Documentation for Health ABC Biospecimens Data

1. General description 2

- **a. Biospecimens dataset** 2
- **b. Biospecimens listings** 2

2. Cross reference of variable names with exact source 2

3. Dataset structure and contents 4

4. Condition of data 4

- **a. Important note about Year 1 Hemoglobin A1c (Hg_A1C1) data** 4
- **b. Major differences with previously released data** 5
- **c. Grouping variables** 7
- **d. DRAWFLAGx** 9
- **e. Assay details** 9

5. Dataset index formulation and key variable mapping 10

6. General strategies for manipulating and merging the data 10

7. Special missing value codes 10

Appendix I Complete Blood Count 11

Appendix II Explanation of Drawflags 13

Appendix III Biospecimen Assay Details 14

Appendix IV Mapping of Old Variable Name/Dataset to New Variable Name 15

Appendix V Data Added to Biospecimens Dataset Since Release Version 2012-05-01 19

1. General description

a. Biospecimens dataset

The Biospecimens SAS dataset is now posted on the Health ABC website under Datasets/Documentation, then Current Datasets. This dataset contains all data derived from biospecimen (blood or urine) analyses in Health ABC to date, including complete blood counts done by labs local to the clinics in years 3 and 10/11. Before version 11.0 of the Biospecimens dataset, these data were scattered among a number of datasets (YxRead, YxLabAnc, YxCytok), depending on their origin, visit year of collection, and type. Throughout this documentation, a lower case x denotes any applicable visit year x. As the origin ("core" measurements vs Ancillary Study measurements) became more blurred and irrelevant, and as data began pouring in from samples collected in many different years, this became ungainly and certainly confusing. Despite the need to rename practically every biospecimen variable released to date, we decided that combining visit years, origins, and types (cytokine vs other assays) was the only way to go. This also makes it clearer where variable naming across years has been inconsistent, so it was harder to identify identical assays done in more than one visit year. With this combined dataset, all these problems have been cleared up and a new table documenting what assays have been completed and released has been added (see Appendix III).

Note that after version 11.0 of the Biospecimens dataset was released, the versioning for Health ABC datasets switched to using dates. The next version of the Biospecimens dataset after version 11.0 was version 2012-05-01.

In addition the following SAS files, not specific to any year but updated each time data are released, can be found at the top of the Current Datasets listing:

PH	Participant History File
Formats	SAS Format Library

b. Biospecimens listings

Proc Contents and Frequencies/Univariates are zipped together in a self-extracting document (Biospecimens_Listings) that is available on the Health ABC website under the Current Datasets listing.

Please note that the metabolomics biospecimen data associated with Dr. Rachel Murphy's Health ABC ancillary study (AS13-147AM) are in a separate dataset that is available on the Health ABC study website.

2. Cross reference of variable names with exact source

As mentioned above, almost all previously released variables have been renamed as of version 11.0, either simply by adding one or two digits to the end to denote which visit year the specimen was collected in, or to make the variable names between visit years more consistent. Appendix III gives the variable names for all assay variables, along with grouping variables, where needed to analyze the data (see below). A few other variables are included in the dataset that are not documented in Appendix III.

One group of variables not documented there are the variables COMMENTSx. These variables represent a combination of all previous comment variables used to explain why data are missing from a particular assay for a particular person. For example, when assay values are denoted .H (above assay range) or .B (below assay range), the comment variable gives the upper or lower limit of the assay range, respectively.

Another group of variables not documented in Appendix III are the alert variables that indicate whether an alert was issued for an assay out of the healthy range. For example CALERT1 is the variable that indicates whether a creatinine alert was issued in Year 1.

A complete blood count (CBC) was done by a lab local to each clinic in visit year 3 and again starting a number of weeks into visit year 10 and continuing into year 11 to catch participants missed in year 10. In Year 11, a questionnaire was administered at the time of the blood collection for the CBC. Thus only some of the Year 10/11 CBC data have corresponding questionnaire data. The CBC questionnaire data are included in the Y11Visit data file. The data from the Year 3 blood counts, which were formerly in Y3Read, are included in the Biospecimens dataset and the year 10/11 data are released here for the first time as of version 11.0. None of this was recorded on the original CBC results form that was used to get the local lab reports into the data system. Some time later, the year 3 differentials were input by the Project Office, and these are included, as is a red cell distribution width measurement (RDW3) also collected after the fact from the hard-copy reports.

As of version 2012-05-01, additional data were input by the Project Office, and are included. Some previously-released data has been revised and some data values have changed as part of this effort. Given this, analyses previously run on CBC data should be rerun using the most recent version of the Biospecimens dataset (release version 2012-05-01 or later).

The following variables are new as of version 2012-05-01: mean corpuscular hemoglobin (MCHx), mean corpuscular HGB concentration (MCHCx), red blood cell count (RBCx), WBC differential - bands (neutrophils) (PT_BNDNPHLx), and WBC differential - total neutrophils (PT_NPHLx). PT_BNDNPHLx is not available for Pittsburgh participants at Year 10. Mean platelet volume (MPVx) is also new as of this release, and includes data for Memphis participants only at Years 3 and 10/11.

Finally, the differential variables previously released for Year 3 and RDWx now include Year 10/11 data as of version 2012-05-01. PT_SEGNPHL1011 is not available for Pittsburgh participants at Year 10.

The CBC variables are listed together in Appendix III. Appendix I gives further details of what each variable represents.

