any modeling.** Also included in the data file are the results from the participants in the Observational Study Measurement Precision Study (OS-MPS). This is approximately 1% of the OS.

4.11 ECG Results

The data set "ecg_ct_inv" contains baseline (one record per ppt) and follow-up (multiple records per ppt) ECG result. It was updated in September of 2010 and now contains 510 additional measurement variables supplied by Epicare.

The data set "ecg_mi_nova_ct_inv" contains serial comparison of baseline and follow-up ECGs for the likelihood of MI using the Novacode classification system.

4.12 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.

4.13 Outcomes

The current release of outcomes data includes centrally verified, locally verified and self-reported outcomes collected through August 14, 2009 for CT, OS and CaD. Summary data is divided into four files: *outc_self_ctos_inv.dat* (CT+OS self reported outcomes), *outc_self_cad_inv.dat* (CaD self reported outcomes), *outc_adj_ctos_inv.dat* (CT+OS adjudicated outcomes), and *outc_adj_cad_inv.dat* (CaD adjudicated outcomes). Detail information collected on verified outcomes are in additional data files by outcome type, separately for CT+OS and CaD.

The adjudication process, set of outcomes adjudicated, self-reported outcomes, and information on screening procedures collected on Form 33 changed at the start of the Extension Study. The following Table describes the changes, as do the "Usage Notes" in the data dictionary files.

	WHI 1993-2005				**Extension 2005-2010**	
Outcome	HT	DM	CaD	OS	CT/OS	Form
CARDIOVASCULAR:						
MI	C	L	L	L	C	121
Stroke	C	L[1]	L[1]	C[1]	C	121/132
Congestive heart failure	C	L	L	L	S	121
Angina	C	L	L	L	S	121
Peripheral artery disease (PAD)	L	L	L	L	C	121
Carotid artery disease (CAD)	L	L	L	L	C	121/132
Coronary revascularization (PTCA/CABG)	L	L	L	L	C	121
TIA	C	L	L	L	S	121/132
CANCER:						
Breast cancer	C	C	C	C	C	122/130
Endometrial cancer	C	C	C	C	C	122/130
Colorectal cancer	C	C	C	C	C	122/130
Ovarian cancer	C	C	C	C	C	122/130
Multiple myeloma	C	C	C	C	C	122/130
Leukemia	C	C	C	C	C	122/130
Lung	C	C	C	C	C	122/130
Lymphoma, Hodgkin's	C	C	C	C	C	122/130
Lymphoma, Non-Hodgkin's	C	C	C	C	C	122/130
Pancreas	C	C	C	C	C	122/130
Other cancers	L	L	L	L	C[2]	122/130
FRACTURES:						
Hip	C	C	C	C	C	123
Non-hip fractures	L	L	L	L[3]	S	123
OTHER:						
Pulmonary embolism	C	S	S	S	C (HT)	126
Deep vein thrombosis	C	S	S	S	C (HT)	126
Hysterectomy	C	S	S	S	C (HT)	131
Death from any cause	C	C	C	L	C	124

C - Centrally adjudicated
L - Locally adjudicated
S - Self-reported

[1] used central adjudication data when available, and local adjudication data otherwise
[2] not SEER coded
[3] done only at the BMD centers

Self-Reported Outcomes for CT/OS Participants

Outcome	WHI 1993-2005	Extension 2005-2010
Condition		
Cataracts*	X	-
Colorectal polyps/adenomas	X	X
Dementia, Alzheimer's	-	X
Diabetes diagnosis, ever	-	X
Fractures (non-hip)	X	X
Gallbladder disease/stones	X	-
Glaucoma	X	-
Kidney/bladder stones*	X	-
Macular degeneration	-	X
Osteoarthritis*	X	X
Osteoporosis	X	-
Parkinson's disease	-	X
Rheumatoid arthritis	X	-
SLE	X	X
Exams/Procedures		
Barium enema X-ray*	X	X
Blood in stool	X	X
Blood pressure*	X	-
Bone density scan	-	X
Breast biopsy/aspiration	X	X
Breast exam	X	X
Breast exam, other (MRI/ ultrasound)	-	X
Cholesterol	X	-
D&C	X	X
ECG	X	-
Endometrial biopsy	X	X
Eye exam*	X	-
Flex sig/colonoscopy	X	X
Hysterectomy	X	X
Mammogram	X	X
PAP smear	X	-
Physical exam*	X	-
Rectal exam	X	X
Medication/treatments		
Anxiety pills	-	X
Depression pills/therapy	-	X
Diabetes mellitus requiring therapy	X	X
Diabetes, diet/exercise	-	X
Diabetes, insulin	X	X
Diabetes, pills*	X	X
Estrogen pills	-	X
High blood pressure pills*	X	X
High cholesterol pills	-	X
Shots for DVT*	X	-
Osteoporosis calcium pills	-	X
Osteoporosis non-calcium pills	-	X

