Updated 6/23/23

WHI Specimen Test Results (Read Me)

Table of Contents

1. Overview

The WHI Specimen Test Results data files include the following:

1. CBC File: WHI Clinic CBC Results
2. CORE File: CT Subsample Core Analyte Test Results
3. Specimen Results File: Test results from the WHI Clinic CBCs, CT Subsample Core Analytes, Long Life Study, most other Core studies, and Ancillary/Broad Agency Announcement studies

Test results from almost all of the non-DNA Core (W) and Broad Agency Announcement Studies (BAAs or BAs) studies are included. Test results from Ancillary Studies (AS or M) with funding end dates one year or more prior to each data release are added on an annual or semi-annual basis. The WHI Clinic CBC and CT Subsample Core Analyte results files have been part of the investigators dataset for many years. In November 2013, the WHI Clinic CBC results and the CT Subsample Core Analyte test results were also added to the Specimen Results file in order to establish a single resource for all WHI Specimen Test Results.

For a listing of tests included in the Specimen Results along with many details about the assays, please click on 'Specimen Result Descriptions'. NOTE: This link provides *only a description* of the specimen results. The actual test results for participants are located in the WHI Specimen Results data file. Access to the WHI data files is limited to individuals in one of these categories:

- Current WHI Principal Investigator
- Ancillary Study Principal Investigator meeting the requirements for data access per Ancillary Study Policy
- Former Principal Investigator who was active during the first WHI Extension (2005-2010) for at least one year
- Lead author on an approved paper

For information on how to use and interpret the Specimen Result Descriptions, see #4, below.

2. Cautions

- *Test Versions*

  WHI assigns a new Test Version for an analyte for every new test method, sample type used, unit of measure used (e.g., ng/ml, pg/ml), or laboratory that tests the analyte.

  Examples:

  - If five different laboratories perform testing of CRP, there will be five Version numbers.
  - If one of these five CRP laboratories reported CRP as mg/dL and later changed to reporting CRP results in mg/L, the results reported in mg/L would have a separate Version number.
  - If one of these five CRP laboratories changed its testing method for CRP, the new method would have a separate Version number.
  - If one of these five CRP laboratories tested samples using serum for 3 WHI ancillary studies, but for another 6 studies tested EDTA plasma, the latter would have a separate Version number.

  Use caution when combining data for an analyte from several different Test Versions. (See Statistical Methods for Biomarker Data, #5, below.)

- *Test Units*

Most Test Versions for an analyte use the same units of measure, but not always. Make sure that you convert the test results to the same units of measure.

For information about how to convert units, use an online converter, such as this one: http://www.endmemo.com/medical/unitconvert/, or contact the WHI Help Desk (helpdesk@WHI.org).

- *Case/Control/Cohort Selection*
  The participant selection for each Study is specific to that study. Take care to avoid selection bias when combining data from more than one Study. Some information about the participant selection for each study may be found on the WHI Study Pages: https://www.whi.org/studies.

- *Sample Storage Time*
  WHI has tested several key analytes over time to assess the impact of long-term storage on assay results. The results, which almost certainly will vary by analyte, will be posted to whi.org/data at some point in the future. For now*, investigators should be very cautious about combining assay results from samples stored for varying amounts of time*. The 'TESTDY' variable in the 'Specimen Test Results File' provides the number of days between the draw date and the date samples were shipped to the testing lab.

- *Specimen Processing Protocol*
  Because the blood was not collected in a clinical setting, the blood collection/processing protocol for the Long Life Study was substantially different than that of the WHI clinic visits. There is a variable in the Specimen Draws file, PROCPROT, that indicated which blood collection/processing protocol was used. A synopsis of the Long Life Study blood collection/processing protocol may be found here, and the WHI Clinic protocol here.

- *When in doubt…*
  Contact the PI for the study that generated the test results. It is the study PI who has first-hand knowledge about participant selection criteria, the test method, the laboratory, and use of the test results in the analysis for her/his study.
  - *How to find contact information for an Ancillary Study PI:*
    - Determine the WHI Study ID number(s) of interest from the 'Specimen Result Descriptions'. (For information on how to use and interpret this resource, see #4, below.)
    - Contact the WHI Help Desk (helpdesk@WHI.org) to request the contact information (institution name and email address) for the Study PI.

