INCIDENT AND PREVALENT DISEASE ALGORITHMS (PrevIncDz.sas7bdat)

INCIDENT AND PREVALENT DISEASE ALGORITHMS (PrevIncDz.sas7bdat)

1. General description

The PrevIncDz dataset contains calculated variables categorizing Health ABC participants by the presence or absence of specific prevalent diseases at baseline, as well as indicators of the incidence of these same diseases during follow-up.

There are basically two types of incident disease variables, those based on periodic questionnaires (annual contacts), and those based on Health ABC events data, which are collected and documented whenever the clinics are notified of a new event. The former do not lend themselves well to time-to-event-based analyses, but can be used as outcome variables at specific time points (the time of the Year 2 visit, the time of the Year 3 visit, etc.). The latter are documented using hospital records so that an accurate time-to-event variable can be calculated, and Cox proportional hazard modeling employed in analysis.

For periodic questionnaire-based algorithms, there are yearly variables for each year in which the necessary information was collected. These include YxDEPR1, YxDEPR4, YxADAEPI, YxADA2H, YxHBP1, YxHBP2, YxSHBP, and Y6METSYN (and the component variables Y6METSAB, Y6METSBP, Y6METSGL, Y6METSHD, and Y6METSTG). If there is no variable for a particular year, that is because one or more components needed for the calculation is missing (see also Strengths and Weaknesses, below). For this final release, 2016/01/01, questionnaire-based data are included through the Year 17 Quarter 3 visit.

For events-based algorithms, there is a single variable for each type of event, generally classified as never, baseline prevalent, incident, or recurrent, depending on whether an event occurred during Health ABC and whether the participant was classified as already having disease at baseline. Each events-based incident disease variable also has an associated days-to-event variable as well as an event-date variable (see also below, under Strengths and Weaknesses). Thus, the incident cancer (any type except non-melanoma skin cancer) algorithm includes the indicator variable CANANYi, the associated days-to-event variable CANANYds, and an associated event-date variable CANANYdt. In addition, these events-based algorithms generally have an adjudication variable (e.g., CANANYad) that classifies an incident event as either definite or possible, depending on how it was classified by the Health ABC Events Adjudicator.

The original prevalent disease algorithms were specified by the prime movers listed in the documentation and were reviewed extensively by the Diagnosis and Disease Ascertainment Committee. Since then, a group of investigators interested in diabetes has repeatedly worked on creating incident diabetes algorithms and revising the prevalent diabetes algorithms. For other diseases, the original prime mover, for the most part, took on expanding on the prevalent disease algorithms to create new incident disease algorithms with the same basic strategy. In some cases, due to decisions made throughout Health ABC about which measurements would be done each year, it was necessary to create new prevalent disease algorithms that would be comparable to the incident disease algorithms that could be constructed. This was the case, for example, for risk for depression based on CES-D. While creating the incident risk for depression algorithm, which uses the short (10-item) version of CES-D, a new prevalent risk for depression algorithm (Y1PDEPR4) was created that is analogous to Y1PDEPR3, but uses the short form of the CES-D.

Prevalent and incident disease algorithms are documented in the following ways:

- A calculated variable worksheet giving a written description of how the variables were coded, including how missing values were treated. These worksheets may include medication variables from YxRxCalc that are used in the definition of prevalent or incident disease. These variables are in parentheses to indicate that documentation for them can be found with the documentation for YxRxCalc.
- A flow chart showing the decision tree used to determine presence or absence of disease at baseline, including variable names.
- SAS code (Appendices II and IV) showing the actual SAS code used (for those analysts wishing to look more deeply into exactly how a particular participant might end up coded as having or not having prevalent or incident disease.

Diseases are grouped by types of disease and prime mover for compactness of documentation. See Grouping columns in Appendices I and III. These groupings are linked to the corresponding calculated variable worksheets for convenience in navigating the documentation. In addition, the individual variable names in the listing and the calculated variable worksheets are linked to the corresponding flowcharts.