*Not on all versions of Form 33

If the central adjudication was closed as of August 14, 2009, the central adjudication result was used; otherwise if a local adjudication exists, the local adjudication was used. In addition, for participants not enrolled in the Extension Study, any outcomes occurring after the study close-out date are censored. For CT and OS participants, the close-out date is April 8, 2005; for CaD participants, close-out date is the earliest of the unblinding date or April 8, 2005.

For each adjudicated outcome, three variables are provided in the summary files: one indicates the occurrence of the outcome since enrollment, the second variable provides the number of days from enrollment to the **first occurrence** of the outcome, and the third indicates if the outcome was verified centrally or locally, or comes from the cause of death only. For the CaD trial outcomes, the 'number of days' variable indicates the number of days since the CaD randomization date. 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.

Self-reported outcomes included are all non-hip fractures, those outcomes routinely reported in Tables 5.5 and 6.3 for OS and CT, respectively, in the August 14, 2009 Annual Progress Report, and new outcomes added to Form 33 in the extension.

A few of the self-reported outcomes were not included on early versions of Form 33. Others were dropped on subsequent versions. 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.

Eight verified outcomes have a "subsequent condition" rule (angina, CABG, PTCA, possible silent MI, definite silent MI, TIA, carotid artery disease, and in situ breast cancer). This rule means that an angina, CABG or PTCA occurring on the same date or after a clinical 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. A possible or silent MI determined from the ECG data occurring on the same date or after a clinical MI is not counted, nor is a possible occurring after a definite silent MI. For CaD data, outcomes are counted after CaD enrollment, and the subsequent condition takes affect after CaD enrollment date; that is, outcomes that occur prior to CaD enrollment are not taken into account for the subsequent condition rule.

Information on death and last contact is also provided. All deaths occurring before August 14, 2009 have been included even if they have not yet been adjudicated. Those deaths not yet adjudicated do not have a cause of death and are not included in the death detail files. 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 (LAST33DY) is included in the self-reported outcomes files.

A small number of OS/CT participants (n=878) have no Form 33 or 33D in the study database. Also, a small number of CaD participant (n=57) have no Form 33 or 33D after CaD enrollment. These participants have missing values for the outcomes reported on Form 33D and last contact date. For CT/OS, an additional 45 participants have a Form 33D but no Form 33 after enrollment; for CaD, the number is 5. 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.

4.13.1 Detail Files

There are separate detail files for the main outcomes disease types: breast cancer, cancer (non-breast), cardiovascular (non-stroke/carotid), death, DVT/PE, fracture, and stroke/carotid. If a participant has had at least one adjudicated event of a disease type, they will have a record in the corresponding data file. These files include an indicator variable for each outcome that matches the same variable in the outcomes summary files. The corresponding "Ascertainment Source" variable in the summary file indicates which form provides the detail information. However, if the source is "cause of death", no detail information is available except that related to the death in the Death detail file.

The structure of each detail file differs slightly, but in general, if a participant had greater than one event reported on the same form, the record in the detail file will include variables for all events. Check the "Usage notes" for clarification. The

most straightforward way to use the detail files is to select the variables for one outcome type at a time. Refer back to the appropriate outcomes form to help identify the variables related to the outcome of interest.

The following example demonstrates how to obtain detail information on participants with a stroke. Of interest are the Oxfordshire and TOAST classifications available from the central adjudication, and the question reporting the type of stroke available on both locally and centrally adjudicated strokes.

