Brief Overview

Data was collected on 50,000 pregnant women, their pregnancies, deliveries and all of their subsequent 58,000 children enrolled between 1959 and 1965 at 12 sites in the U.S. The children were then followed for up to eight years.

The most important data collection points were: the pregnancy/before birth; the neonatal 'nursery' period; a physical and neurologic examination at 4 months; a psychological exam focusing on development at 8 months, and another at 12 months; formal assessment of speech, language and hearing at 3 years; a psychological exam and testing at 4 and 7 years; and speech, language, and hearing at 8 years. Data was collected by well-trained; lay interviewers, professionals, or clinicians on clinical interview forms, direct observational assessments, and test result schedules.

The primary focus of interest in the NCPP was physical and biological: the pregnancy, birth, the child, and that child's physical and neurodevelopment, but there is also considerable information collected from the mother about herself, father of the child, family structure and household compositions and child's older siblings born before the study began, and sociodemographic information when the child was 7 and 8 years old.

The mode of data collection was filling out forms for each mother and/or child at each examination. The data from every form, or in some cases multiple forms, were then transferred to punch cards. **Electronic datasets were stored on 32 IBM reel-to-reel magnetic tapes mostly in records corresponding to one punchcard of over 140 possible 80-column punchcards per child. This data was also stored on 75 microfiche. The data was transferred to ASCII files using _____??**

Current State of Data

There are currently 63 ASCII files ranging in size from 1-400 Megabytes, totaling 631Mb. Most ASCII files correspond to a form or group of forms pertaining to a measurement event, such as

an obstetric visit (ob form), a record of placental pathology (path form), or a psychological test administered at age 7 (ps form). Because each punch card contained only 80 columns, in many cases, there were multiple cards per form. As there was one ASCII file per form, there were also multiple cards per ASCII file. Our goal was to have one observation per person that included all forms and all cards, i.e. all data sets for each individual, in a rectangular dataset. A mother and her child(ren) were considered separate individuals. Please see Appendix A for the list of forms and corresponding number of cards.

Programming Information

All programming was done using SAS, version 8.0. One dataset for each group of forms (i.e. card) was created. In all, there were 5 datasets, including pathology, pediatric examinations, psychological tests, and socioeconomic variables of the family, **and obstetrics**. Within each dataset program, multiple forms and cards per form were read in as separate mini-datasets then merged at the end of the program.

For example, Pathology dataset with 3 forms: path1, path2, and path3 consisted of the following mini-datasets:

Path1, card1
Path1, card2
Path2, card1
Path3, card1
Path3, card2
Path3, card4

Although the programs were written as above, if requests are made for only certain parts of data (i.e. only Path1, card2 and Path3, card1 above), the programmer can easily cut and paste the dialogue *of each card* , to match the researchers requests then merge them together (see below for explanation of how cards and forms were merged).

How Data Was Read In

Input Style

Column Input was used for all programs. This form of input (explain)……..

Conditional Input

Each card had a maximum of 80 columns. Columns 1-5 were for the card and revision number, and columns 6-14 of each card were to identify a single participant, and therefore were the same on each card corresponding to the same participant (see below for explanation of NINDB number). Columns 15 to the end of the card concerned questions on the corresponding form. Thus, each card per form had different questions and answers after column 14, and had to be read in separately using a conditional input.
For example, from path program, cards 1 and 2:

```
data path1_card1; *Card 1 of 2 path1 cards;
infile 'c:\ncpp\ascii\path1' lrecl=81;
* uses column input;
input cardnumb 1 @; * @ allows the conditional input;
if cardnumb=1;
input /* condition*/
nindb          6-14

data path1_card2; *card 2 of 2 path1 cards;
infile 'c:\ncpp\ascii\path1' lrecl=80;
* uses column input;
input cardnumb 1 @; * @ allows the conditional input;
if cardnumb=2;
input /* condition*/
nindb          6-14
```

The main difference between these two programs is that the first reads in the data only if the card number is one, and the second reads in the data only if the card umber is two.

Card, Form and Revision Numbers

Column 1 was for the card number, 2-4 the form number, and 5 the revision number. The form number remained the same for all cards of a single form, with only the card number changing. For example, Form 240 with 5 cards might have column numbers of 1240, 2240, 3240, 4240,

and 5240. Revision numbers may not be the same for all cards per form, or within the same card.

Unique Identifier (NINDB number)

Each participant in the study had his/her own eight-digit unique identifier number. This was located at columns 6-14 in the ASCII texts. The first fourteen columns correspond to the following:

			NINDB unique identifier number			
					Person	
	Card Number	Revision Number	Site Number	Family Number	Pregnancy	Plurality
Column number	1,2,3,4	5	6,7	8,9,10,11,12	13	14
Fictitious number	1201	3	05	54320	2	2

The NINDB unique identifier number consists of the site number, family number, plurality, and pregnancy:

The *site number* refers to one of the 12 sites.

