# Using the Public Data Archive from the Registry of Patients with Alpha$_1$-Antitrypsin Deficiency

Ralph O'Brien, Susan Sherer, Shannon Neeley, Gerald Roberson, William Grasser,
Aimee Wahle, Jana Shepherd, Alvin Van Orden, Ben Neibaur
Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation

This public archive of the Alphal-Antitrypsin Deficiency Registry Database was developed by the Cleveland Clinic Foundation's Department of Biostatistics and Epidemiology and is being distributed by the Registry's sponsor, the National Heart Lung and Blood Institute (NHLBI). We have strived to make the CD complete and self-documenting.

Creating this archive called for us to maximize its utility to the research community while protecting patient confidentiality ("de-identification") according to the October 1999 NHLBI guidelines, which are copied on the CD (Misc\NHLBIguidelines.pdf). Accordingly, each variable was examined to see if it could prove useful to researchers and if it posed a significant risk to patient confidentially. Written rules, summarized herein, have been used to change some variables in a consistent manner across all of the databases. All such decisions are documented in the file Misc\FieldComments.txt.

All code was checked by at least one other programmer.

Our goal was to make the structure of this archive as simple and as self-explanatory as possible. Accordingly, the individual databases correspond to the 17 case report forms (CRFs) that are appropriate for further analyses by the public. Thus the structure of each SAS dataset corresponds directly to a given CRF. All CRFs and their instructions have all been scanned to form PDF (Acrobat) files. They can be viewed and printed using Acrobat Reader, which is downloadable free (www.adobe.com) and comes in versions for all major computing systems. Notes have been made on the scanned CRFs to give the SAS variable names and to indicate which fields have been changed or dropped to protect patient confidentiality or to assure the research integrity of the data archive.

We also give three SAS programs that link to and merge datasets, perform data manipulation, and execute analyses.

## 1 Do this first

We advise users to copy the entire contents of this CD to their computer's harddrive or to some server's harddrive. Our Windows-based examples given below assume that everything was copied to a directory called C:\Alpha1CD and that the CD's directory structure was kept intact.

The CD was created on the Windows 2000 platform (ISO 9660 compliant).

## 2 Directories and files

The CD's contents are best described by tabling its file structure. You should check this with what you have copied over to you local machine and become familiar with it.

| Directory\File | Description |
|---|---|
| ReadMe.txt | Merely instructs users to read UsingDatabase.pdf |
| UsingDatabase.pdf | What you are reading now |
| RecodingRules.pdf | Rules used to adapt the database for public use |
| WindowsFormats\ | |
| CreateFormats.sas | SAS program to create formats library in Windows |
| formats.sas7bcat | Windows version of formats library |
| CaseReportForms\ | |
| form00.pdf | "Central Laboratory, Patient Information" |
| form01.pdf | "Screening" |
| form02a.pdf | "Initial Visit Form, Part A" |
| form02b.pdf | "Initial Visit Form, Part B" |
| form03.pdf | "Pulmonary Function Test Results" |
| form04.pdf | "Modified Dyspnea Index" |
| form05a.pdf | "Follow-up Visit Form, Part A" |
| form05b.pdf | "Follow-up Visit Form, Part B" |
| form06a.pdf | "Dropout/Death Notification" |
| form06b.pdf | "Final Death Notification" |
| form06c.pdf | "Release of Medical Information" [no dataset] |
| form07.pdf | "Cause of Death Review" |
| form08a.pdf | "Telephone Contact Follow-up" |
| form09.pdf | "PFT — Retrospective Data" |
| form10.pdf | "Augmentation Therapy Record" |
| form11.pdf | "Adverse Reaction Form" |
| form12.pdf | "Data Change Form" [no dataset] |
| form13.pdf | "Data Query Form" [no dataset] |
| form14.pdf | "Pulmonary Function Test Equipment" [no dataset] |
| form15.pdf | "Pulmonary Function Survey" [no dataset] |