Sometimes the CBC sample clumped and the laboratory was unable to get an accurate platelet count, but they were able to give a qualitative reading, such as “sufficient.” When this happened, Pittsburgh’s lab did not report a platelet count, but categorized the platelets as “normal,” “sufficient,” or “insufficient.” Memphis’ lab <u>did</u> report a platelet count, but added a footnote warning that these results may not be reliable. If the platelet count reported by the Memphis lab was ≥50,000, the report warned that platelet clumping was observed and re-evaluation may be helpful. If the count reported was <50,000, the report warned that an accurate platelet count could

not be determined. In the event that the qualitative reading was “normal” or “sufficient” (Pittsburgh) or ≥50,000 with a warning (Memphis), the clinics were asked to record -66 for CHPLATE. If the qualitative reading was “insufficient” (Pittsburgh) or <50,000 with a warning (Memphis), they were asked to record –55. These values have been replaced by the special missing value codes .T and .U, respectively.

If mean corpuscular volume was reported to decimal accuracy, the Field Centers were asked to round to the nearest whole number before recording MCV on the Year 3 CBC Results form.

When the CBC original results were searched for the second time to transcribe WBC differential, these original results could not be found for 428 people. Percent monocytes was either missing or not transcribed for one additional person. When there was a complete blood count but incomplete or missing WBC differential, the missing WBC differential variables were set to .M.

Finally, DRAWFLAGx (formerly LCBRFLAG, see below), is a multi-digit binary flag indicating which blood draw (first/repeat) each core assay result came from, where x denotes the year the specimen was collected. Appendix II gives a detailed explanation of what each digit represents for each year’s variable.

For analysts accustomed to the old variable names and dataset locations, Appendix IV maps old variable name and dataset to new variable name.

3. Dataset structure and contents

The BIOSPECIMENS file contains a single observation per participant.

Key variable:

HABCID HABC Enrollment ID without the 2-letter prefix

4. Condition of data

a. Important note about Year 1 Hemoglobin A1c (Hg_A1C1) data

Since the release of version 11.0 of the Biospecimens dataset in February 2010, the Coordinating Center has learned of analyses at Yale and the Pittsburgh Field Center that have uncovered a problem with the Year 1 hemoglobin A1c (HG_A1C1) data. Hemoglobin A1c levels tended to be higher at baseline than at Year 4. The assays were performed at the same lab but using different methods. After thorough investigation between the lab and the manufacturer, it was determined that an error may have been made during the baseline measurements that did not allow unstable Schiff bases to be completely removed before the assay. In addition, the baseline assay used a 1-point calibration, whereas the Year 4 assay used a 2-point calibration, giving greater precision. The Coordinating Center is working with the laboratory to determine whether these values are suitable for use in cross-sectional Year 1 analyses, but they should not be used in longitudinal analyses. If it is determined that a calibration can be applied to the Year 1 values to make them comparable to Year 4 and later, the data will be re-released with the calibration applied. Until then, analysts should be wary of using the Hg_A1C1 data in analyses.

b. Major differences with previously released data

As of release version 2012-05-01, a listing of new data added to the Biospecimens dataset can be found in Appendix V.

The following variable names were changed as of release version 2012-05-01 in order to make the variable names more accurate: 1) VITD250H_2 (Y2 Serum 25(OH)Vitamin D (ng/mL)) was changed to VITD25OH_2; 2) IGF1 (Y1 Serum insulin-like growth factor 1 (ng/mL)) was changed to IGF1_1; 3) HR8GLUC1 (Y1 Fasting serum glucose, 2 hour post-GTT (mg/dL, .T if not fasting)) was changed to HR2GLUC1. Several variables had a very small number of participants where the assay value was recorded as zero. When appropriate, these zeros have been changed to a special missing value and the assay detection limit noted in the COMMENTSx variable. The variable HG_A1C4 had the value of "U" for 38 participants. This value has now been explained in HGCODE4 as "comprised specimen, use data with caution." Analysts using data from this variable should be wary of using the HB_A14 values for these participants. Also, as noted in section #2 above, changes have been made to some CBC data and new CBC data has been added to the dataset. Given this, analyses previously run on CBC data should be rerun using the most recent version of the Biospecimens dataset (release version 2012-05-01 or later). All of these changes were made as of release version 2012-05-01.

All assay variables have had a number appended to the variable name to indicate the visit year in which the specimen was collected as of release version 11.0. When the variable name already ended in a number, or simply adding a number might cause confusion, the visit year number was preceded with an underscore. In addition, many variables have been renamed so that all assays of the same or very similar analytes would fall together in an alphabetic list such as the Proc Contents. Sample type (serum, plasma, citrated plasma, or urine) and units have been added to all labels for clarity. In most cases, urine assays are reported as normalized to urinary creatinine. Previous raw (non-normalized) values have been dropped. The exception to this is the Year 1 urinary data for Dr. Mark Sarnak's ancillary study AS09-121, where both normalized and non-normalized values were released.

