Data Preparation and Use of WHI Datasets
BioLINCC Submission
December, 2018

Data as of March 31, 2018

Table of Contents

1. Introduction

The WHI datasets prepared for BioLINCC include most baseline and follow-up data for all Observational Study (OS) and Clinical Trial (CT) participants, including outcomes, adherence, CT unblindings and blood analytes from core, ancillary and BAA studies.

The WHI study ended March 31, 2005, and the closeout date for data collection was April 8, 2005. The Extension Study 1 (2005-2010) ended September 30, 2010. Participants consenting to join the Extension Study 2 (2010-2020) continue to be followed primarily for outcomes data collection. The outcomes and follow-up forms datasets have been updated and include all outcomes occurring and follow-up forms collected through **March 31, 2018**. The adjudicated outcomes from all phases of WHI are combined in the same datasets. Because of changes affecting which participants' outcomes are adjudicated, the self-report outcomes are split into datasets with those occurring in the main WHI Study and Extension Study 1 versus Extension Study 2.

In the Extension Study 2, cardiovascular and hip fracture outcomes are only adjudicated in a subset of participants referred to as the Medical Records Cohort (MRC). The MRC consists of all former hormone trial participants and all African American and Hispanic participants from all study components. For the remaining participants (referred to as the Self-Report Cohort or SRC), only self-reported outcomes are collected, except all cancers are still adjudicated.

Section 4.18 contains information on the outcomes datasets, including details on which outcomes are adjudicated versus self-reported in the three phases of WHI**. Please review section 4.18 carefully before using the outcomes datasets.**

2. Data File Setup

Each data set is provided as a tab-delimited ASCII file with a header row containing the variable names. The code needed to create SAS datasets 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.

Not all data files have the same number of records because not every form was completed by each participant. When multiple screening forms were submitted for a participant, we have included only the latest form prior to randomization or enrollment. The first variable in each file, called *ID* (sometimes referred to as the "common ID"), is a unique participant identifier that replaces the WHI Member ID. All files include *ID* and it MUST be used to merge multiple files. For form-based datasets, the order of the variables after ID generally 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 datasets. The form questions used in to create computed variables have been noted in the variable descriptions. If you would like a copy of the SAS code used to create a computed variable, contact helpdesk@whi.org.

The following extensions are often used in variable names:

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 tab-delimited data files are represented by a blank. 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 was 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 was set to “missing”. If a question is a sub-question, it is 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 and Study Membership

There are six computed variables in the demographics dataset that indicate region and sunlight exposure based on latitude at WHI enrollment and CaD randomization. Understandably, these may seem out of place in this dataset. When the variables were added in 2014, they did not fit into any existing category or dataset. We decided to include them in the demographics dataset rather than create a new dataset.

The variables are:

- LATREGION "Latitude (degrees N) of CC at CT randomization/OS enrollment"
- WATTSCAT "Watts ((J/s) per m2) of CC at CT randomization/OS enrollment"
- LANGLEYSCAT "Langleys (g-cal per cm2) of CC at CT randomization/OS enrollment"
- LATREGIONCAD "Latitude (degrees N) of CC at CaD randomization"
- WATTSCATCAD "Watts ((J/s) per m2) of CC at CaD randomization"
- LANGLEYSCATCAD "Langleys (g-cal per cm2) of CC at CaD randomization"

4.2 Extension Study Membership

The EXTFLAG variable indicates if a participant was enrolled in the first WHI Extension (2005-2010). The EXT2FLAG variable indicates if a participant is enrolled in the second WHI Extension (2010-2020). Flags pertaining to the second Extension were added to this data set to indicate if a participant is part of the Extension 2 Medical Record Cohort (EXT2MRC) or the Extension 2 Self-report Cohort (EXT2SRC) as defined in the introduction part of this document.

4.3 Computed Variables

Many computed variables that have been commonly used in data analyses are included in various datasets. In general a computed variable resides in the data set which contains 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. Note: Users agree not to use the data on American Indians or Alaskan Natives to infer tribal status or affiliation.