Users of the Specimen Test Results data are responsible for determining the appropriateness of combining data from multiple studies, for excluding questionable and/or extreme values from analyses, and for careful inspection of the data before starting an analysis project.

3. Descriptions of Data Files

- *Specimen Test Results*
  The file named "`spec_results_ctos_inv.dat`" contains the test result values. There is one row in this file per test result. Investigators will need to merge the Specimen Test Results file with the Specimen Test file to get details about the test version (lab name, specimen type, etc.) that corresponds to the result value. You may also want to merge The Specimen Test Results file with the Specimen Draws file to get details about the blood draw (e.g. timing, hours fasting).

- *Specimen Draws*
  Additional information regarding the specimen collection (draw), such as days since randomization, visit, visit year, timing of draw, number of hours fasting, is provided in the Specimen Draws data file called "`spec_draws_ctos_inv.dat`" This file holds one row per Specimen Draw. Join to this file using the ID and PPTDRAW columns.

- *Specimen Tests*
  The file called "`spec_tests_ctos_inv.dat`" contains detailed information about a single Version of a Specimen Test, including the method (when available), the units of measure, the lab name, the number of results for this Version, and the median and standard deviation of the test results. Also included is the number of quality assurance samples (Blind

Duplicates) that WHI included among the participant samples. (Please see #7.a., below, for a description of the WHI Blind Duplicate Quality Assurance samples.)

NOTE: The WHI Clinic CBCs were performed at several local clinical laboratories. For logistical reasons, this file shows a single lab name for the Clinic CBCs: WHI Clinic Labs.

Two measures of quality are provided in this file:

1. Coefficient of Variation (CV%). The CV% is calculated for each blind duplicate pair (i.e., standard deviation of the two results ÷ the average of the two results * 100). The average of these individual pair CV%s is reported.
2. Correlation Coefficient (Corr) or Log Correlation (Log Corr)*. The Corr or Log Corr is calculated for all blind duplicate pairs tested with this Version. Hence, the Corr or Log Corr is an overall measure of the linear association between blind duplicate paired samples. The Corr or Log Corr is not calculated if fewer than three blind duplicate pairs were tested.
   *Log Correlation Coefficients are calculated instead of Correlation coefficients for tests with positively skewed data (skew ≥+2).

- *Study Case-Control Types*
  The file containing the case/control information of the study that performed the test is called "`spec_case_control_types_ctos_inv.dat`". Please be careful when using these data. The cases and controls were selected for each specific study based on the study requirements. The case/control designation in the file only designates the final case/control selection for the specific study and does not describe the complex selection criteria. Some participants may have served as more than one type of case or control for the same study, depending on study design. For example, Study #129 investigated colorectal, breast, and endometrial cancers. Participants in this study with more than one of these cancers have a row in this dataset corresponding to each case designation. Study #362 had multiple study aims and some participants served as controls for more than one aim. Therefore, the case/control dataset will have a row for every distinct control designation that a participant had in this study. In addition, cases and controls selected at the time of the study may have developed conditions or outcomes after they were included in the study. For example, a control for a breast cancer study may have developed breast cancer after being selected as a control.

  For some studies, information about the case/control selection is provided in WHI Study Pages. Check here first to assess the selection criteria for a study. However, it is wise to contact the PI to determine case/control assignments appropriate for analysis. For example, Study #W15 (Vitamin D levels in CaD participants with colorectal cancer or fractures) selected both colorectal cancer and fracture cases and matched controls. The variable called CASEFLG identifies the cases and controls but does not distinguish between the types of cases.