"Fasting" glucose was measured in Years 1, 2, 4, 6, 10, and 11; "fasting" insulin in Years 1, 4, 10, and 11; and "fasting" lipids in Years 1, 2, 4, 6, 8, 10 and 11. Participants were instructed to fast for 8 hours before coming to the clinic and blood was drawn before they were given anything to eat. Not surprisingly, not all participants followed these instructions, and some returned for a second draw on a different day after fasting for at least 8 hours. In these cases, the fasting value was chosen over the non-fasting value (see also DRAWFLAGx below). In other cases, no fasting value was obtained. To deal with these non-fasting "fasting" glucose and insulin values in past datasets, a second variable, FAST8GLU(x) or FAST8INS(x), respectively, was added in which the non-fasting values of glucose and insulin were set to .T (missing due to technical reasons). However, the original variables, FAST_GLU and FAST_INS remained somewhat confusing, as they actually represent a mixture of fasting and non-fasting values. These original (mixed fasting and non-fasting) variables have now been renamed GLUCOSEx and INSULINx, respectively. As of version 11.0, additional new variables, HR8GLUC1, FAST8CHOLx, FAST8LDLx, FAST8HDLx, and FAST8TRIGx, have also now been added to accomplish the same goal for 2-hour glucose and lipids. However, since the original variables did not imply fasting, their variable names have not changed, except to add the year suffix. 34 Year 4 participants were incorrectly classified as not

having fasted in FAST8GLU4 and FAST8CHOL4. These participants did fast for their Year 4 blood draw, and FAST8GLU4 and FAST8CHOL4 were corrected as of version 2012-05-01. Also, several of the Year 11 fasting variables (FAST8xxxx11) had inadvertently not been populated with data values, this has been corrected as of version 2012-05-01.

More importantly, in some cases, assays were reported in one unit (e.g., pg/mL) in one visit year and in a different unit (e.g., pmol/L) in another visit year. A major effort was made to convert the assays from one visit year to the units of the other visit year, to make year-to-year comparisons more straightforward. <u>However, not all assays in different years were done using the same methodology, and thus the analyst must ensure that the values are comparable across visit years</u>. Details of the analysis method can be found in Appendix III by clicking on the link in the Laboratory column. In a few cases, where the lack of comparability was well understood, the data have been standardized across years already. For example, the Years 4 and 11 insulin data were converted from the raw values to values comparable to Years 1and 10 using a calibration equation (Years 4 and 11 values multiplied by 0.6544) provided by LCBR, the core lab, which did the assays in all four years.

In contrast, Year 1 cystatin C and Years 3, 4, 6, and 10 cystatin C, are known to be <u>non</u>-comparable, but have not yet been standardized to each other. Due to a standardization change on the part of the instrument and reagent manufacturer, years 3, 4, 6, and 10 cystatin C values run approximately 17% lower than year 1 values. The central lab recommends that the years 3, 4, 6, and 10 values be considered "correct" (although without a consensus standard, true "correctness" cannot be determined), and prior values be adjusted by a harmonization factor of 1.2. However, investigators working with Health ABC cystatin C data may decide to work out their own process in collaboration with the LCBR for adjusting levels over time. Because these variables as released are known not to be usable in longitudinal analysis without calibrating one against the other, they have been given different variable names (CYSTAT_UNCAL1, CYSTATIN3, CYSTATIN4, CYSTATIN6 and CYSTATIN10).

Similarly, serum creatinine in Year 1 and Years 3 and 10 are known to be non-comparable and have not yet been standardized to each other. Years 3 and 10 creatinine was measured with a method that is traceable to IDMS standard, but year1 creatinine was not. The central lab recommends that a harmonization process be done for creatinine to know what harmonization factor to use. One example of such a process would be to re-run year1 samples with the current IDMS-standardized method. Discussions are in process of how best to accomplish this goal. Again, for clarity, the Year 1 and Years 3 and 10 variables have been given different variables names (CREATIN1 and CREAT_IDMS3, CREAT_IDMS10, respectively).

Two cytokines (CRP and IL6) were measured in three different sample types (serum, citrated plasma, and EDTA plasma) by two different labs over multiple years. Yongmei Liu and Kurt Lohman, who provided the Year 2, 4, and 6 values (measured at Wake Forest on EDTA plasma samples), also developed a calibration based on pilot studies to convert the Year 1 values (measured at LCBR in serum) and the Year 8 values (measured at Wake Forest on citrated plasma) to values comparable to those measured in the other 3 years. In this case, the raw data (CRP_LCBR1, IL6_LCBR1, CRP_CITR8, IL6_CITR8) were left in the dataset for cross-sectional analyses, and calibrated variables for longitudinal analysis (CRP_CALIB1, IL6_CALIB1, CRP_CALIB8, IL6_CALIB8) were added. Note that negative values for CRP_CALIB1 are expected. In year 2,

LCBR measured IL6 and CRP only in a subgroup, while Wake Forest measured the entire cohort, so the _WAKE version of the data should be used for longitudinal analysis, and no calibrated version was required. In Year 10, Wake Forest measured IL6 on serum only in a subgroup (IL6_WAKE10), and no calibration version was required.