4.5 Current Medications (Form 44)

Included with the Current Medications dataset 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 23 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 23 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 Lifestyle Questionnaire (Form 155)

Form 155 was designed to collect information on a broad range of psychosocial, physical activity and physical function topics. The form was administered once in place of Form 151, during the Extension Study 2010-2020 year 2 annual mailing. All the questions on Form 151 were included on Form 155 plus many of the same questions asked on Forms 34/35, 36/37 and 38 during WHI, and new questions not asked previously. In addition to the individual questions, the dataset includes constructed variables using the same algorithms as before when applicable and new constructs for perceived stress, personal growth, life purpose, satisfaction with life and resilience. There are two different variables for personal growth; PGROWTH1 constructed using a 3-item scale, and PGROWTH2 using a 7-item scale. Similarly, life purpose has two variables; PURPOSE1 based on a 3-item scale, and PURPOSE2 using a 7-item scale. In the past, the questions related to recreational physical activity asked about moderate and strenuous activity separately; Form 155 combined them. For this reason, the variables for energy expenditure from strenuous (HARDEXP), moderate (MODEXP) and total recreational physical activity (TEXPWK) could not be computed; nor could strenuous activity episodes per week (SEPIWK) or minutes of strenuous physical activity per week (SMINWK).

4.8 Diet

4.8.1 Food Frequency Questionnaire (FFQ, Form 60)

There are three datasets that contain Food Frequency Questionnaire (FFQ) data. One data set contains item level data while the other contains over 100 nutrients calculated from participant responses to the FFQ. The 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). Be sure to read the guidelines for diet-disease analyses posted to the 'dataset documentation' page and the 'propose a paper' page. 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 third dataset that contains FFQ data is the My Pyramid Equivalents Database (MPEDs). See Section *4.8.2 MPEDs and Diet Quality Indices.*

There are a number of vitamin A related variables in the WHI nutrient data set that use different units. Analysts 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.

There are also a number of dietary variables related to folate intake. Folate is present in foods in a naturally occurring form (typically called "folate") and a synthetic form (typically called "folic acid"). The latter is added to foods in fortification and in dietary supplements. Before using this data, please read the separate document on our website (Folate and folic acid variables - WHI data) that explains which variables to use and how to compute a summary variable that correctly accounts for the fortification and different bioavailability of the two forms of folate.

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.

4.8.2 MPEDs and Diet Quality Indices

The variables in the new dataset, *f60_mpeds_inv*, were derived from the same FFQs used to create the two datasets described above. The MPEDs data includes 32 measures of per 100 grams equivalents of food groups calculated using the *MyPyramid Equivalents Database 2.0* (MPED). These MPEDs food groups-measures are based on the USDA's 2005 release of the American food guide pyramid.

The component scores and total score for the Healthy Eating Index-2005 (HEI-2005) can be found in the dataset, *f60_hei_2005_inv*. The HEI-2005 scores are measures of diet quality that assess conformance to Federal guidance and the published Dietary Guidelines for Americans 2005; see accompanying dataset documentation for details.

Likewise, datasets for component scores and total scores of the HEI-2010, HEI-2015 and Alternate Healthy Eating Index (AHEI)-2010 can be found in *f60_hei_2010_inv, f60_hei_2015_inv, and f60_ahei_2010_inv*, respectively, along with corresponding documentation. Guidance and pseudo-code (SAS) is also available for Alternate Mediterranean Diet (aMed) and Dietary Approaches to Stop Hypertension (DASH). Data are not provided for the latter two indices because aMed and DASH are reliant on quantile values (e.g., quintiles) from the sample population of interest, so depend on an investigator's specific research objectives. Investigators are welcome to create aMed or DASH components and total score using these guidelines.

All diet quality indices are based directly or indirectly (via MPEDs) on FFQ data, and therefore possess inherent limitations of FFQ derived data. 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.

Additional details regarding the MyPyramid Equivalents database or any of the diet quality indices can be found in corresponding documentation included with each dataset.