1) Determine who had a stroke by using the variable called STROKE in the "outc_ct_os_inv.dat"summary data file. In the combined CT and OS, 4978 participants have had a stroke, and based on the STROKESRC variable, 1066 came from local adjudications, 3511 from central adjudications, and 401 from the cause of death only.
2) Retrieve data items coded on the adjudicated strokes from "outc_stroke_carotid_inv.dat" by selecting records with STROKE=1. Note that the 401 strokes that came from the cause of death will have no detail information.
3) Merge the two sources of data and check to make sure the correct participants and data items have been selected.

Sample SAS code (assumes the data files have already been read into SAS):

```
PROC SORT DATA=outc_ct_os_inv OUT=allstrk (WHERE=(STROKE=1)) ; BY id ; RUN;
PROC SORT DATA=outc_stroke_carotid_inv OUT=strkdet (WHERE=(STROKE=1)) ; BY id ; RUN;
DATA strokes ;
  MERGE allstrk (in=insumm keep=id stroke strokesrc strokedy)
              strkdet (in=indetail keep=stroke ascsource strokedx strokeoxford stroketoast ) ;
  BY id ;
  IF insumm and indetail ;
RUN;

/* Note that since the OXFORD and TOAST classifications came from the centrally adjudicated strokes only, the variables
"strokeoxford" and "stroketoast" are missing for those locally adjudicated. */
```

4.13.2 SEER Cancer Coding

Detail information on all cancer outcomes is provided in two files: one for in situ and invasive breast cancer, and another with all other cancer sites. During the WHI study, only the primary cancers (breast, ovary, endometrial, colon, rectum, rectosigmoid junction) were centrally adjudicated and SEER coded. Cancers from all other sites were locally adjudicated. In the Extension Study, all cancers are centrally adjudicated, and the primary cancers continue to be SEER coded. All other cancers will eventually be SEER coded. The CCC is retrospectively coding one cancer site at a time. When all cases form one site are complete, all new cancers at that site will be coded prospectively. As of August 14, 2009, SEER coding is complete for the following sites: esophagus, lung, lymphoma (Hodgkin's and non-Hodgkin's), leukemia, multiple myeloma and pancreas. Form 130 is used to record the SEER coding, as well as estrogen and progestin receptor assays and Her2/Neu results for breast cancers. The local adjudication of cancers recorded only the site, tumor behavior, reporting source and diagnostic confirmation status on Form 122.

Because the values of many of the Form 130 variables differ by cancer site, we have not provided format values. For two variables, MRPHHISTB and ICDCODE, we provide reference files that can be reviewed or merged to obtain the value labels. For SIZE, EXTENSION and INVOLVE, it will be necessary to refer to the SEER coding manual to obtain the correct values using the following link:

(http://seer.cancer.gov/manuals/historic/EOD_2nd.pdf).

This link takes you to the main page of the coding manual. Use the table of contents or the bookmarks to find the section relevant to each specific cancer. In each section are the descriptions for all values of the variables for tumor size, extension and lymph node involvement. Some cancer sites may have more than one section. For example, cancers of the pancreas are split into two parts: head/body/tail and other/unspecified. Also, in section IV of the General Instructions is a description of the coding of the regional lymph nodes. Some of the information is repeated below.

<u>Tumor Size</u>
Codes the exact size of the primary tumor (in millimeter). If size is unknown/not stated, '999' is coded.

Exceptions (Refer to the SEER coding manual):

For the following sites, size is not applicable.
- Hodgkin's lymphoma
- Non- Hodgkin's lymphoma
- Leukemia
- Multiple myeloma

For the following sites, code '998' has a specific meaning.
- Breast
- Colorectal
- Esophagus
- Lung
- Stomach

Extension of Tumor

Codes the status of the tumor growth within the organ or origin, or its extension to neighboring organs, or its metastasis to distant sites. Code '99' is reserved for unknown tumor extension.

Lymph Node Involvement

Codes the status of the regional and distant lymph nodes. Code '9' is reserved for unknown lymph node status.

Exceptions (Refer to the SEER coding manual):

For the following sites, lymph node status is not applicable.
- Hodgkin's lymphoma
- Non- Hodgkin's lymphoma
- Leukemia
- Multiple myeloma

Number of lymph nodes positive/number of lymph nodes examined

Codes the number of lymph nodes that were positive by pathology review, and the total number of lymph nodes examined by the pathologist. Refer to the SEER coding manual for the meaning of codes 96, 97, 98 and 99.