- *WHI Clinic Complete Blood Count (CBC) Results*
  - The data file named "`cbc_ctos_inv.dat`" includes the results from blood collected at a *screening* visit (and, for the OS participants, Annual Visit 3) and analyzed at each Clinical Center's (CC) local laboratory. Data are missing if the lab was unable to process the sample. Values were reported for the following tests: white blood cell (WBC) count (Kcell/ul), platelet count (Kcell/ul), and hematocrit (%), and, when available, hemoglobin (gm/dl).
    - NOTE: WBC differential data are not available in the WHI Clinic CBC Results. While individual clinics may have assessed WBC differential, the results 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 may still exist.
  - Careful inspection of the CBC result data is recommended before using these results in analyses.

- *Core Analyte Results (CT Subsample)*
  - The "`core_ctos_inv.dat`" 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 Core Analyte subsample includes approximately 8.6% of the HT and 4.3% of the DM participants. Also included in the data file are the results from the participants in the Observational Study Measurement Precision Study (OS-MPS). The OS-MPS blood was drawn within 6 months after the baseline blood draw and includes approximately 1% of the OS.

- Because the sub-sampling 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.

4. Use/Interpretation of the 'Specimen Result Descriptions' Reports

*Use of the **Specimen Result Descriptions***

This site is limited to descriptions of specimen results that are included in the Specimen Tests file. Test data from Core/Ancillary/Broad Agency Announcement studies are eligible for release one year after a study's end date.

- Clicking on Specimen Result Descriptions will bring up a list containing all of the Tests included in the dataset. Use the 'Search' field to find the Test of interest or scroll 'Next' to view Tests.
- If you know the WHI Study number (e.g., W58) of interest, you may also search by Study.
- Click on the Test name to view a summary of the participant results and the Quality Assurance information for the Test ('[Test Name] Result Details').

*Interpretation of the **Result Details** Report*

- When you click on a Test Name, the first report that is displayed is the Result Details Report. (E.g., this is the Result Details Report for CRP.) The Result Details Report has two charts: (1) Result Distribution and (2) QA Measures.
  - The ***Result Distribution chart*** presents a synopsis of the participant result data for each Version of a Test. (If there is only one Version, it means that no other lab, sample type, or units of measure has been used for this test.) The Result Distribution Box-and-Whiskers chart shows the minimum and maximum values for each test version (the whiskers) and the middle 50% of the values (the boxes). If the units of measure vary by version, this is noted in red on the chart. The number on the X-axis represents the Version number.
  - The ***QA Measures chart*** presents two measures of quality for each Test Version: the Coefficient of Variation (CV%) and the Correlation Coefficient (Corr) or Log Correlation Coefficient (Log Corr). (Click here for a description of the WHI quality measures.)
  - For tests with positively skewed data (skew ≥+2), the ***Result Distribution chart*** and ***Correlation chart*** are on the multiplicative (log-transformed) scale.
- The table below the charts shows more details about each Test Version (lab, specimen type, units of measure, number of WHI participant samples with results (N), the median of the participant results (Med), the average of the participant results (Avg), standard deviation of the participant results (Stddev), the skew of the participant results (Skew), the number of QA duplicate pairs tested (Dups), the Correlation Coefficient for the QA pairs tested (Corr) or Log Correlation Coefficient (Log Corr)*, the average Coefficient of Variation for each QA pair (Avg cv%), the maximum Coefficient of Variation for any QA pair (Max cv%), and the minimum correlation for any Pull of samples tested (Min pull Corr). *(Note: A "Pull" is a set of vials pulled from the repository and shipped to one or more labs.*)
- By clicking on a Version number in the table or on the chart, or by clicking a point of interest on a chart, a 'Version Details' page will open with Version-specific details.

Please note:

1. Correlation Coefficient (Corr) or Log Correlation (Log Corr)*. The Corr or Log Corr is calculated for all blind duplicate pairs included in a sample pull. Hence, the Corr or Log Corr is an overall measure of the linear association between blind duplicate paired samples.
   *Log Correlation Coefficients are calculated instead of Correlation coefficients for tests with positively skewed data (skew ≥+2).

2. Coefficient of Variation (CV%). The CV% is calculated for each blind duplicate pair (i.e., standard deviation of the two results ÷ the average of the two results * 100). The table presents the Average CV% (which is the average of the individual pair CV%s) and the Maximum CV% (which is the highest individual CV% for a single blind duplicate pair).