| Directory\File | Description |
|---|---|
| form16.pdf | "Laboratory Results — Normal Ranges"[no dataset] |
| form17.pdf | 'Quality Control of PFT Data"[no dataset] |
| form18.pdf | "Authorization Form"(for paying centers) [no dataset] |
| form19.pdf | "Family Relationship Form"[no dataset] |
| form20.pdf | "Organ Transplantation Form" |
| form21.pdf | "Participant Survey Form"[no Public dataset] |
| ProcContents\ | Results of SAS PROC CONTENTS for each dataset. |
| form00_PublicContents.lst | "Central Laboratory, Patient Information" |
|  | "Screening" |
| **form01_PublicContents.lst** |  |
| form02a_PublicContents.lst | 'Initial Visit Form, Part A" |
| **form02b_PublicContents.lst** | 'Initial Visit Form, Part B" |
|  | 'Pulmonary Function Test Results" |
| **form03_PublicContents.lst** |  |
| form04_PublicContents.lst | "Modified Dyspnea Index" |
| form05a_PublicContents.lst | "Follow-up Visit Form, Part A" |
| **form05b_PublicContents.lst** | "Follow-up Visit Form, Part B" |
| form06a_PublicContents.lst | 'DropouVDeath Notification" |
|  | "Final Death Notification" |
| **form06b_PublicContents.lst** |  |
| form07_PublicContents.lst | 'Cause of Death Review" |
|  | "Telephone Contact Follow-up" |
| **form08a_PublicContents.lst** |  |
| form09_PublicContents.lst | "PFT — Retrospective Data" |
| form10_PublicContents.lst | "Augmentation Therapy Record" |
| form11_PublicContents.lst | "Adverse Reaction Form" |
|  | "Organ Transplantation Form" |
| **form20_PublicContents.lst** |  |
|  |  |
| Public\ | Data released to the public. |
| WindowsDatasets\ | Windows versions of SAS datasets. |

| Directory\File | Description |
|---|---|
| form00.sas7bdat | "Central Laboratory, Patient Information" |
| form01.sas7bdat | "Screening" |
| form02a.sas7bdat | "Initial Visit Form, Part A" |
| form02b.sas7bdat | "Initial Visit Form, Part B" |
| form03.sas7bdat | "Pulmonary Function Test Results" |
| form04.sas7bdat | "Modified Dyspnea Index" |
| form05a.sas7bdat | "Follow-up Visit Form, Part A" |
| form05b.sas7bdat | "Follow-up Visit Form, Part B" |
| form06a.sas7bdat | "Dropout/Death Notification" |
| form06b.sas7bdat | "Final Death Notification" |
| form07.sas7bdat | "Cause of Death Review" |
| form08a.sas7bdat | "Telephone Contact Follow-up" |
| form09.sas7bdat | "PFT — Retrospective Data" |
| form10.sas7bdat | "Augmentation Therapy Record" |
| form11.sas7bdat | "Adverse Reaction Form" |
| form20.sas7bdat | "Organ Transplantation Form" |
| icd9data.sas7bdat | ICD9 codings needed to build formats library |
| SAScode\ | SAS code used to compile each dataset. |
| F00code.sas | "Central Laboratory, Patient Information" |
| F01code.sas | 'Screening" |
| F02acode.sas | 'Initial Visit Form, Part A" |
| F02bcode.sas | 'Initial Visit Form, Part B" |
| F03code.sas | "Pulmonary Function Test Results" |
| F04code.sas | "Modified Dyspnea Index" |
| F05acode.sas | "Follow-up Visit Form, Part A" |
| F05bcode.sas | "Follow-up Visit Form, Part B" |

| Directory\File | Description |
|---|---|
| F06acode.sas | "Dropout/Death Notification" |
| F06bcode.sas | "Final Death Notification" |
| F07code.sas | "Cause of Death Review" |
| F08code.sas | "Telephone Contact Follow-up" |
| F09code.sas | "PFT – Retrospective Data" |
| F10code.sas | "Augmentation Therapy Record" |
| F11code.sas | "Adverse Reaction Form" |
| F20code.sas | "Organ Transplantation Form" |
| F21code.sas | "Participant Survey Form" |