c. Grouping variables

While many laboratory analytes measured in Health ABC Ancillary Studies were measured on all participants with available samples, some were measured in only one or more subgroups, depending on the investigator's study design. Analysts using these data for other analyses should thoroughly understand the way the participants measured were selected and be very careful how they handle the selection bias inherent in these designs. Details of the sub-samples are given in Appendix III. Some studies exclude a large group of participants from the selection process, and therefore may not be suitable at all for analyses outside the original study design. For example, in AS01-13, which measured serum bone-specific alkaline phosphatase, serum calcium, serum CTX, and serum PTH in Year 1, samples were chosen according to a case-control design, with cases of diabetes being defined as those taking a diabetes drug or with fasting glucose higher than 126 mg/dL. Normal-glucose controls were defined as those who were not taking a diabetes drug and fasting glucose less than 110 mg/dL, 2-hour glucose tolerance test less than 140 mg/dL, and no self-report of diabetes. Anyone in between these groups, or who was not fasting, or who was missing fasting glucose or 2-hour glucose values, or whose Hg_A1c values were missing, or who was categorized as having fetal hemoglobin, was excluded from consideration. And of course, the 29 participants who did not have baseline biological specimens were also excluded. This amounted to 928 or 30% of the Health ABC study cohort being excluded from both groups. For AS01-13AM, which measured serum PINP, urinary NTX-1, urinary microalbumin, and urinary creatinine, 10 cases and 102 controls from the previous design were excluded because it was discovered that they had not been fasting long enough before the fasting glucose measurement, so that a total of 1030 or 33.5% of the Health ABC study cohort were excluded from consideration. Therefore, no analysis is possible outside the original case-control design. Three variables, DIAB_SGRP, DIAB_UGRP, AND PINPGRP, have been created to allow proper categorization of participants by their case/control status. The correct grouping variable must be used for each analyte, depending on which selection paradigm was used for that analyte. The SAS label of these variables lists which analytes were measured with each selection scheme.

In those cases where the study design is more straightforward, it is still necessary to control for the selection process in the analysis. Indicator variables have now been added for all analytes that were not measured in the entire cohort (other than loss due to missing samples or missed visits). **To properly understand these indicator variables, it is necessary to download and apply the latest SAS formats catalog.** Each indicator variable includes a label that describes which Ancillary Study and analyte(s) it applies to and value labels that describe what weight must be applied to that group during analysis to make inferences about the Health ABC population in general. For example, IGFGRP has the label "Colbert AS01-24AM stratified sample (measured IGF1, IGFBP1_1, IGFBP2_1)." Since for this Ancillary Study the two strata are 1) all those with non-missing values of the covariates of interest, and 2) a random sample of the remaining participants, the values of IGFGRP are "1=Has non-missing covariates, wt=1.00" and "0=Random sample of rest of participants, wt=7.68." To use these data in an analysis claiming to represent the Health ABC population in general, the analyst should give participants with IGFGRP=0 a weight of 7.68 and participants with IGFGRP=1 a weight of 1 **(note that the usual weight option in SAS can give**

incorrect results). Weighting of case-control and case-cohort data are more complicated, and the investigator not well versed in this type of analyses are advised to consult with someone who is, before embarking upon analysis of data using any of these non-random study designs.

Some of these subgroup analyses cover more than one year. For example, a number of cytokines and their receptors were measured in a random sample of 502 participants in Year 1, plus in both Year 1 and Year 2 for a sample of participants functionally limited by 6, 12, 18 or 24 months, plus matching non-limited controls. The grouping variables SRRANDOM and SRGRP indicates which participants fall into each of these groups. When these variables apply to more than one visit year, the analytes are listed in the labels with "x" in place of the visit-year indicator in the variable names.

In at least one case, a sample group was chosen from baseline attributes, and an analyte (the 65kD isoform of glutamic acid decarboxylase) was measured in multiple years (Years 1, 2, 3, 4, and 6). In this case, the grouping variable GADGROUP was created from the original sampling design, and the weight relates to the proportion of participants in each subgroup at baseline. As participants were lost to follow up over the succeeding years, the proportions may have changed differentially between groups. The user is cautioned to use some of the modern missing data techniques, such as multiple imputation, to deal with this issue separately. At a later date a randomly selected subset of gad65aa positive and negative participant from the GADGROUP were tested for c-peptide, and this data has been released.

For those wishing to avoid the complication of weighting their analysis, it is sometimes possible to subset the data for some analytes to participants who were chosen as a random sample of the entire Health ABC cohort. For example, SRRrandom=1 will pick out the 502 participants who were randomly sampled for soluble cytokine receptor assays at baseline. No weighting is necessary if the analysis is restricted to this group. In an ordinary case-cohort study, the subcohort can be used in a similar way. However, although FetSubC and IsopSubC have been created to identify the subcohorts for the fetuin and F2-isoprostane assays, respectively, the subcohort in both of these studies was not a straightforward random sample. In both cases, the Health ABC cohort was first stratified by race, gender, and prevalent diabetes status. For each race/gender subgroup, the number of participants with prevalent diabetes was restricted. Thus, there are 8 substrata of the subcohort in these studies, and weights must be applied even if only the subcohort will be used in analysis.

When two grouping variables exist for an assay, it is important to understand the difference. In general, one gives a category for everyone who has assay results, and the value labels give the appropriate weights to use when all participants will be analyzed. The other gives a category only for those in the subcohort. The weights are different between these two variables for corresponding categories because the cases are first removed from the subcohort to make a non-redundant category when all participants are included. This changes both the numerator (the pool of participants available to select in that category) and the denominator (the number selected) used to calculate the weights.For more information about sampling, please see:

Lee AJ, McMurchy L, Scott AJ. Re-using data from case-control studies. Stat Med. 1997 Jun30;16(12):1377-89. PMID: 9232759.
Scott A, Wild C. Case-Control Studies with Complex Sampling. J R Stat Soc Ser C Appl Stat. 2001;50(3):389-401.