Exceptions (Refer to the SEER coding manual):

For the following sites, number of lymph nodes positive/examined is not applicable.
- Hodgkin's lymphoma
- Non- Hodgkin's lymphoma
- Leukemia
- Multiple myeloma

Summary Stage

Codes the cancer summary stage taking into account the tumor site, size, multiplicity, depth of invasion, and extension to regional or distant tissues, involvement of regional lymph nodes, and distant metastases. See the following link for coding specifics:

(http://seer.cancer.gov/manuals/historic/ssm_1977.pdf).

5. 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 or enrollment
Days since randomization or 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 participants continuing in the Extension Study, the visit number for forms collected during the extension follow consecutively from their last visit during WHI.

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 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)
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.

Appendix A Previous Releases

Changes in the 6/19/2008 Release

BMD	Modified the algorithm for the "expected at visit" variable in all three data sets
Clinic CBC Results	Updated from the 9/12/05 database freeze.
ECG Results	Follow-up ECG results added to the data set
Form 2 - Eligibility Screen	Updated from the 9/12/05 database freeze
Form 4 - HRT Washout	First release of this data set
Form 20 - Personal Information	Updated from the 9/12/05 database freeze
Form 30 - Medical History	Updated from the 9/12/05 database freeze
Form 31 - Reproductive History	Updated from the 9/12/05 database freeze
Form 32 - Family History	Updated from the 9/12/05 database freeze
Form 34 - Personal Habits	Updated from the 9/12/05 database freeze
Form 37 - Thoughts and Feelings	Modified the algorithm for the "expected for visit" variable.
Form 42 - OS Questionnaire	Updated from the 9/12/05 database freeze
Form 43 - Hormone Use	Updated from the 9/12/05 database freeze
Form 44 - Current Medications	Data structure is now 1 row per med; "Expected at visit" variable was updated to include AV1; "Visit year" variable is now set to null when visit year was 0 for a semi-annual or annual visit
Form 45 - Current Supplements	"Expected at visit" variable was updated to include AV1
Form 48 - OS Follow-Up Year 1	Updated from the 9/12/05 database freeze; Data is now restricted to forms dated post-randomization
Form 60 - FFQ (sets a, b, c, d)	The PDF and SAS files were updated; The description and SAS name for visit year was changed to be consistent with other data sets; The data files did not change
Form 81 - Pelvic Exam	Modified the algorithm for the "expected at visit" variable
Form 82 - Endometrial Aspiration	Data set now restricted to EP ppts only instead of all HT; Modified the algorithm for the "expected at visit" variable
Form 90 - Functional Status	The functional status subsample was updated to include only ppts >= 65
Form 92 - Pap Smear	Modified the algorithm for the "expected at visit" variable
Form 143 - OS Follow-Up Year 3	Data now restricted to forms dated post-randomization; Descriptions of some METS variables were updated
Form 144 - OS Follow-Up Year 4	Same as Form 143

Form 145 - OS Follow-Up Year 5	Same as Form 143
Form 146 - OS Follow-Up Year 6	Same as Form 143
Form 147 - OS Follow-Up Year 7	Same as Form 143
Form 148 - OS Follow-Up Year 8	Same as Form 143
Form 149 – Sup. to OS Follow-Up	Same as Form 143
Outcomes - CaD Self-Reported	Variables "f33hosp" and "f33hosptimes" were added; See the data dictionary for documentation
Outcomes - CT+OS Self-Reported	Same as CaD Self-Reported

Changes in the 12/21/06 Release

New data sets:

- CaD Medication Adherence
- CaD Outcomes, Adjudicated
- CaD Outcomes, Self-Reported
- CT+OS Outcomes, Adjudicated
- CT+OS Outcomes, Self-Reported
- ECG – MI Novacode Serial Comparison
- Form 54 - Change of Medications
- HRT and CaD Unblindings Prior to Study Closure
- HRT Medication Adherence

Updated data sets:

- Core Analyte Results (follow-up data added)
- ECG baseline results (range check edits updated; files renamed; see section 4)
- Form 45 - Current Supplements (CT follow-up added; combined into one file)
- Form 92 - Pap Smear (follow-up added)

Changes in the 11/30/06 Release

New data sets:

- 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:

- 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)