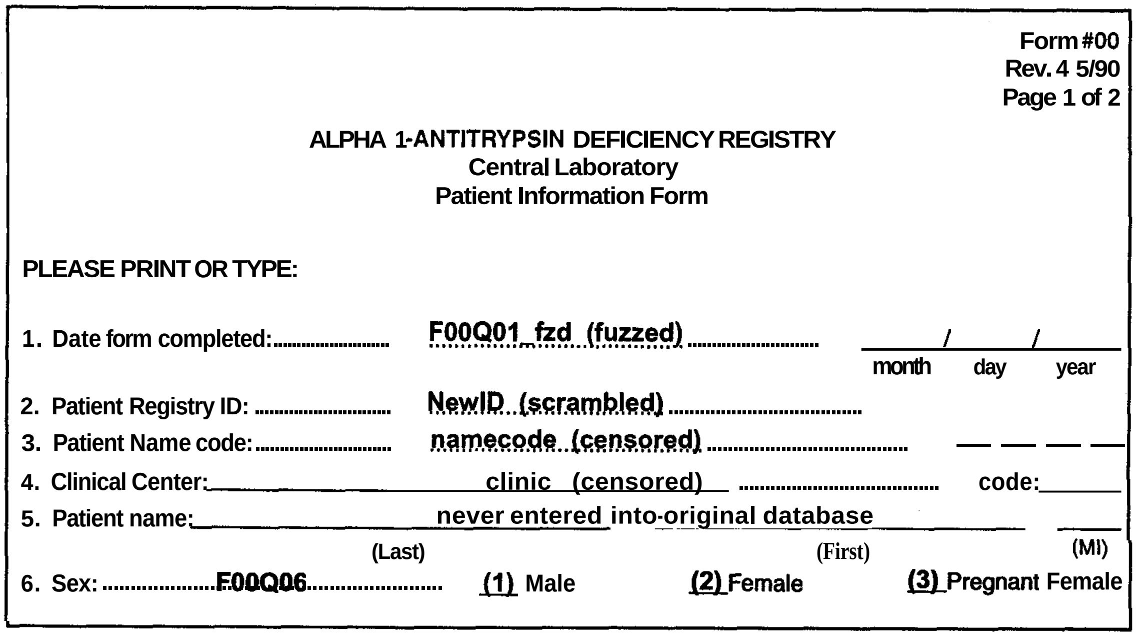
| | |
|---|---|
| Examples\ | Three data analyses using SAS |
| InitialBlood.sas | Basic use of a single dataset |

| Directory\File | Description |
|---|---|
| SmokingDropDeath.sas | Match merges two datasets using `newID` as primary key |
| VisitsPerYear.sas | Handles multiple records per case using PROC **SQL;** match merges two datasets; computes time interval between two fuzzed dates |
| Misc\ | Miscellaneous files |
| NHLBIguidelines.pdf | "Guidelines for Preparation of Data Sets for Delivery to NHLBI for Eventual Public Release" (October 14, 1999) |
| FieldComments.txt | Variable by variable log of all reviews and modifications. |
| OperationsManual\ | Operations manual for the Registry, by chapters. |
| 0_Protocol.pdf | |
| 1_Introduction.pdf | |
| 2_Design.pdf | |
| 3_RegistryStructure.pdf | |
| 4_Activities.pdf | |
| 5_CCCStructure.pdf | |
| 6-LabProcedures.pdf | |
| 7_FormIntro.pdf | |
| 8_References.pdf | |
| 9_PFDataSubmission.pdf | |
| RegistryPersonnel.pdf | Listing of Registry personnel. |
| PublicatonsToDate.pdf | The first pages of all Registry-based publications to date (July 2001). |

## 3 Public Files

The several types of public files are:

- *Annotated case report forms (CRFs) and the instructions usedfor completing them.* The annotations include the variable names, *so* this set serves as a codebook for the **SAS** datasets. For example, the **first** part of Form 00 **looks like:**



Form #00
Rev. 4 5/90
Page 1 of 2

ALPHA 1-ANTITRYPSIN DEFICIENCY REGISTRY
Central Laboratory
Patient Information Form

PLEASE PRINT OR TYPE:

1. Date form completed:........................ F00Q01_fzd (fuzzed).......................... ____/____/____ month day year
2. Patient Registry ID: .......................... NewID (scrambled)....................................
3. Patient Name code:.......................... namecode (censored).................................... — — — —
4. Clinical Center:________ clinic (censored) ...................................... code:______
5. Patient name:________ never entered into original database ________ ____
(Last) (First) (MI)
6. Sex: .....................F00Q06........................ (1) Male (2) Female (3) Pregnant Female

- *Public SAS datasets.*

SEE APPENDIX

Note: Subjects were measured repeatedly over time, so some of the datasets have multiple records per subject. See the example "**8.3** VisitsPerYear.sas" on page 15.