d. DRAWFLAGx

Some participants returned for repeat blood draws for a variety of reasons (not fasting at the first visit, problems with the draw, etc.) and consequently had two records in the core lab dataset. These records have been combined to give one set of results for each participant using the following rules:

1. Missing values were replaced by valid assay results wherever possible.

2. When there was a non-missing value for the same assay in both records, the fasting value was taken.

3. A multi-digit binary flag (DRAWFLAGx) was created to indicate which blood draw each assay result came from. 1 indicates that the repeat blood draw or repeat urine specimen was used, while 0 indicates that the original (day-of-clinic-visit) specimen was used. A complete list of which digit represents which assay can be found in Appendix II.

e. Assay details

The assay documentation has been totally revamped to make it more user friendly and easier to update when new assays are added to the dataset. Appendix III is a report based on the Biospecimens Tracking log that gives the following information for each assay:

1) Visit year in which the specimen was collected
2) The SAS variable name (there may be more than one listed if there are related variables for the same analyte, e.g., GLUCOSEx and FAST8GLUx)
3) The type of biospecimen (urine, serum, plasma or citrated plasma – when "plasma" is noted in this column or SAS labels, EDTA plasma is implied)
4) The analyte measured
5) Unit of measure
6) The Health ABC sub-sample in which the assay was done
7) The grouping variable that describes that sub-sample, when appropriate
8) The blind duplicate reproducibility. In most cases, investigators were required to include a 5% duplicate sample (provided in a blinded manner by the Coordinating Center), and the resulting duplicate results have been analyzed. The coefficient of variation (CV) of the paired results is given in this column. Usually, the CV is based on the original variable. If the distribution of the original variable is skewed, the CV may be based on the log transformation of the original variable.
9) Skewness. For the convenience of the analyst and to make sense of the high CVs found for some assays in the previous column, the skewness result from a SAS Proc Univariate for each variable is included (when more than one version of the variable is listed in Appendix III, the skewness value is for the "raw" first value listed). Each analyst will have their own preference for how to deal with highly skewed data, so the raw assay results are released without transformation. Any highly skewed data should be log transformed; used as quantiles, with cutpoints, or in a non-parametric analysis; or dealt with in some other way when these data are being used as dependent variables or when extreme values are driving the linear relationship obtained.
10) The investigator for whom the assay was conducted (or Core, for core Health ABC assays). Clicking on the investigator's name will provide a link to the investigator's last known email

address.

11) The Ancillary Study reference number of the Ancillary Study for which the assay was done. Clicking on this link will bring up a copy of the Ancillary Study Proposal.
12) The laboratory at which the assay was done. Clicking here will bring up a description of the assay methodology.
13) The date the dataset was released.

5. Dataset index formulation and key variable mapping

The BIOSPECIMEN file is sorted by HABCID, which is a unique identifier for each participant.

6. General strategies for manipulating and merging the data

Because the Health ABC datasets are sorted by Health ABC Enrollment ID, the HABCID variable is most useful for merging with other datasets.

7. Special missing value codes

SAS allows for stratification of missing values. The following special missing values have been assigned:

. : Missing Form
Used when a value is missing because the entire form has not been entered or the participant does not exist in the database from the corresponding Reading Center.

B: Below Assay Range
Used when a measurement had such a low detection level that the value was recorded as less than a specific number instead of an exact number. An effort has been made to supply the assay limit in the COMMENTSx variable corresponding to the visit year of the analyte.

H: Above Assay Range
Used when the concentration of an analyte was so high that the value was recorded as above a specific number, instead of an exact number. Again, an effort has been made to supply the assay limit in the COMMENTSx variable corresponding to the visit year of the analyte.

M:Missing
Used to flag missing values when the value is required (i.e., true missing values).

T:Missing Due to Technical Problems
Used when a value is missing from the Reading Center dataset due to technical difficulties. For example, some values have to been set to .T when the participant was not fasting (e.g., FAST8GLUx). In other cases, the laboratory's internal QC indicated a problem with the assay value and the value was not reported to the Coordinating Center by the lab.

U:Unacceptable
Used with certain Reading Center data when the data exist but cannot be used; for example, assay data that do not meet internal CV standards at the laboratory.

Appendix I Complete Blood Count

Variable Name	Variable Description	Variable Label	Value Label	Year
CBCYR	Visit at which Y10/11 CBC was done	Year (10 or 11) CBC was done)	10=Year 10 11=Year 11	10/11
HCTx	Year x Hematocrit	Yx Hematocrit (%)	percent	3, 10/11
HGBx	Year x Hemoglobin	Yx Hemoglobin (g/dL)	g/dL	3, 10/11
MCHx	Year x Mean corpuscular hemoglobin	Yx Mean corpuscular hemoglobin (pg)	pg	3, 10/11
MCHCx	Year x Mean corpuscular HGB concentration	Yx Mean corpuscular HGB concentration (%)	%	3, 10/11
MCVx	Year x Mean corpuscular volume	Yx Mean corpuscular volume (fL)	fL	3, 10/11
MPVx	Year x Mean platelet volume	Yx Mean platelet volume (fL)	fL	3, 10/11
PLATEx	Year x Platelet count	Yx Platelets (thousands/uL)	thousands/µL	3, 10/11
RBCx	Year x Red blood cell count	Year x Red blood cell count (millions/uL)	millions/uL	3, 10/11
RDWx	Year x Red blood cell distribution width	Yx Red blood cell distribution width (%)	%	3, 10/11
WBCx	Year x White blood count	Yx White blood count (thousands/uL)	thousands/µL	3, 10/11
PT_BNDNPHLx	Year x WBC differential - bands (neutrophils)	Year x WBC differential - bands (neutrophils)	%	3, 10/11
PT_BSPHLx	Year x WBC Differential – percent basophils	Yx WBC Differential – percent basophils	%	3, 10/11
PT_EOPHLx	Year x WBC Differential – percent esoinophils	Yx WBC Differential – percent esoinophils	%	3, 10/11
PT_LCTx	Year x WBC Differential – percent lymphocytes	Yx WBC Differential – percent lymphocytes	%	3, 10/11
PT_MONOCTx	Year x WBC Differential – percent monocytes	Yx WBC Differential – percent monocytes	%	3, 10/11