*Interpretation of the **<u>Version Details</u> Report***

- When you click on a Version number or point of interest in the Result Details Report, a '<u>Version Details</u>' page will open with Version-specific details. (E.g., this is the Version Details Report for CRP Version 4.)
- At the top of the Version Details page, information about the Test Version is provided (i.e., Lab, Specimen Type, Test Method, and Comments).
- Version Details may be viewed 'By Pull' or 'By Study'. If you are interested in details for only one Study, click the 'By Study' tab. If you are interested in details for all of the results for this Test Version, click the 'Results by Pull' tab.
- The Version Details Report '**By Pull**' presents similar information as the Result Details Report, only at the level of Pull rather than Version.
- The Version Details Report '**By Study**' presents a table of facts about the participant results and QA measures for the entire study.
- By clicking on the number in the "Dupes" column in the 'By Pull' QA Measures table, additional detail will be presented: Duplicate Sample Results Report.

*Interpretation of the **<u>Duplicate Sample</u> Results Report***

- When you click on the number in the "Dupes" column in the 'By Pull' QA Measures table, additional detail will be presented in the <u>Duplicate Sample Results Report</u>. (E.g., this is the Duplicate Sample Results Report for CRP, Version 4, Pull 207-1.)
- The Duplicate Sample Results Report presents a scatter chart of the test results for the blind duplicate pairs included in the sample Pull of interest and a table showing the CV% for each blind duplicate pair and the test values for each QA sample tested.

Note: WHI includes 10% Quality Assurance (QA) Blind Duplicate samples (5% pairs) in each sample Pull. For more information about the WHI QA Procedures and interpretation of the QA Measures charts, see 'Description of WHI QA procedures'. Also, please see #7.a., below, for a full description of the WHI Blind Duplicate QA samples.

5. <u>Statistical methods for biomarker data</u>

The specimen test results data come from numerous studies that analyzed biospecimens for a subset of WHI participants. Various sampling schemes were used to select participants for each individual study, primarily nested case-control, case-cohort or cohort only. If you plan to use results from only one study, then the statistical methods used should be appropriate for the study design. For example, a conditional logistic regression model could be used for a nested case-control study. However, if you plan to combine results from more than one study, from studies with different designs, or you plan to use a set of combined results for a different outcome than the studies intended, appropriate actions need to be made in the analysis to avoid biased results.

For example, there were three nested case-control studies in the CaD trial designed to examine the association between serum vitamin D (25(OH)D) and breast cancer, colorectal cancer, and fracture. If you plan to use all available 25(OH)D results from these studies to determine if there is an association between 25(OH)D and diabetes, it is recommended that you use inverse probability weighting so that the sample for the analysis is representative of the WHI population as a whole in terms of the rate of incident diabetes. But since the case group for most of the outcomes in these studies constitutes only a small fraction of the cohort, and would be down-weighted to such an extent as to have little influence on the results, use of the control data only is recommended. An alternative could be to analyze the case and control groups separately for associations with other outcomes. However, the analyses in the case group would lack a ready interpretation, but could add to the overall presentation.

Analysts might also consider inter-laboratory standardization of biospecimens. A method that uses linear regression to account for gross systematic differences in location (e.g., mean) and/or scale (e.g., standard deviation) of measurements between labs are described in Standardization of Assays from Multiple Labs for a Given Analyte.

The aforementioned methods of inverse probability weighting and inter-laboratory standardization can be used alone or in conjunction with each other, or other methods. Analysts should perform analyses that are within their abilities, which yield statistically valid estimates. Study specific variables are provided in the "Study/Case/Control/Types" dataset that indicate the study each sample came from, and the type of sample (i.e., case, control, cohort). These variables along with other relevant variables (e.g., age at enrollment, race/ethnicity) can be used to take the appropriate statistical actions.