4.9 Medication and Supplement Inventory (Form 153) collected in WHI Extension

Included with the Form 153 – Medication and Supplement Inventory datasets are a number of reference files, including a PDF called *F153_ReadMe.pdf*. The 153_ReadMe document provides further details about the collection and analyses of the Medication and Supplement Inventory collected in the WHI Extensions 1 and 2.

4.10 Bone Densitometry Results: BMD

In the November 30, 2006 release, the BMD data were reorganized into three files by scan type: Hip, Spine and Whole body. 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 sub study. 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 datasets. 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	Whole body total fat
Total spine BMD	Whole body total percent fat
Whole body BMD	Whole body total lean
Whole body BMC	Whole body total fat free mass
Whole body total 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, provided the BMD values of L2, L3 and L4 are all known. Conversely, it is set to missing if any of L2, L3 or L4 is missing.

We have included both the uncorrected and corrected values in the BMD datasets. Previous releases of the BMD data included corrections to trochanter BMD and inter-trochanter 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 are included and can be identified by the SAS variable names HIPQDR, SPNQDR, and WBQDR.

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).

In the June 2014 release, four computed T-score variables were added to the BMD datasets:

HIPTSCORE	"Total Hip BMD T-score"
HIPNKTSCOR	"Hip Femoral Neck BMD T-score"
SPNTSCOR	"Spine (L2-L4) BMD T-score"
WBTSCOR	"Whole Body BMD T-score"

The T-score represents the number of standard deviations the participant's BMD is above or below the mean for a healthy 30 year old woman of the same race/ethnicity. Reference values were provided by UCSF and are machine specific.

4.11 Blood Results: Clinic CBC

The data file named *CBC* includes the results from blood collected and analyzed at each Clinical Center's (CC) local laboratory. All clinical trial and observational study participants were to have serum collected and analyzed once during screening. In addition, all observational study participants had another serum collection and CBC at their year 3 annual visit. Data is missing if the lab was unable to process the sample. Values were reported for the following tests: white blood cell count (Kcell/microliter), platelet count (Kcell/microliter), hematocrit (%), and (when available) hemoglobin (gm/dl).

The clinic CBC data does not include any white blood cell differential data. While those data may have been collected as part of the WHI CBC, they were not entered into the WHI database.

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.**

4.12 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 this dataset are the test results from the participants in the Observational Study Measurement Precision Study (OS-MPS). This is approximately 1% of the OS.

4.13 Specimen Results from Core and Broad Agency/Ancillary Studies (SPEC)

Currently, the Specimen Results data set contains results from serum, plasma, RBCs, and urine from nearly all Core studies and from Ancillary and Broad Agency Announcement Studies with funding end dates of over one year prior to the date of the current release, including the Core Analytes and Clinic CBC data.

When using specimen results data it is important to note that subsampling may have been applied for selecting participants for blood collection and/or for testing. Also more than one lab, assay method, or specimen type could be used to determine a result for the same type of test.

Although all WHI participants had blood collected at baseline, only active CT participants (about 95%) had blood collected at annual visit 1, and only active OS participants (about 80%) had blood collected at annual visit 3. Only a small subsample of participants had their blood collected after these collections. A 6% subsample of CT participants selected at random with minorities oversampled, which included approximately 8.6% of the HT and 4.3% of the DM participants, had their blood collected at annual visits 3, 6, and 9. A 1% sample of OS participants had blood collected at baseline and within 6 months after baseline. Approximately 5% of the original WHI population had blood collected as part of the WHI Long Life Study (WHILLS) that took place approximately 18 years post-enrollment.

For details about the files associated with Specimen Results and use of the data, please refer to the Specimen Results Read Me document on our website. This document is also included with the Specimen Results data set.

There is a variable in the Specimen Draws file, PROCPROT, indicating which blood processing protocol was used. The WHI Long Life Study (WHILLS) blood collection and processing protocol (WHI Long Life Study: Synopsis of Blood Protocol) was substantially different than that of the WHI clinic protocol (Biospecimen Collection). The fasting blood specimen collected for the Nutritional Biomarker Study (NBS) and Nutrition and Physical Activity Association Study (NPAAS) were collected and processed according to the WHI clinic protocol.