Variable Name	Variable Description	Variable Label	Value Label	Year
PT_NPHLx	Year x WBC differential - total neutrophils	Yx WBC differential - total neutrophils	%	3, 10/11
PT_SEGNPHLx	Year x WBC Differential – percent segmented neutrophils	Yx WBC Differential – percent segmented neutrophils	%	3, 10/11

Appendix II Explanation of Drawflags

Variable	Assays in order they appear in flag
DRAWFLAG1	1st digit: albumin 2nd digit: alkaline phosphatase 3rd digit: cholesterol 4th digit: creatinine data 5th digit: fasting glucose 6th digit: fasting insulin 7th digit: HDL 8th digit: Hemoglobin A1c 9th digit: 2-hr glucose 10th digit: LDL 11th digit: triglycerides 12th digit: Hg code
DRAWFLAG2	1st digit: cholesterol 2nd digit: fasting glucose 3rd digit: TSH 4th digit: FT4
DRAWFLAG4	1st digit: HgA1c 2nd digit: cholesterol 3rd digit: fasting glucose
DRAWFLAG6	1st digit: cholesterol 2nd digit: triglycerides 3rd digit: HDL 4th digit: LDL 5th digit: fasting glucose 6th digit: Hg A1c 7th digit: Hg code
DRAWFLAG10	1st digit: cystatin C 2nd digit: fasting glucose 3rd digit: fasting insulin 4th digit: Hg A1c 5th digit: serum creatinine
DRAWFLAG11	1st digit: Vitamin B12 2nd digit: fasting glucose 3rd digit: fasting insulin 4th digit: Hg A1c 5th digit: cholesterol 6th digit: HDL 7th digit: LDL 8th digit: Triglycerides

Appendix III Biospecimen Assay Details

Appendix III can be accessed in the Biospecimens Dataset Documentation section of the Health ABC study website.

Appendix IV Mapping of Old Variable Name/Dataset to New Variable Name

Old Variable Name	Old Dataset	New Variable Name
ACARO	Y6LABANC	ACAROT6
ACAROTENE	Y2LABANC	ACAROT2
ADIPON	Y1LABANC	ADIPON1
ALBUMIN	Y1READ	ALBUMIN1
ALK_PHOS	Y1READ	ALK_PHOS1
ALPHA_T	Y6LABANC	ATOCOPHRL6
APALERT	Y1READ	APALERT1
ATOCOPHEROL	Y2LABANC	ATOCOPHRL2
BCAROT_9CIS	Y6LABANC	BCAROT_9CIS6
BCAROT_13CIS	Y6LABANC	BCAROT_13CIS6
BCAROTENE_13CIS	Y2LABANC	BCAROT_13CIS2
BCAROTENE_TRANS	Y2LABANC	BCAROT_TRANS2
BCRYPTOX	Y2LABANC	CRYPTOX2
CALERT	Y1READ	CALERT1
CHHCT	Y3READ	HCT3
CHHGB	Y3READ	HGB3
CHLRATIO	Y8READ	CHLRATIO8
CHMCV	Y3READ	MCV3
CHOLES	Y1READ,Y2READ,Y4READ, Y6READ,Y8READ	CHOLES1,CHOLES2,CHOLES4, CHOLES6,CHOLES8
CHPLATE	Y3READ	PLATE3
CHWBC	Y3READ	WBC3
COMMENT1	Y1CYTOK	COMMENTS1
COMMENT2	Y1CYTOK	COMMENTS1
COMMENT3	Y1CYTOK	COMMENTS1
COMMENT4	Y1CYTOK	COMMENTS1
COMMENTS	Y3LABANC	COMMENTS3
CPTHCOMM	Y1LABANC	COMMENTS1
CREATIN	Y10READ	CREAT_IDMS10
CREATIN	Y1READ	CREATIN1
CRP	YxCYTOK	CRP_LCBRx
CRPCOM	Y2CYTOK	COMMENTS2
CRYP	Y6LABANC	CRYPTOX6
CYSCOMM	Y1LABANC	COMMENTS1
CYSTATIN	Y10READ	CYSTATIN10
CYSTATIN	Y1LABANC	CYST_UNCAL1
DELTA_T	Y6LABANC	DTOCOPHRL6
DHEAS	Y1LABANC	DHEAS1
ESTRADIOL	Y3LABANC	ESTRADIOL3
F2_ISOPROST	Y1LABANC	F2_ISOPROST1
FAST_GLU	Y1READ,Y2READ, Y4READ,Y6READ,Y10REA	GLUCOSE1,GLUCOSE2, GLUCOSE4,GLUCOSE6,GLUCOSE1