References (* less theoretical):

- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the Whole Cohort in the Analysis of Case-Cohort Data. *American Journal of Epidemiology* 2009; 169 (11): 1398–1405.
- Hernan MA, Robins JM. Estimating Causal Effects from Epidemiological Data. *Journal of Epidemiology and Community Health* 2006; 60: 578–596.
- Mark SD, Katki HA. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (sampled) cohort studies with missing case data. *Journal of the American Statistical Association* 2006; 101(474):460-471.
- *Reilly M, Torrang A, Klint A. Re-use of case-control data for analysis of new outcome variables. *Statistics in Medicine* 2005; 24:4009-4019.
- Robins JM, Hernan MA, Brumback BA. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 2000;11: 550-560.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; 89(427): 846–866.

6. Examples

1. An investigator wants to know if an Ancillary Study has already been done assessing the relationship between CRP and colorectal cancer.

   The best source of information to address this question is the Biospecimen Analytes by Outcome and Study ID. Unlike the Specimen Result Descriptions, this sortable/filterable list includes all Approved, Funded, and Completed studies, not just those whose biomarker data are eligible for release at this time.

2. An investigator wants to propose an exome chip genotyping project on as many participants as possible who have data on lipids, glucose, and insulin.

   Information to help with this proposal can be found in the Specimen Result Descriptions, which has counts of participants with test results that are currently available in the investigators' dataset as well as details about the tests. *What is not available are counts of participants with multiple test results*.
   - *If the investigator has WHI Data Acce*ss, s/he can use the Query Builder to obtain the information required.
   - *If the investigator does not have WHI Data Access*, s/he will need to search for the tests one at a time to check for counts and test details – and estimate the number of participants for the proposed exome chip project.
     a) In www.whi.org, go to the **Researchers/Data/Dataset Documentation/Specimen Results** page and click on Specimen Result Descriptions.
     b) Search for one of the tests of interest, e.g., Insulin.
     c) Note the number of Test Versions and review the results presented to get a feel for the differences among the Versions.
     d) One at a time, click on a specific Version to review the details for that Version.
     e) Decide if all the results for a Version would be of interest, or if some sample Pulls for a Version would need to be excluded (e.g., due to poor correlation between the blind duplicate QA samples).
     f) Add up the number of participants with insulin results that are likely to be usable.
     g) Move on to the next test.
     h) Contact the PI of a study if there are questions about the study's test method, participant selection, etc.

3. An investigator is considering forming a collaboration with the PI of Study BA9 (i.e., Broad Agency Announcement Study #9) to look at hormones, but she wants more information about the test results of BA9.

Information about the tests for a particular study can be found in the Specimen Result Descriptions by searching for the WHI Study Number, in this case BA9.

a) In www.whi.org, go to the **Researchers/Data/Dataset Documentation/Specimen Results** page and click on Specimen Result Descriptions.
b) Search for BA9.
c) Note the number of Tests and review the information presented.
d) One at a time, view the details for each Test of interest that was performed in BA9, viewing details by Pull as desired.
e) Contact the BA9 PI if there are questions about the study's test method, participant selection, etc.

4. An investigator has an approved paper proposal to use all existing Vitamin D test data in a paper about breast cancer.

Before performing the data analysis using Vitamin D test data, an investigator is strongly advised to review the Test Version details to decide whether or not all of the Vitamin D data should be used in the analysis.

a) In www.whi.org, go to the **Researchers/Data/Dataset Documentation/Specimen Results** page and click on WHI Data, and then Specimen Result Descriptions.
b) Search for Vitamin D.
c) Note the number of Test Versions and review the results presented.
d) One at a time, click on a specific Test Version and view the Version details.
e) For each version, decide if any of the Vitamin D results should be excluded from the analysis due to quality or other concerns.
f) Contact the study PI if there are questions about a study's test method, participant selection, etc.
g) After signing a Data Distribution Agreement and obtaining access to the WHI investigators' dataset, complete the project.

Please see #8 for a hypothetical example using the WHI Specimen Results Data Files.

7. Questions and Answers

a) What are WHI Blind Duplicate Quality Assurance (QA) samples?
The WHI blind duplicate QA samples are not split participant samples. Instead, they are samples from women who went through the initial WHI screening, but for various reasons were not enrolled. WHI has IRB approval to use biospecimen from these non-enrolled women only as QA samples. No covariate data are available for them.