There are several variables in the Specimen Draws file that are specific to the WHILLS Processing Protocol. One of them, SSTRESPUN, flags serum vials that are from Serum Separator Tubes (SST) that were centrifuged twice – in error. It is known that such double centrifugation can alter glucose and potassium concentrations. It is possible that other test results would also be compromised. *Exercise caution when using test results from SST vials when SSTRESPUN = 1.*

4.15 ECG Results

The data file *ecg_ct_inv.dat* contains baseline (one record per participant) and follow-up (multiple records per participant) 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.

4.16 Risk Scores

In the June 2014 release the new data file *risk_scores_inv.dat* was added to the Medical History category containing the Gail Model 5-year breast cancer risk score and the WHO FRAX risk scores.

The Breast Cancer Risk Assessment Tool (i.e., the Gail model) is used to predict risk of invasive breast cancer in women 35 years of age or older. Gail model variables include age, ethnicity, age at menarche, age of the mother at the birth of her first live child,

number of first-degree relatives with breast cancer (0, 1, or >1) and the number of previous breast biopsy examinations (0, 1, or >1). Calculations of 5-year risk estimates for WHI women were made by the National Surgical Adjuvant Breast and Bowel Project statistical center by following their usual coding procedures on baseline data from individual WHI participants. Because historical information on atypical hyperplasia was not collected in the WHI, all women with previous breast biopsy examinations are coded as "unknown" for this variable.

Reference:
Gail M, Costantino J, Bryant J, et al. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *J Natl Cancer Inst.* 1999;91:1829-1846.

The FRAX® tool has been developed by WHO to evaluate fracture risk of patients. It is based on individual patient models that integrate the risks associated with clinical risk factors as well as bone mineral density (BMD) at the femoral neck. The FRAX® algorithms give the 10-year probability of fracture. The output is a 10-year probability of hip fracture and the 10-year probability of a major osteoporotic fracture (clinical spine, forearm, hip or shoulder fracture). Calculation of these risk probabilities for WHI women are based on baseline data.

References:
For complete information on the development and use of these scores, and a list of references, see:
http://www.shef.ac.uk/FRAX

4.17 Observational Study Follow-up Questionnaires

The OS follow-up datasets 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 datasets 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 datasets. The Form 60 datasets 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 datasets.

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.

4.18 Outcomes

The current release of outcomes data includes centrally verified, locally verified and self-reported outcomes occurring on or before and adjudicated through March 31, 2018 for CT and OS, and occurring on or before September 30, 2010 for CaD. Summary data for verified outcomes is divided into two files: *outc_ct_os_inv.dat* (CT+OS adjudicated outcomes), and *outc_cad_inv.dat* (CaD adjudicated outcomes). If a participant is not part of the MRC, they will have no confirmed outcomes occurring after the end of Extension Study 1 in these datasets, except for cancer outcomes which continue to be adjudicated in all participants. Detail information collected on verified outcomes are in additional datasets by outcome type, separately for CT+OS and CaD.

The self-reported outcomes data is now in two parts. The datasets including all self-reported outcomes through Extension Study 1 are: *outc_self_ctos_inv.dat* (CT+OS self- reported outcomes), *outc_self_cad_inv.dat* (CaD self-reported outcomes), and those including Extension Study 2 self-reported outcomes data only are: *outc_self_x2_mrc_inv* and *outc_self_x2_src_inv.*

In these new self-reported datasets for Extension Study 2, only first outcomes occurring after September 30, 2010 and through March 31, 2018 are included. Because outcomes other than cancer and death are no longer adjudicated in the SRC, the SRC self-report dataset now includes outcomes that were previously adjudicated.

The adjudication process, set of outcomes adjudicated, self-reported outcomes, and information on screening procedures collected on Form 33 changed at the start of Extension Study 1, and again at the start of Extension Study 2. The following tables describe the changes, as do the "Usage Notes" in the data dictionary files.

