The CPP ‘Variable.asc’ electronic data file

Novice data managers and researchers new to cpp data or pressed for time are strongly encouraged to begin their work with the variable file: 59,000 records, 1225 variables, .

The variable file (varfile) is/was

1. A **selection of 1200 variables** drawn from or created from the total 6,700 in the 30 work files and master data file (6.1 million punchcard records) which was

2. structured with **1 record per child id or pregnancy id** and was NOT in 80-column punchcard record format and contained

3. data from **many areas** of substantive interest but

4. no serological test result data.

For novices, the varfile is the easiest to use because it contains information across the life-course and from many substantive areas of interest. Information on both the mother and child are in one record. Novice data managers do not have to address linking the ids (known as match-merging in sas, adding variables in spss) with the problems of pairs where either mother or child is absent, individuals wrongly id’ed, or the duplicate record problems brought by multiple births (not a unique key).

The varfile has the **clearest, most easily understood original documentation** (use **varfile 1972 docs + variable namenumbers.pdf**) but NOT the microfiche documentation (not **III Variable file Part A Index of data items fiche 41-44.pdf,III, Variable file Part B Data item creation method fiche 45-51.pdf**). Using the varfile 1972 documentation requires less computing knowledge and searching than using the ‘fiche documentation.

While its ostensible purpose was to aid the research ends of the CPP in an ongoing, timely manner, it also supplied information to various external watchdog committees which were not always favorable to the project’s continuation. Thus, there are **non-research influences on the variable selection and electronic file structure.** For our purposes, note that the varfile

5. variables are not in life-course chronological order or clearly grouped in scientifically coherent sections.

Since there are so many variables (1225 or so - the exact number depending chiefly on whether dates are treated as mm, dd, yy and/or mm/dd/yy), from diverse work data files and the master data file, with created as well as punched variables, the sections of the CPP documentation about the variable file on the microfiche are particularly large and cumbersome to search, even in PDF format, even if definitive. ( )

For ease of use by most readers, and particularly inexperienced analysts (JHU Pediatric Fellows) or even experienced researchers unfamiliar with the CPP data, will find the

6. enhanced working documentation from 1972 (**varfile 1972 docs + variable namenumbers.pdf**) easiest to use and sufficient for many. Enhanced by adding variable name-numbers (V001-V1179)

to the item description. However, the 1972 documentation

7. does NOT follow the usual modern convention of listing/describing the items in the order in which they appear in the electronic ASCII or other dataset (**NOT column order**) i.e.

V100 and V101, adjacent in the documentation, are NOT necessarily adjacent in the dataset as the SAS input program (varfilenumv.sas) follows the 1972 documentation order which is NOT the usual column order in the ASCII format VARFILE.ASC. as shown by the SAS output text file varfile.con[tents].

Also, the SOURCE of the varfile variable is USUALLY, but not always, obvious in the 1972 documentation, especially when there are explicit variable labels and unambiguous value labels. Again, there is

8. **NOT ALWAYS clear information on VARIABLE SOURCE** or how it was created, if not punched from a card in the 1972 documentation. In these cases, the analyst MUST refer to the later definitive microfiche documentation.**(III Variable file Part A Index of data items fiche 41-44.pdf, III, Variable file Part B Data item creation method fiche 45-51.pdf).**

The position of the NINDB is the first 9 columns, NOT the usual punchcard structure of cardnumber/version 1-5, NINDB 6-14 and note well that

9. There are NINDB numbers where the last digit is blank i.e. there is no child/plurality information and the NINDB is 12345678_ not 123456789.

The NINDB is 8 columns (1-8) in the VARFILE and a variable called 'plurality' is column 9.

Substantively, these NINDB8/No Plurality records may be losses to follow-up in the sense of the mother never-came back but a baby was born or the mother never came back because the pregnancy ended before full-term, but the variable(s) known as the 'cohort switch' explicitly addresses this issue.

These NINDB numbers may be analytically ignored as representing mothers lost to follow ups, or they may be included. If included, the last (9th) digit may be set to 9 or, as is the author's preference, to 8, which does not apply anywhere else in the CPP dataset(s). If this missing plurality is not explicitly treated most software packages will not read the indented number as a nine digit number, important if using the NINDB as a means of subsetting by site with a range of NINDB numbers. For example , all 'JHU numbers' begin with 37 so can range from 370000000 to 379999999. SAS, by default, reads 37345678_ as 37,345,678 which is below 370,000,000. Of course, a site variable could also be used.

The original VARFILE.ASC electronic file transcribed from IBM reel-reel tape by NARA was over 1600 columns wide, but it only contained 1491 columns of data. The surplus white space has been elided.

A broad contents page is given in variable file contents.pdf but again, the user is cautioned that it is not definitive and if the source of the variable or its content is not clear then the user MUST refer to the microfiche documentation (III Variable file Part A Index of data items fiche 41-44.pdf, III, Variable file Part B Data item creation method fiche 45-51.pdf).