WHI includes blind duplicate QA samples in every sample pull sent to a laboratory for testing. The members of a blind duplicate pair have different Draw IDs, and their vial labels look identical to the other samples in a pull. Hence, the testing labs are 'blind' as to which samples are QA samples. The QA samples are evenly distributed throughout a sample pull.

b) How is the Coefficient of Variation Percent (CV%) calculated?
A CV% is calculated for each blind duplicate pair by taking the standard deviation for the results from two members of a pair, dividing that by the average of the two results, and multiplying by 100. The Average CV% is calculated by taking the average of the CV%s for the individual blind duplicate pairs that were tested with the Version (for the Version CV%) and for the Pull (for the Pull CV%).

Notes:
i. CV% for a blind duplicate pair is only calculated when there is a valid test result for both members of the pair.
ii. A few studies used samples from a QA Pool rather than the Blind Duplicate QA samples.

c) What test result data are currently available to qualified investigators?
- WHI Clinic CBC results (i.e., WBC count (but not WBC differential), hematocrit, and platelet count on both OS and CT participants at baseline and on the OS participants at Annual Visit 3)
- CT Subsample Core Analyte Data
- Data from Ancillary and BAA Studies with funding End Dates at least one year prior to the current data release

- Data from Core biomarker projects designed to enhance the use of WHI Data (e.g., SHARe Biomarkers [W54] and Hormone Trial European American Biomarkers [W58])
- For availability of specific analytes, see Specimen Result Descriptions
- WHI Long Life Study test results on ~7,400 participants whose blood was collected at ~18 years post enrollment were included in the Fall 2013 release: CBCs, lipids, CRP, Creatinine, glucose, and insulin. For a description of the Long Life Study, please see this page.

d) *When will the test result data be updated?*

- Anytime there is a new data release to the investigators dataset.

e) *It says on WHI Biospecimen Analytes by Outcome and Study ID that xx,xxx participants have had YYY test done. Why are there only x,xxx results available for YYY?*

The 'Biospecimen Analytes by Outcome and Study' shows an *approximate* count of all tests from *approved*, funded, and completed studies. The only test results that are available to release in the investigators' dataset are from studies that had a funding end date of at least a year ago.

f) *How was the WHI blood collected/processed/stored?*

- WHI 1993-2005
- Long Life Study (2012-2013)

g) *Some participants have more than one value for a test. Which one should I use?*

- There are times when a lab provides more than one result for a test for a participant. There are other times that a participant's blood has been tested for the same analyte by more than one lab. It is up to you to decide how to handle the multiple results.

h) *The Test Method is missing for the test I am interested in. How can I get it?*

- WHI Studies and Laboratories did not always report their test method to the WHI Clinical Coordinating Center, so some test methods are missing. The best way to obtain missing test method information and other detailed information is to contact the PI of the Study that generated the data. (Contact the WHI Help Desk, helpdesk@whi.org, to request PI contact information.)

i) *How can I get access to the biomarker data or the Query Builder?*

Please refer to these pages:

- Propose a Paper
- Plan a Study

j) *I have more questions. Who do I contact?*

- Check the Study Page for additional information about a study that generated test results (https://www.whi.org/studies).
- For specific questions about laboratories, test methods, participant selection criteria, and test results, contact the PI of the Ancillary Study that generated the test results. Contact the WHI Help Desk, helpdesk@whi.org, to request PI contact information.
- For questions about biospecimen collection, processing, storage, etc., contact the WHI Help Desk (helpdesk@WHI.org).
- To obtain the contact information for an Ancillary Study PI, contact the WHI Help Desk (helpdesk@WHI.org).