Rfc dmjjmugle Q?Q t_pg_`jcq ucpc _bbcb rm rfc dgl_j npct_jclacZglagbclac bgqc_qc b_r_qcr &Clb md Qrsbw* Pcjc_qc 0./4-./-./', Rfcpc ucpc lmr _lw af_lecq rm rfc ambgle md rfcqc t_pg_`jcq dpm k rfc k mqr pcaclr npgmp pcjc_qc &0./.-/.-./', Njc_qc qcc rfc djmuaf_pr _lb jmega qcargmlq* _lb rfc Q?Q npmep_k kgle ambc dmp rfcqc t_pg_`jcq gl rfc ?nnclbgacq,

Lcu Glagbclr Q?Q T_pg_`jcq dmp Dgl_j Pcjc_qc &0./4+./+./' md Npct_jclr - Glagbclr Bgqc_qc B_r_qcr

Q?Q L_kc	Q?Q J_`cj
W/3O/?B?0F	W/3O/8EJSAMQC QR?RSQ&?B?CNG)MERR bcd md W/ BK'
W/3O1?B?0F	W/3O18EJSAMQC QR?RSQ&?B?CNG)MERR bcd md W/ BK'
W/4O/?B?0F	W/4O/8EJSAMQC QR?RSQ&?B?CNG)MERR bcd md W/ BK'
W/4O1?B?0F	W/4O18EJSAMQC QR?RSQ&?B?CNG)MERR bcd md W/ BK'
W/5O/?B?0F	W/5O/8EJSAMQC QR?RSQ&?B?CNG)MERR bcd md W/ BK'
W/5O3?B?0F	W/5O18EJSAMQC QR?RSQ&?B?CNG)MERR bcd md W/ BK'
W/3O/?B?CNG	W/3O/8EJSAMQC QR?RSQ&?B?CNG bcd md W/ BK'
W/3O1?B?CNG	W/3O18EJSAMQC QR?RSQ&?B?CNG bcd md W/ BK'
W/4O/?B?CNG	W/4O/8EJSAMQC QR?RSQ&?B?CNG bcd md W/ BK'
W/4O1?B?CNG	W/4O18EJSAMQC QR?RSQ&?B?CNG bcd md W/ BK'
W/5O/?B?CNG	W/5O/8EJSAMQC QR?RSQ&?B?CNG bcd md W/ BK'
W/5O1?B?CNG	W/5O18EJSAMQC QR?RSQ&?B?CNG bcd md W/ BK'
W/3O/F@N/	W/3O/ Glagb FRL pcnmprcb
W/3O1F@N/	W/3O1 Glagb FRL pcnmprcb
W/4O/F@N/	W/4O/ Glagb FRL pcnmprcb
W/4O1F@N/	W/4O1 Glagb FRL pcnmprcb
W/5O/F@N/	W/5O/ Glagb FRL pcnmprcb
W/5O1F@N/	W/5O1 Glagb FRL pcnmprcb
W/4F@N2	W/4 Glagb FRL &Nfwqgmjmega_j'
W/4QF@N	W/4 Glagb gqmj_rcb Qwqrmjga FRL

2. Cross reference of dataset names with exact source

A complete list of variable names can be found in the Proc Contents (Appendix V, search under PrevIncDz) and in Appendices I and III.

3. Dataset structure and contents

The PrevIncDz file contains a single observation per participant. There are 3075 observations in the PrevIncDz file.

Key variables:

HABCID	HABC Enrollment ID without the 2-letter prefix

4. Condition of data

a. Known data errors:
None at this time

b. Strength and weaknesses of dataset items:

Prevalent disease variables: Most of the prevalent disease algorithms are based on self-report of disease at baseline from questions in the Baseline Questionnaire. Some of the algorithms, such as Y1ADAEPI, Y1ADA2H, Y1PFTCAT, Y1PHBP2, Y1OSTBMD, are Y1PCHD2 are based on more objective measurements at baseline (glucose tolerance, PFT, blood pressure, BMD, EKG, etc.). And finally, there are some variables coding for risk of certain conditions, such as depression or benign prostatic hypertrophy, based on standardized questionnaires.

See also Year 1 Calculated Variable Dataset for other calculated variables, such as ankle-arm index, abnormal heartrate, and Teng Mini-Mental score that are related to baseline disease status.

As additional visit years have been added to the original diabetes algorithms,the number of participants whose diabetic status is unknown increases with each year of the HealthABC study. This increase is partly explained by participants missing a visit or dropping out of the study, but it is also due to the fact that fasting glucose is not collected for certain years or not available for some participants in study years where glucose tests were performed. To reduce the number of missing the following major changes were implemented in release 2010/10/01 for prevalent and incident diabetes variables:

1. Where a participant was not fasting but their glucose is within normal levels they are classified as "Not Impaired".
2. Where glucose results are not available or no medication history was collected, participants who self reported as not having diabetes are put into a new category, "Self report not diabetic - no fasting glucose".