Old Variable Name	Old Dataset	New Variable Name
	D	0
FAST_INS	Y1READ, Y10READ	INSULIN1, INSULIN10
FAST8GLU	Y1READ,Y2READ, Y4READ,Y6READ	FAST8GLU1,FAST8GLU2, FAST6GLU4,FAST8GLU6
FAST8INS	Y1READ	FAST8INS1
FETUIN	Y1LABANC	FETUIN1
FETUINCOMM	Y1LABANC	COMMENTS1
FOLATE	Y3LABANC	FOLATE3
FREE_T	Y1LABANC	FREE_TEST1
FREET	Y3LABANC	FREE_TEST3
FT4	Y2READ	FT4_2
G2ALERT	Y1READ	G2ALERT1
GALERT	Y1READ,Y2READ, Y4READ,Y6READ,Y10READ	GALERT1,GALERT2, GALDERT4,GALERT6,GALERT10
GAMMA_T	Y6LABANC	GTOCOPHRL6
GTOCOPHEROL	Y2LABANC	GTOCOPHRL2
HDL	Y1READ,Y6READ,Y8READ	HDL1, HDL6, HDL8
HG_A1C	Y1READ,Y4READ,Y6READ, Y10READ	HG_A1C1,HG_A1C4, HG_A1C6, HG_A1C10
HGCODE	Y1READ,Y4READ,Y10READ	HGCODE1,HGCODE4,HGCODE10
HG_CODE	Y6READ	HGCODE6
HOLOTC	Y3LABANC	HOLOTC3
HR_GLUC	Y1READ	HR_GLUC1
IGF	Y1LABANC	IGF1_1
IGFBP1	Y1LABANC	IGFBP1_1
IGFBP2	Y1LABANC	IGFBP2_1
IGFBP3	Y1LABANC	IGFBP3_1
IL2SR	YxCYTOK	IL2SRx
IL6	YxCYTOK	IL6_LCBRx
IL6COM	Y2CYTOK	COMMENTS2
IL6SR	YxCYTOK	IL6SRx
ISOLUT1	Y6LABANC	ISOLUT1_6
ISOLUT2	Y6LABANC	ISOLUT2_6
ISOLUT3	Y6LABANC	ISOLUT3_6
ISOZEAX	Y6LABANC	ZEAX_CIS6
LCBRFLAG	Y1READ,Y2READ,Y4READ, Y6READ,Y10READ	DRAWFLAG1, DRAWFLAG2, DRAWFLAG4, DRAWFLAG6, DRAWFLAG10
LDL	Y1READ,Y6READ,Y8READ	LDL1, LDL6, LDL8
LEPTCOMM	Y1LABANC	COMMENTS1
LEPTIN	Y1LABANC	LEPTIN1
LUT	Y6LABANC	TRANS_LUT6
LUTEIN	Y2LABANC	TOT_LUT2

Old Variable Name	Old Dataset	New Variable Name
LYCO_5CIS	Y6LABANC	LYCL_5CIS6
LYCO_13CIS	Y6LABANC	LYCO_13CIS6
LYCO_15CIS	Y6LABANC	LYCO_15CIS6
LYCO_9CIS	Y6LABANC	LYCO_9CIS6
LYCOPENE	Y2LABANC	LYCO_TOT2
OLDL	Y1LABANC	OLDL1
OLDLCOMM	Y1LABANC	COMMENTS1
PAI	Y1CYTOK	PAI1_1
PCT_BSPHLS	Y3READ	PT_BSPHL3
PCT_EOSNPHLS	Y3READ	PT_EOPHL3
PCT_LYMPHCTS	Y3READ	PT_LCT3
PCT_MONOCTS	Y3READ	PT_MONOCT3
PCT_SEGNTPHLS	Y3READ	PT_SEGNPHL3
PCTSATTC	Y3LABANC	PCTSATTC3
PENTOSIDINE	Y1LABANC	PENTOSDN1
RESISTIN	Y1LABANC	RESISTIN1
ROH	Y6LABANC	ROH6
RP	Y6LABANC	RP6
SBONEALP	Y1LABANC	SBONEALP1
SCA	Y1LABANC, Y2LABANC	SCA1, SCA2
SCTXI	Y1LABANC	SCTX1
SHBG	Y1LABANC, Y3LABANC	SHBG1, SHBG3
S_PINP	Y1LABANC	S_PINP1
SPTH	Y1LABANC	SPTH1
SRCOMM	Y1CYTOK	COMMENTS1
STNFRI	YxCYTOK	STNFRI_x
STNFRII	YxCYTOK	STNFRII_x
TALERT	Y1READ, Y6READ	TALERT1, TALERT6
TAP	Y6LABANC	TAP6
TBC	Y6LABANC	BCAROT_TRANS6
THCY	Y3LABANC	THCY3
TLYCO	Y6LABANC	LYCO_TRANS6
TNF	Y1CYTOK	TNFA1
TNFA	Y2CYTOK	TNFA2
TNFACOM	Y2CYTOK	COMMENTS2
TNFRCOM	Y2CYTOK	COMMENTS2
TNFRIICM	Y2CYTOK	COMMENTS2
TOT_TC	Y3LABANC	TOT_TC3
TOTAL_T	Y1LABANC	TOTAL_TEST1
TOTALLYCO	Y6LABANC	LYCO_TOT6
TRIG	Y6READ, Y8READ	TRIGLYC6, TRIGLYC8
TRIGLYC	Y1READ	TRIGLYC1
TSH	Y2READ	TSH2
TSHALERT	Y2READ	TSHALERT2
U_CA	Y1LABANC	Deleted (see U_CACR1)