Table 1: WHI adjudicated outcomes by study period and WHI cohort

	WHI 1993-2005				**Extension 1 2005-2010**	**Extension 2 2010-2020**	**Extension 2 2010-2020**	
Outcome	HT	DM	CaD	OS	CT/OS	MRC	SRC	**Form**
CARDIOVASCULAR:								
MI	C	L	L	L	C	C	S	121
Stroke	C	L[1]	L[1]	C[1]	C	C	S	121/132
WHI congestive heart failure	C	L	L	L	S		S	121
UNC heart failure	C[2]	C[2]	C[2]	C[2]	C[2]	C[2]		135/136
Angina	C	L	L	L	S	S	S	121
Peripheral artery disease	L	L	L	L	C	C	S	121
Carotid artery disease	L	L	L	L	C	C	S	121/132
Coronary revascularization	L	L	L	L	C	C	S	121
TIA	C	L	L	L	S	S	S	121/132
Heart valve disease						C	S	121
Aortic aneurysm/dissection						C	S	121
CANCER:								
Breast cancer	C	C	C	C	C	C[3]	C[3]	122/130
Endometrial cancer	C	C	C	C	C	C[3]	C[3]	122/130
Colorectal cancer	C	C	C	C	C	C[3]	C[3]	122/130
Ovarian cancer	C	C	C	C	C	C[3]	C[3]	122/130
All other cancers	C	C	C	C	C	C[3]	C[3]	122/130
FRACTURES:								
Hip	C	C	C	C	C	C	S	123
Non-hip fractures	L	L	L	L[4]	S	S	S	123
OTHER:								
Pulmonary embolism	C	S	S	S	C (HT)	C	S	126
Deep vein thrombosis	C	S	S	S	C (HT)	C	S	126
Hysterectomy	C	S	S	S	C (HT)	S	S	131
Death from any cause	C	C	C	L	C	C	S	124

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

[1] Used central adjudication data when available, and local adjudication data otherwise
[2] Adjudication by UNC is only done for the HT and Black/Hispanic participants.
[3] All cancers centrally adjudicated on all Extension Study 2 participants
[4] Done only at the BMD centers

Table 2: WHI self-reported outcomes for CT/OS participants by study period

Outcome	WHI 1993-2005	Extension 1 2005-2010	Extension 2 2010-2020
Condition			
Atrial fibrillation	-	-	X
Cataracts*	X	-	-
Colorectal polyps/adenomas	X	X	X
COPD	-	-	X
Dementia, Alzheimer's	-	X	X
Diabetes diagnosis, ever	-	X	-
Fainted/blacked out	X	-	-
Falls	X	-	X
Fractures (non-hip)	X	X	X
Gallbladder disease/stones	X	-	-
Glaucoma	X	-	-
Kidney/bladder stones*	X	-	-
Macular degeneration	-	X	X
Osteoarthritis*	X	X	X
Osteoporosis	X	-	-
Parkinson's disease	-	X	X
Rheumatoid arthritis	X	-	-
SLE	X	X	X
Exams/Procedures			
Barium enema X-ray*	X	X	-
Blood in stool	X	X	X
Blood pressure*	X	-	-
Bone density scan	-	X	X
Breast biopsy/aspiration	X	X	X
Breast exam	X	X	X
Breast exam, other (MRI/ ultrasound)	-	X	X
Cholesterol	X	-	-
D&C	X	X	-
ECG	X	-	-
Endometrial biopsy	X	X	X
Eye exam*	X	-	-
Flex sig/colonoscopy	X	X	X
Hysterectomy	X (non HT)	X (non HT)	X
Mammogram	X	X	X
PAP smear	X	-	-
Physical exam*	X	-	-
Rectal exam	X	X	-
Medication/treatments			
Anxiety pills	-	X	-
Depression pills/therapy	-	X	-
Diabetes, diet/exercise	-	X	X
Diabetes, insulin	X	X	X
Diabetes, pills*	X	X	X
Estrogen pills	-	X	-
High blood pressure pills*	X	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 March 31, 2018, the central adjudication result was used; otherwise if a local adjudication exists, the local adjudication was used. In addition, for participants not enrolled in Extension Study 1, 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. Similarly, for participants not enrolled in the Extension Study 2, any outcomes occurring after the study close-out date of September 30, 2010 are censored.