The details of these changes for prevalent diabetes are specified below:

> Y1PDIAB1: Note: This variable assigns diabetes status based on self-report and diabetes medication use without taking into account any glucose assays. If participants report that they do not have diabetes and there is no record of their taking diabetic medications, they are now coded as "Not Diabetic."
>
> Y1ADAEPI required a fasting glucose to assign participants into the category of "Not impaired". However, a normal glucose result when NOT fasting is now also considered acceptable.
>
> Y1ADAEPI has a new category of "Self-reported not diabetic, no fasting glucose" for participants coded as not diabetic in Y1PDIAB1 (based on self-report or med use) and whose glucose levels were not measured or have a <u>non-fasting</u> glucose greater than or equal to 100.

Y1ADA2H has also been updated to allow participants with a normal <u>non-fasting</u> two-hour glucose level to be assigned to "Not Impaired." The new "Self -reported not diabetic, no fasting glucose" classification has also been added to this variable.

In Release 2010/10/01 there are two new prevalent metabolic syndrome component variables, Y1METS8TG and Y1METS8GL. A positive indicator in either of these new components requires that a participant must have fasted for at least eight hours. For example, a Yes value indicating elevated glucose for Y1METS8GL is based on a fasting glucose test. A non-fasting value is accepted if it is in the normal range (i.e., Y1METS8GL=0 even if FAST8GLU1=.T if GLUCOSE1<110). These new variables are set to missing when the participant was not fasting and the test result was elevated, since it is impossible to know whether the value would be elevated if the participant had been fasting. The original Y1METSGL was set to Yes if there was an elevated value whether or not a participant was fasting. Also, please note that the corresponding new summary variables, Y1METS8YN and Y1METS8NO, are calculated using Y1METS8TG and Y1METS8GL, along with the original variables Y1METSBP, Y1METSAB, Y1METSHD, which are not seriously affected when the participant was not fasting. Please see algorithm flow charts for details.

Previous to release 2010/10/01, the algorithm for prevalent hypertension (Y1PHBP1) made use of the medication calculated variable Y1HBPDRG in two ways:

1) To confirm a self-reported diagnosis of hypertension, resulting in Y1PHBP1=1 (confirmed hypertension) or Y1PHBP1=2 (possible hypertension--if not using a hypertensive drug)
2) To screen for possible people who deny hypertension on self-report, but are taking a hypertensive med (resulting in Y1PHBP=3, treated but not reported). In this context, MIFREAS, the reason for medication use, was used to narrow down the criterion to those actually taking these medications for hypertension, although they are commonly used for other problems such as heart disease.

The first use we consider justified because the participant reports that they have hypertension, so if they're taking a medication that is often used for hypertension, it's probably confirmation that they really do have it. However, we are no longer comfortable with the second use, which suggests that they might be hypertensive just because they take a drug that is sometimes used for hypertension. Furthermore, once categorized as "treated but not reported" those participants were previously coded in all subsequent study years as having a history of hypertension. To test this concern , we dropped the "reported but treated" category and compared the results to how participants were classified before. We found that the majority of participants coded as "HTN treated but not reported" ended up as "No Prevalent HTN Reported" in following years (99 ppts out of 103 in year 1 with Y1PHBP1=3). Consequentially in Release 2010/10/01, we have stopped using MIFREAS and have eliminated the "HTN treated but not reported" category. This brings the Year 1-6 variables in line with later years, where the "treated but not reported" category had already been dropped because MIFREAS was no longer collected. Please see the hypertensive algorithm flow charts for details.

The remainder of this section is left in place to document the changes that have been made over time to some of the prevalent disease algorithms that have been modified since the initial release of the dataset.

The prevalent diabetes algorithms based on fasting glucose and/or OGTT assays were previously updated to keep them in line with the current, more strict ADA criterion for fasting glucose (≤100 mg/mL) see Appendix I).