Old Variable Name	Old Dataset	New Variable Name
U_MALBCR	Y1LABANC	U_MALBCR1
U_MICALB	Y1LABANC	Deleted (see U_MALBCR1)
U_NTXI	Y1LABANC	Deleted (see U_NTXICR1)
U_NTXICR	Y1LABANC	U_NTXICR1
VITB12	Y3LABANC	VITB12_3
VITC_COMMENTS	Y2LABANC	COMMENTS2
VITC_TOTAL	Y2LABANC	VITC_TOTAL2
VITD25OH	Y2LABANC	VITD25OH_2
ZEAX	Y6LABANC	ZEAX_TRANS6
ZEAXANTHIN	Y2LABANC	ZEAX_TOT2

Appendix V Data Added to Biospecimens Dataset Since Release Version 2012-05-01

Version / Release Date	Analyte	Year(s)	Investigator	Health ABC Ancillary Study Reference Number
2016-05-01	Urinary Alpha-1-microglobulin, NGAL, PIIINP, Uromodulin, Chloride, Potassium, Sodium	1	Mark Sarnak	AS09-121AM2
2016-05-01	Aldosterone	1	Amanda Leonberg-Yoo	AS14-154AM
2016-05-01	Myostatin	6	Peggy Cawthon	AS10-135AM
2015-08-01	Vitamin D and Fibroblast Growth Factor 23 with grouping variable	2	Mark Sarnak	AS09-121
2014-09-01	AMD status and 2-kiloDalton (kDa) intervals (KDA_X_X_6, where X_X denotes the interval)	6	Alex Iannaccone	AS07-102
2014-01-01	Apolipoprotein-E (apoE)	2	Eric van Exel	AS12-145
2014-01-01	Klotho	2	Richard Semba	AS10-134
2014-01-01	Insulin-like growth factor 1 (IGF-1)	3	Anne Schafer	AS13-146
2014-01-01	Insulin-like growth factor binding protein 3	3	Anne Schafer	AS13-146
2014-01-01	N-propeptide of type I procollagen (P1NP)	3	Anne Schafer	AS13-146
2013-11-01	Dihydro Vitamin K1	2	Stephen Kritchevsky	AS10-131
2013-11-01	ucMGP	2	Stephen Kritchevsky	AS10-131
2013-11-01	Vitamin K1	2	Stephen Kritchevsky	AS10-131
2013-11-01	Cystatin C	6	Mark Sarnak	AS09-121AM
2013-02-01	C-peptide	1-6	Lewis Kuller	AS04-51
2013-02-01	Creatinine, Creatinine ratios*	1	Mark Sarnak	AS09-121
2013-02-01	Interleukin 18	1	Mark Sarnak	AS09-121
2013-02-01	Kidney injury molecule 1	1	Mark Sarnak	AS09-121
2013-02-01	Microalbumin	1	Mark Sarnak	AS09-121
2012-10-01	Soluble receptor for advanced glycation end products (sRAGE)	1	Richard Semba	AS08-112
2012-05-01	Carboxymethyl-lysine (CML)	1	Richard Semba	AS08-112
2012-05-01	CTX	1	Ann Schwartz	AS07-108
2012-05-01	Fluorescent Oxidation Products	1	Alka Kanaya	AS03-43
2012-05-01	P1NP	1	Ann Schwartz	AS07-108
2012-05-01	Undercarboxylated Osteocalcin	1	Ann Schwartz	AS07-108

Version / Release Date	Analyte	Year(s)	Investigator	Health ABC Ancillary Study Reference Number
	(ucOC)			
2012-05-01	Complete blood count	3	Core	NA
2012-05-01	Creatinine	3	Kushang Patel	AS07-105
2012-05-01	Cystatin C	3	Kushang Patel	AS07-105
2012-05-01	Erythropoietin	3	Kushang Patel	AS07-105
2012-05-01	Ferritin	3	Kushang Patel	AS07-105
2012-05-01	Iron	3	Kushang Patel	AS07-105
2012-05-01	Soluble Transferrin Receptor	3	Kushang Patel	AS07-105
2012-05-01	Vitamin B12	3	Kushang Patel	AS07-105
2012-05-01	Cystatin C	4	Elsa Strotmeyer	AS04-64
2012-05-01	Insulin, fasting	4	Elsa Strotmeyer	AS04-64
2012-05-01	Cholesterol	10	Joseph Hanlon	AS05-73AM
2012-05-01	HDL	10	Joseph Hanlon	AS05-73AM
2012-05-01	Interleukin 6	10	Caterina Rosano	AS05-79AM6
2012-05-01	LDL	10	Joseph Hanlon	AS05-73AM
2012-05-01	Triglycerides	10	Joseph Hanlon	AS05-73AM
2012-05-01	Complete blood count	10/11	Core	NA

*The Year 1 urinary data was released as both non normalized, and normalized to creatinine. The formula used for the microalbumin/creatinine ratio (mg/g) is microalbumin *0.1/creatinine*1000. The formula used for the interleukin 18/creatinine ratio (pg/mg) is interleukin 18/creatinine*0.01. The formula used for the kidney injury molecule 1 /creatinine ratio (pg/mg) is kidney injury molecule 1 /creatinine*0.01.