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 the Annual Progress Reports, and new outcomes added to Form 33 in both Extension Studies.

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 dictionaries.

Six verified outcomes have a "subsequent condition" rule (angina, possible silent MI, definite silent MI, TIA, carotid artery disease, and in situ breast cancer). This rule means that an angina 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 on or before March 31, 2018 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. A variable with the days from enrollment to last contact (LAST33DY) is included in the self-reported outcomes files.

The variable ENDFOLLOWDY in the adjudicated outcomes summary data file (*outc_ct_os_inv.dat*) represents the days from enrollment to death if it occurred, or end of follow-up, based on the entire available follow-up period. For a time-to-event analysis using adjudicated outcomes, this variable should be used as the censoring variable. For a time-to-event analysis using self-reported outcomes, the days from enrollment to death or days from enrollment to the last Form 33/33D if no death occurred (called LAST33WHIDY, LAST33EXT1DY and LAST33DY for the latest Form 33/33D during WHI, Extension Study 1 and Extension Study 2, respectively, see above) should be used as the censoring time for those participants without the event.

To allow for analyses that only include outcomes occurring through the end of WHI, or the end of the Extension Study 1, two variables are provided that represent the days from enrollment to the end of WHI, and end of Extension Study 1 follow-up. These variables are located in the adjudicated outcomes summary data file (*outc_ct_os_inv.dat*), and are named ENDWHIDY and ENDEXT1DY. For example, to restrict an analysis to just the time period of WHI, any outcomes with "days to outcome" after the "days from enrollment to the end of WHI" would not be counted, and the participant's follow-up time would be censored using this new variable. Similarly, to restrict analyses to the CaD trial follow-up period, the variable called ENDWHICADDY in the *outc_cad_inv.dat* data file should be used, and to restrict a CaD trial analysis through the end of Extension Study 1, the ENDEXT1DY variable in the *outc_cad_inv.dat* data file should be used. All of the above mentioned variables combine death and last contact information for the relevant time period. Similar methods should be employed when utilizing the latest Form 33/33D information for self-reported outcomes.

A small number of CT/OS participants have no Form 33 or 33D in the study database. Also, a small number of CaD participant 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 and CaD, an additional 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.

4.18.1 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 Death detail file includes the cause of death as categorized on Form 124 plus an extended list of specific cancer types, and the ICD code for the underlying cause of death. Both ICD-9-CM and ICD-10-CM codes are used, where the ICD-10-CM codes consist of a letter in the first position which is not present in the ICD-9-CM codes. The following links can be used to look up specific codes:

ICD9 Data and Medical Billing and Coding search

The data file *outc_death_all_discovered_inv.dat* includes all reported and discovered deaths. For consistency of follow up timing, the death variables in the adjudicated outcomes files and the main death detail file only include deaths that occurred in a study phase the participant consented to, e.g., a participant whose death occurred in Extension Study 2 but who did not consent to Extension Study 2 would not be marked as dead. This new file includes all reported and discovered deaths regardless of consent. This file also includes follow up time variables (ENDFOLLOWALLDY and ENDEXT1ALLDY) that should be used in conjunction with the DEATHALL variable for analysis purposes. These follow up time variables incorporate the added follow up time from National Death Index searches for participants who were included in a given search, but were not found as deceased, as well as using the death date for those that were found deceased. **Note:** This data should be used only for analyses that use death as the outcome of interest, i.e., events from cause of death that occur after a participant's consent period should not be incorporated into the corresponding outcome (e.g., breast cancer or CHD) since ascertainment of the outcome after the consent period but prior to death is not possible.