The prevalent depression algorithm using CES-D scores (Y1PDEPR2) was previously changed from a cutoff of 15 to the correct cutoff of 16 (Radloff 1977 App Psych Meas 1(3)385-401), and those with exactly the cutoff value of 15 correctly included in the "at risk" group for both Y1PDEPR2 and Y1PDEPR3. In addition, the algorithm for determining whether a participant was taking an antidepressant has been modified to take reason for use into account. Since many of these drugs are prescribed for reasons other than depression, the reason for use is now taken into account in determining whether to set this indicator variable to 1 (taking antidepressant) through Year 6. After Year 6, MIFREAS is not collected and without it YxDEPDRG can not be created. Without YxDEPDRG, YxDEPR1 can also not be created and therefore is not available after Year 6. And finally, in order to have a standardized variable that remained consistent from year to year for determining risk for depression, a 10-point scale (YxCES_D10, also included in this dataset) based on the short form of the CES-D was calculated for all years (except Year 2, where the GDS was used instead of CES-D, and Years 7 and 9, where no depression scale questions were included.) The prevalent disease variable Y1PDEPR4 and incident disease variables YxDEPR4 are based on this scale.

After comparing the number of participants with a history of cardiovascular disease or cancer using the HCFA data from HPrevDis and OPrevDis to the number found by self-report at the baseline visit, the prime movers decided to use both as criteria to determine baseline prevalent disease. This ensures that cases classified as incident during Health ABC are more likely to be true new cases. New prevalent disease variables (Y1PCHD3, Y1PCANANY, Y1PCANBRST, Y1PCANCOLN, Y1PCANLUNG, Y1PCANPRS) were created and are included in the dataset. In addition, since non-melanoma skin cancer is not considered an event in Health ABC, these cancers were omitted from the new prevalent disease variable Y1PCANANY. A baseline prevalent cardiovascular disease variable (Y1PCVD) was also added, which includes prevalent coronary heart disease or prevalent cerebrovascular disease, by HCFA or self-report.

Thus it is very important for the analyst to study the flow charts and calculated variable worksheets (Appendix I), and possibly even the SAS code (Appendix II) to be sure they understand what each variable represents, what constitutes an appropriate control group for their purpose, etc.

Periodic-questionnaire-based incident disease variables

With release 2010/10/01, incident diabetes indicators include changes similar to those made in the prevalent diabetes indicators. Non-fasting glucose results within the normal range are assigned to "Not impaired" in YxADAEPI and YxADA2H variables and these variables include a "Self-reported not diabetic, no fasting glucose" category. In years where fasting glucose tests were conducted, YxADAEPI and YxADA2H have seven options (recall that the only difference between these two sets of variables is the definition of prevalent diabetes at baseline):

0: Not impaired
0.5: Self-Reported not diabetic, no fasting glucose [NEW]
1: Impaired fasting glucose
2: Diabetic by fasting glucose
3: Newly diagnosed diabetes
4: Previously diagnosed diabetes
Missing

In years when fasting glucose tests were not done, YxADAEPI and YxADA2H categories are limited to:

0.5: Self- reported not diabetic, no fasting glucose
3: Newly diagnosed diabetes
4: Previously diagnosed diabetes
Missing

Please note that the former category "0: self-reported not diabetic" has been changed to "0.5: Self-reported not diabetic, no fasting glucose," so visit years without fasting glucose are consistent with visit years when glucose tests were performed.

With Release 2010/10/01 there are two new incident metabolic syndrome component variables for year 6, Y6METS8TG and Y6METS8GL. As explained above for the new prevalent variables, a positive indicator in these new components requires that a participant must have fasted for at least eight hours, although non-fasting results are accepted if they are in the normal range. Also, corresponding summary variables Y1METS8YN and Y1METS8NO were added.

As is the case with prevalent hypertensive variable Y1PHBP1 (see explanation above in prevalence section) with release 2010/10/01, we have stopped using MIF Reason and have eliminated the "HTN treated but not reported" category from incident hypertensive variables YxHBP1. Again, please see the hypertensive algorithm flow charts for details.

Many of the incident disease variables are based upon periodic questionnaires or measurements administered at annual contacts. Due to limited time for each visit, however, the content of the visits was not identical from year to year. In addition, some participants had a phone, proxy, or home visit in lieu of a clinic visit in some years, so data necessary to determine their incident disease status may be missing. Fortunately, most of the questions or measurements needed for most of the algorithms were considered important enough to be asked/administered at every type of contact and in every year. There are some exceptions, however.

For example, the GDS depression scale, rather than the CES-D depression scale was used in year 2, so there is no Y2DEPR4. It was also not considered appropriate to ask the depression questions by phone or through a proxy, so these types of visits may lead to missing values (in the end, a number of phone contacts did ask the battery of CES-D questions). And, finally, the version of the CES-D used alternated between the short form (10 items) and the long form (20 items). For best comparisons across years, a 10-item score was calculated for all years with CES-D and used in the YxDEPR4 algorithm.