The *family number* is unique for each family but is the same for each person of the family. Thus, a mother and all her child(ren) enrolled in the study will have the same family number.

Pregnancy refers to the number (order) of the child enrolled by the mother into this study. For example, if the mother is in her first pregnancy and enrolls in the study, the pregnancy will be one. If the mother had a child previously, but was not in the study at the time, then becomes pregnant again and enrolls in this study, pregnancy will still be 1 because it is the mother's first child enrolled in the study. If the mother is pregnant twice and enrolls in the study for both pregnancies, the second child will have a pregnancy of 2.

Plurality refers to whether the data corresponds to a child or mother. This is also the only number that distinguishes between twins, triplets, and quadruplets. Plurality will always be 9 for data corresponding to a mother and 0 ***(or 1 in some cases????)*** for a single birth. Plurality will either be 1 or 2 for twins, 1,2, or 3 for triplets, and 1, 2, 3, or 4 for quadruplets.

Together the pregnancy and plurality correspond to the individual *person* the NINDB and data represent.

Field Names

In order to standardize the data for public use, the names of the variables correspond to the field and form numbers, and therefore are not intuitive in terms of data analysis. The data is labeled by field and card number. For example, field 27 of card 1202 would be f27_1202. However, some fields have multiple parts. In these cases, each part is consecutively labeled a, b, and so on after the field number, in the order given in the codebook (ex. f27a1202). For example, field 8 of form PS20, card 1120 concerns clinical impression at the 4 year psychological examination, and has 5 parts to it: Intelligence (column 19), fine motor development (col. 20), gross motor development (col. 21), concept formation (col. 22), and behavioral (col. 23). This example would be written in the SAS program as follows, with the corresponding field name and column number:

F08a1120	19
F08b1120	20
F08c1120	21
F08d1120	22
F08e1120	23

Date Fields

All date fields were split up into separate variables for the month, day (sometimes week) and year. For example, if for field 5, card 3201, the following numbers for a date of examination in the ASCII text file were 112168, then the month (11) would correspond to f05a3201, the day (21) to f05b3201, and the year (1968) to f05c3201.

Multiple Measurements/Multiple Cards

In some cases, there were multiple measurements resulting in additional cards for some of the participants. An example is the PED8 form, card 1408. This form is the newborn diagnostic summary, which includes abnormalities, diagnoses, and procedures. The diagnoses and procedures are recorded at the end of the form. If there were more then 10-18 diagnoses, card 2 was required, which was coded the exact same as card 1 except column 1 is 2, and columns 57-

80 (the numbers corresponding to diagnoses/procedures) are different. If 19-27 diagnoses were reported, card 3 was required and the same procedure as above was incorporated. This was repeated up to 46-54 diagnoses, for which a card 6 was required.

Merging

As mentioned above, for each dataset (ex. PEDS, PATH), each card per form was read in separately, forming small datasets for each card. In order to combine the small datasets into one large dataset relating to pediatrics, psychological exams, etc., we had to merge the individual cards, for that particular dataset, at the end of the SAS program. Although we merged the data via this method, because all the programs are already written, one could merge different datasets then those given.

First, the basics of merging using SAS will be _____, followed by an example we used, and an example merging various small datasets from different forms/different topics into one dataset.

```
data 'c:\ncpp\sas\pedforms_allobs';
merge ped01_card1 ped01_card2 ped01_card3
      ped02_card1 ped02_card2 ped02_card3 ped02_card4 ped02_card5
      ped03_card1 ped03_card2
      ped05_card0
      ped06_card1 ped06_card2 ped06_card3 ped06_card4
      ped07_card0
      ped08_card1 ped08_card2 ped08_card3 ped08_card4 ped08_card5
      ped10_card1 ped10_card2 ped10_card3
      ped11_card1 ped11_card2 ped11_card3
      ped12_card1 ped12_card2 ped12_card3 ped12_card4
      ped14_card1
      ped75_card0
      ped76_card1 ped76_card2 ped76_card3 ped76_card4;
by nindb;
run;
```

2. Problem with Peds dataset: multiple observations per card.

    a. How this was solved.

One problem we ran across in PED 14 was that there were multiple observations per person in the only card corresponding with PED 14. As a result, a person was listed up to seven times in seven rows of the dataset, whereas we want one row, i.e. one observation per person. The reason there were more than one observation per person in that PED-14 corresponded to physical growth measurements at each exam. Because participants often had exams at different times, exam number did not correspond with card number. Rather, all exams were written as having the same card number.