8. Continuation of Example #4 (SAS Code)

```
/***************************************************************************/
/* This is a hypothetical example using the WHI Specimen Results Data Files   */
/*                                                                 */
/* Setting:  analyze cohort of participants with vitamin D (25(OH)D) results; */
/*          based on QC information (see example 4 above), want to exclude   */
/*          results from study W15 - pull #6.                  */
/* Note:     This example assumes that all four data files have been read     */
/* into SAS using the provided SAS code.                          */
```

```
/*****************************************************************************/

** Step 1: use the TESTS data file to get information on the vitamin D tests **;
PROC FREQ DATA=spec_tests_inv ; table TESTABBR ; RUN;
/* note that the abbreviated test name for vitamin D is VITD */
DATA spec_tests_vitd ; set spec_tests_inv ; if TESTABBR="VITD" ;
 KEEP TESTABBR TESTNAME TESTVER TESTVERID SPECTYPE ;
RUN;

** Step 2: merge the information on the vitamin D tests with the specimen results **; **         data file and keep just the results for vitamin D.                  **;
PROC SORT DATA=spec_tests_vitd ; BY TESTVERID ; RUN;
PROC SORT DATA=spec_results_ctos_inv ; BY TESTVERID ; RUN;
DATA spec_results_vitd ;
  MERGE spec_results_ctos_inv spec_tests_vitd (in=inkeep) ;
  BY TESTVERID ;
  IF inkeep ;
/* Based on QC information, remove results from pull #6 in W15 (PULLID=W15-6) */
IF PULLID = "W15-6" THEN DELETE ;
RUN;
/* check that results from W15 - pull #6 are not in file */
PROC FREQ DATA=spec_results_vitd ; table PULLID; RUN;

** Step 3: find out from what study visits blood samples were taken by **;
**         merging in information from DRAWS data file.              **;
PROC SORT DATA=spec_results_vitd; BY ID PPTDRW ; RUN;
PROC SORT DATA=spec_draws_ctos_inv ; BY ID PPTDRW; RUN ;
DATA spec_results_vitd ;
  MERGE spec_results_vitd (in=inkeep) spec_draws_ctos_inv (KEEP=ID PPTDRW DRAWDAYS DRAWVTYP DRAWVY) ;
  BY ID PPTDRW ;
  IF inkeep ;
RUN;
PROC FREQ DATA=spec_results_vitd; table studyid*drawvtyp*drawvy / missing;RUN;
/* Note that studies 105, 181 and BA9 used screening, and W15 used year 1 samples */
```

Important note: If you have determined that you need to add information about case-control status to your merged dataset, proceed to Step 4. Do this only after examining the "`spec_case_control_types_ctos_inv.dat`" dataset to rule out duplicates on ID, STUDYID, and CASEFLG, or to decide how you will handle duplicates if present. For this example, there are no duplicates to be concerned about.

```
PROC SORT DATA=spec_case_control_types_ctos_inv ; BY ID STUDYID CASEFLG ; RUN ;
DATA DUPS;
  SET spec_case_control_types_ctos_inv ; BY ID STUDYID CASEFLG ; IF ~(FIRST.CASEFLG and LAST.CASEFLG); RUN ;
/* check which studies have duplicates on ID, STUDYID, and CASEFLG */
PROC FREQ data=dups; table STUDYID ; RUN;

** Step 4:  add information about case-control type for each study **;
PROC SORT DATA=spec_results_vitd; BY ID STUDYID ; RUN;
PROC SORT DATA=spec_case_control_types_ctos_inv ; BY ID STUDYID; RUN ;
DATA spec_results_vitd ;
  MERGE spec_results_vitd (in=inkeep) spec_case_control_types_ctos_inv ;
  BY ID STUDYID ;
  IF inkeep;
RUN;
PROC FREQ DATA=spec_results_vitd; table studyid*caseflg/missing;RUN;
```

```
/* Note that study 105 was a cohort only, and the others were case-control studies */

/* decide to use only the samples from Study 105 and the controls from Studies 181 */
/* and BA9 since their samples all came from the screening visit */
DATA final_results_vitd ;
  SET spec_results_vitd ;
  IF STUDYID = "105" or (STUDYID in ("181","BA9") and CASEFLG="N") ;
RUN;
/* final checks */
PROC FREQ DATA=final_results_vitd; table studyid*caseflg /missing;RUN;
PROC MEANS DATA=final_results_vitd ; CLASS pullid; VAR testval ;RUN;
```