Similarly, medications were not collected in years 4, 7, or 9, so there is no YxDEPR1 for those years. In other cases the variable exists, but not all possible values exist, for example, Y4HBP1 can only take on values of 0, 2, or 4 because medications were not collected. Similarly, Y3ADAEPI, Y5ADAEPI, Y7ADAEPI, Y8ADAEPI and Y9ADAEPI can only take on values of 0.5, 3 or 4 because fasting glucose was not measured in those years.

Thus it is very important for the analyst to study the flow charts and calculated variable worksheets (Appendix III), and possibly even the SAS code (Appendix IV) to be sure they understand what each variable represents, what constitutes an appropriate control group for their purpose, etc. To avoid throwing away useful information, participants missing baseline information were not coded as missing throughout the follow-up period. Therefore, in some cases, to get the cleanest possible control group for some analyses, it is best to use the group coded as 0 (no incident disease, no prevalent disease) and remove the subset whose correspondent prevalent disease variable is missing.

Event-based incident disease variables

The cancer and cardiovascular disease algorithms are based on Health ABC events data. Although participants are asked at every contact (i.e., approximately every 6 months) whether any of the events of interest to Health ABC have occurred since the last contact, the data eventually used to adjudicate the event consist of hospital records and other dated materials, so it is possible to nail down exactly when an event occurred. When a single participant has experienced multiple events of a particular type, the days-to-event variable (XXXds) is set to correspond to the earliest definite event, or the first possible event if no definite events of that type occurred. The cancer variables and the cardiovascular variables treat possible events differently because of the differing nature of these two types of disease. A possible cancer followed later by a definite cancer of the same type is presumed to confirm the initial diagnosis. Thus, the cancer algorithms treat the first event of the appropriate type, whether confirmed then or later, as a definite event (CANXXXad=1) occurring on the earlier date (CANXXXds=(earlier date-CV1DATE)). Only if a possible cancer is never later confirmed is it coded as possible (CANXXXad=2), again with the earliest date.

For cardiovascular disease, however, since a later confirmed event does not necessarily imply that an earlier unconfirmed event occurred, confirmed events are used in preference to unconfirmed events of the same type, but the days to event variable is based on the definite event. So, for example, a possible MI on 9/1/99 followed by a confirmed MI on 12/30/00 is coded as CHDMI=2, CHDMIad=1, CHDMIdt=12/30/00, and CHDMIds=12/30/00-CV1DATE. If only an unconfirmed event has occurred, this is used, but coded as possible (e.g. CHDMI=2, CHDMIad=2), and the days-to-event variable (e.g. CHDMIds) is based on the date of the possible event (MI).

Thus it is very important for the analyst to study the flow charts and calculated variable worksheets (Appendix III), and possibly even the SAS code (Appendix IV) to be sure they understand what the variable they are using represents, what constitutes an appropriate control group for their purpose, etc.

5. Missing values

Care has been taken to try to avoid misleading the analyst when key data are missing, while at the same time not discarding useful information. The prevalent disease algorithms have now undergone three rounds of careful consideration of the consequences of missing information. A participant is not coded as being disease-free at baseline if the information necessary to determine this is missing. Instead, they are coded as missing (SAS special missing value code .M). This allows the investigator to determine a clean control group for case-control analyses. If the analyst prefers to consider these participants as not having disease, then they should look for VAR=0 or VAR=.M (or VAR<=0) where VAR represents the prevalent disease indicator variable in question).

For example, Y1PCHF uses the medication use variables Y1CHFDIU and Y1CHFVAS or Y1CARGLY to confirm a self-reported physician's diagnosis (MHHCCHF) of CHF. If medication information is missing, a positive self-report is treated as possible (Y1PCHF=2), rather than definite (Y1PCHF=1). However, lack of medication information does not affect a determination that the participant was free of CHF at baseline, since the medication use variables alone are not sufficient for a presumed diagnosis of CHF. Missing, or indeterminate (don't know, refused) self-report, on the other hand, does result in the special missing value code .M for Y1PCHF.

The flow charts and calculated variable worksheets attempt to clarify exactly how various missing information is treated in the determination of each of the prevalent disease algorithms. SAS code is also provided (Appendix II) for any analysts wanting to probe further into these special cases.