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 addition, the STROKESRC variable reports the number of strokes based on local versus central adjudication, and those determined from the cause of death only.

2) Retrieve data items coded on the adjudicated (both local and central) strokes from *outc_strk_carotid_inv.dat* by selecting records with STROKE=1. Note that the subset of 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.

4.18.2 Heart Failure

Congestive heart failure (fatal and nonfatal) requiring hospitalization, was monitored during the WHI (1993-2005) for all CT and OS participants. The first adjudicated event is included in the outcomes datasets. In the CT+OS *outc_ct_os_inv* dataset, the indicator variable for a CHF event and corresponding variable for days from enrollment to first confirmed CHF diagnosis are called CHF and CHFDY, respectively. Form 121 questions relevant to the CHF event are in the *outc_cardio_inv* detail file. For CHF events during the CaD trial, the corresponding variables are in *outc_cad_inv* and *outc_cardio_cad_inv*.

Another heart failure outcome is available on the subset of WHI participants randomized to the Hormone Trial and all Black and Hispanic participants (MRC Super Cohort) enrolled in WHI (total N = 44,174). Self-reported cases occurring during WHI and the Extension Studies were adjudicated at UNC. Data on all adjudicated cases, not just the first, are in the dataset, *unc_hf_inv.* This dataset includes the entire MRC Super Cohort subset. The variable HFDIAG is the diagnosis from the case adjudication, and CASEDY is the days from enrollment to the HF case date. These variables are missing if there is no adjudicated case. WHI counts a diagnosis of either definite or possible decompensated heart failure as a UNC heart failure case; these cases are identified by the UNCHF variable.

A small number of self-reported cases were sent to UNC for adjudication, but no information is available yet from the adjudication process, and sufficient documentation (at least 2 essential documents) for another small set of cases could not be obtained, so the case was never sent to UNC to be adjudicated. These participants are identified by the variable UNCMISSING. Users of this data should consider whether to include or exclude these participants from an analysis since the outcome for each case is not known.

More detailed information on both the WHI and UNC heart failure data is available:
https://www.whi.org/researchers/data/Documents/WHI%20and%20UNC%20Heart%20Failure%20Data%20Summary.pdf.

4.18.3 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 non-breast 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. The local adjudication of cancers recorded only the site, tumor behavior, reporting source and diagnostic confirmation status on Form 122. The CCC retrospectively adjudicated and SEER coded all cancer cases from the WHI study. All cancers will be centrally adjudicated and SEER coded in the Extension Study 2. All data for cancers from the Extension Study 1 are centrally adjudicated and SEER coded. Cancer cases from the Extension Study 2 are prospectively being adjudicated and SEER coded on a site-by-site basis.

Form 130 is used to record the SEER coding, which includes estrogen and progestin receptor assays and Her2/Neu results for breast cancers. For consistency, the study has used the 'SEER EOD-88, 2nd edition' of the SEER coding scheme with a few exceptions. The study uses 'SEER EOD-88 3rd edition' for all pancreatic cancer cases since it included a minor edit to the 2nd edition coding scheme. WHI also uses the 'SEER EOD-88, 3rd edition' for all Hodgkin's and non-Hodgkin's lymphoma cases as a result of the 3rd edition upgrade of the lymphoma coding scheme which held relevance for the new classification of lymphoma cases.

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/archive/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 site. 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.

Tumor Size

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
- Mycosis fungoides/Sezary's disease (skin)
- Malignant melanoma
- Unknown site

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 regional lymph nodes positive/number of regional lymph nodes examined

Codes the number of regional lymph nodes that were positive or negative by pathology review, and the total number of regional 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 regional 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/archive/manuals/historic/ssm_1977.pdf).

4.18.4 Aging Variables

The data file called *outc_aging_outc_inv.dat* in the Outcomes category includes four sets of variables for use in studies related to aging and longevity. These variables are defined based on the same "data as of" date as the outcomes datasets, and use all reported and discovered death information, and where the last follow-up contact is the later of the latest Form 33 or NDI search date. These variables allow analyses looking at survival to ages 85, 90, 95 and 100 years.

An example of an analysis using this data would be one interested in comparing characteristics of participants who survived to age 90 or older to those who died before age 90. The variable AGEELIG90 would be used to define the eligible sample, i.e. those participants, based on their birthdate, who had the potential to survive to age 90 or older. And the variable SURVAGE90 would define the longevity outcome of survival to age 90 or older. Participants who were unclassifiable because their follow-up ended before age 90 would be excluded from the analysis.

4.19 WHI Long Life Study Home Visit (Form 301)

7,875 WHI Extension Study participants between the ages of 63 and 99 years old participated in the WHI Long Life Study (WHILLS) home visits between March 2012 and May 2013. The eligible population for this project consisted of 14,081 women in the Medical Records Cohort who were at least 63 years old (by 12/1/11). Women were excluded if they resided in an institution (e.g., skilled nursing facility) or were unable to provide informed consent due to dementia. All participants have baseline biomarker data (glucose, insulin, CRP, creatinine, and lipids) and GWAS data. The WHILLS Form 301 data includes a brief clinical assessment, an assessment of functional status, and information about the blood drawn at the home visit. Please see WHI Long Life Study for additional information about the WHILLS.

The February 2014 update included two Short Physical Performance Battery (SPPB) scores that are based Form 301 data. Form 301 was designed to collect the data needed for the 'Look AHEAD' SPPB (LASPPB), which is based on the 'Health ABC' SPPB[2]. The LASPPB is the sum of 3 ratios, the Standing Balance Ratio, the Chair Stand Ratio and the Usual Walk Ratio[3], and results in a continuous variable ranging between 0 and 3. The 'Established Populations for the Epidemiologic Studies of the Elderly' SPPB[1] (EPESESPPB) can also be scored from the Form 301 data. The EPESESPPB is the sum of 3 individual scores: the Total Balance Score, the Chair Stand Score, and the Gait Speed Score, and yields a continuous score ranging from 0 to 12. For both SPPB scores, if an element (e.g., one of the ratios) of the score is missing, the SPPB score is set to missing. The SPPB scores are also set to missing if the Timed Walk Alert Flag (TWALERTFLG) = 1, which means that the Timed Walk data were not collected per protocol.

Additional details on the SPPB scoring may be found here:
https://www.whi.org/researchers//data/Documents/Long%20Life%20Study%20SPPB%20Scoring.pdf.

SPPB References:

[1] Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol Med Sci. 1994;49,M85–M94.

[2] Simonsick EM, Newman AB, Nevitt MC, et al. Measuring higher level physical function in well-functioning older adults: expanding familiar approaches in the Health ABC Study. J Gerontol Med Sci. 2001;56A,M644-M649.

[3] Houston, D (Assistant Professor, Wake Forest School of Medicine). Personal communication to Sue Mann: August 29, 2012 and April 18, 2013.

Form 301 also includes some informational data items regarding the Objective Physical Activity and Cardiovascular Health study (OPACH). These OPACH data items will be included in a future Form 301 data release.

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 the frequency of data collection documents that are on our website (WHI Data Available on this Site) 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 year, closest to visit within visit type and year, 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 Year

On all forms there is a field for the Clinical Center to enter the year/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 year/number is not required and is set to missing.
Except for Form 44 – Current Medications data, the visit year/number for a screening visit type is set to zero. In the Form 44 dataset 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,
2 - for semi-annual contacts 18 months following randomization,
3 - for semi-annual contacts 30 months following randomization,
etc.

The Visit Year 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 year for forms collected during the extension follow consecutively from their last visit during WHI.

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 datasets. 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 datasets where it is deemed useful. It is not included where the dataset 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 dataset). 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 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:

<u>1. By visit type and year (technique used most often by CCC)</u>

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 1 Form 80s from the Form 80 data, you could restrict the rows in the file to the following:

F80VTYP = 3 and F80VY = 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, 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

The criteria you use when choosing rows within a data file should be based on your analysis objectives.

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