Missing data obviously also affect the incident disease algorithms. Since most of the incident disease algorithms created thus far are for diseases that persist once they occur, the following conventions have been adopted:

For periodic questionnaire-based algorithms, previous history of disease is coded as history, even if there is no current data for that person (e.g., participant is deceased or missed the visit). Barring a history of disease, however, key missing information in a particular year results in the incident disease variable for that year being set to missing. For example, if there is no medication information or information about recent diagnosis of hypertension, YxHBP1 is set to missing (.M), but previously diagnosed hypertension still results in a value of YxHBP1=4.

For events-based algorithms, events not yet reported or adjudicated are assumed not to have occurred. Careful attention needs to be paid by the analyst to the question of appropriate censoring. Adjudication is complete for all events reported through 8/14/2012.

Events reported later than this date are included, but the adjudication for events reported after this date was not completed prior to the end of the Health ABC study. **Therefore, for many analyses, it would be cleanest to censor the events to include only those with event dates no later than 8/14/2012.** A date of last contact (DTLASTCT), which consists of the date of death, the date of the last completed contact (missed visits not included), or the discharge date for the most recent entered event, whichever comes last, is also included for each participant. The date of the event upon which the incident disease variable is based is also included (XXXdt). For greatest flexibility of analysis without misleading anyone, the days-to-event variables (XXXds) in this dataset are set to missing (SAS special missing value code .A) if no event has occurred during Health ABC. Thus before a proportional hazards analysis model can be run, the analyst must decide:

- Should only events through 8/14/2012 be considered for the analysis? If so, any incident disease variable with an event date (XXXdt) later than the cutoff date should be set to either 0:No event or 1:Prev dz, no recurrence depending on baseline disease status, and the associated days-to-event variable (XXXds) should be sent to missing. Then the earlier of (cutoff date-CV1DATE) or (DTLASTCT-CV1DATE) should be used for all missing days-to-event.

- Will all events be used in the analysis (with the caveat that there will be bias introduced by incomplete adjudication)? If so, all missing days-to-event can be replaced with DTLASTCT-CV1DATE.

An occurrence of disease during Health ABC could be either recurrent or incident, depending on whether the participant had a history of the disease at baseline. All incident disease algorithms make an effort to separate these two categories. When prevalent disease data are missing, however, this becomes more difficult. Missing prevalent disease data are ignored in the incident disease algorithms, so that a participant with missing prevalent data may be classified as either incident or none on the basis of information obtained during Health ABC. Since the prevalent disease variables are included in the PrevIncDz dataset, if the analyst wishes to be sure that all participants in their analysis are either completely free of disease or truly incident during Health ABC, they may wish to exclude participants whose corresponding prevalent variable is coded as missing. Appendix III lists the corresponding prevalent disease variable used for each incident disease variable.

Special Missing Value Codes

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

. ='Missing Form'
.A ='A:Not Applicable'
.M ='M:Missing'
.U = 'U:Unacceptable'

Description

. : Missing Form
Used when a value is missing because the entire form has not been entered (e.g., participant missed that contact).

A: Not Applicable
Used when a parent variable value indicates that this variable does not apply. For example women can't have prostate disease, so their values for the prevalent and incident prostate disease variables is set to .A. Similarly, if an event has never occurred during Health ABC for a particular participant, the event-based incident disease variable is 0 (e.g. CANANYi=0), and the corresponding days-to-event (CANANYds), date of event (CANANYdt), and adjudication variables (CANANYad) are all set to .A.

M:Missing
Used to flag missing values when there should be a value, but missing data preclude the possibility of calculating one. For example when medication-use data are missing for a participant, and the participant does not have a history of depression, YxDEPR1 is set to .M.

U:Unacceptable
Used when Reading Center data exist but have been reviewed during QC as unacceptable. The prevalent disease variable PFTCAT was set to .U when the QC score for FEV1 or FVC was 0 or 1.

General Strategies for Using Special Missing Values

In SAS, when using special missing values in logical expressions, the missing value is no longer only equal to '.' To express a value equal to missing, the code should be written: <= .Z or alternately: le .Z

To express a value not equal to missing, the code should be written >.Z or alternately: gt .Z
.Z is the greatest value of missing available in SAS. All negative numbers are larger than .Z. So <0 includes all negative numbers and missing values, while <= .Z includes only missing values.

6. Dataset index formulation and key variable mapping

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

7. 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.