- *PROC CONTENTS summaries of all datasets.* Note that all variable names remain in the database even though some have all values censored. For Form 00, the contents list begins as:

```
                 -----Variables Ordered by Position-----

 # Variable   Type Len Po Format    Informat Label
 ---------------------------------------------------------------------------
 1 F00Q01_fzd Num    8   0 MMDDYY10.          Date Form Completed, fzd
 2 newID      Num    8   8                    Patient Registry ID, Scrmbld
 3 namecode   Char   5 208 $MISSC.    $5.     Patient Name Code
 4 clinic     Num    8  16 MISSF.     6.      Clinical Center
 5 F00Q06     Num    8  24 SEXF.      4.      Sex
 6 F00Q07_fzd Num    8  32 MMDDYY10.          Date of Birth, fzd
```

- All SAS programs used to create **SAS** datasets,
Some code has been censored, such as the method to disguise registry patient identification (ID) numbers and the initial seed numbers used to add "**fuzz**" to all dates. See "6 Date and ID Modifications, Missing Value Codes, Variables Dropped" on page 8.

- Rules for changes and a variable by variable log of all review and modification of the datasets.
- Examples of SAS runs to perform an analyses. See "8 Data Analysis Examples" on page 10.

## 5 Format Library (Important!)

The datasets use internal formats, therefore a libname statement *must* be used in **all** SAS runs. In general, use

```
libname library "file specification";
```

The Windows-based directory on the CD is called "WindowsFormats" Thus, if you have copied the entire contents of the CD to a directory called C:\Alpha1CD on a Windows PC, then you would use

```
libname library "C:\Alpha1CD\WindowsFormats\";
```

Note: The SAS code used to create the formats library is included on the CD in order to facilitate creating new libraries, including **those** for platforms other than Windows and Solaris.

## 6 Date and ID Modifications, Missing Value Codes, Variables Dropped

### 6.1 Fuzzed Dates

All date values have been shifted by a randomly determined "fuzz" number of days that is constant for each subject across all datasets. This disguises the true dates (especially the true birthdate), yet any difference between **two** fuzzed dates is identical to the difference between the true dates.

For example, suppose that patient i was born on 9 June 1979 and s/he completed the initial visit on 23 September 1993. The difference (age at initial visit) would be *5220* days (or 14.3 years). Suppose that patient i's fuzz value is $Fuzz_i$ = -37 days. Then the fuzzed birthdate (variable name: `f01q03_fzd`) would be 03 May 1979 and the fuzzed initial visit date (`F2AQ05_fzd`) would be 17 August 1993. **Using** these fuzzed dates, the difference is still 5220 days. Technically, $Fuzz_i$ is a normally distributed random variable with a mean of 0.0 days and a standard deviation of S days, where S is to remain unknown to the public. S is small enough so that *the fuzzed dates could be used as good surrogates **for** the true dates, should such a need arise.*

### 6.2 Disguised Patient ID Number (newID)

The Registry's original five-digit patient ID numbers have no intrinsic connection to public identifiers of the subjects, e.g., their Social Security numbers. Nevertheless, we decided to disguise these ID numbers anyway. Because questions may arise from time to time about specific data values for specific patients, the new seven-digit patient ID numbers (newID) were formed in a way that makes it easier to link them back to the case report forms and databases. This involved both scrambling the digits and adding random digits to certain positions. The algorithm is to remain unknown to the public.

### 6.3 Missing Value Codes

Six types of missing values have been encoded into the SAS datasets:

| | |
|---|---|
| .c | "censored" variable, because this is too sensitive in terms of patient confidentiality |
| .r | "reliability" issues with this variable make it unsuitable for research purposes |
| .n | variable is not research related |
| .a | variable is "not applicable" for this patient visit |
| .e | variable was never entered into the original database |
| .k | value is "unknown," as determined from case report form |

In the Registry's primary database, some variables use the numbers 8 and 9 to define missing values. For the CD version, we changed to a **SAS** missing value, such as .k (unknown), .a (not applicable) or another type.

### 6.4 Variables On CRF, But Dropped From Datasets

Some variables on the form were dropped entirely because they were not considered to be research related. These include such things as hospital/doctor names, addresses, and phone numbers. The CRFs scanned and stored in the CaseReportForms*.pdf files have been annotated to make this clear.

### 6.5 Variables *Not* **On** CRF, And Thus Now Dropped From Datasets

**A** few variables in the primary database at the Cleveland Clinic Foundation are not on the forms, so they were dropped. For example, some are "housekeeping" variables such as the date the form was entered into the database.

## 8 Data Analysis Examples

To help the user get started, we have developed three examples completely within SAS. Of course, other statistical packages can be used as well. But because SAS has excellent tools for managing data, including PROC **SQL,** people favoring other statistical analysis and graphics environments often still continue to manage the data in SAS **and** create analysis datasets for export to other packages.

*If the contents of the CD have been put into a Windows file called C:\Alpha1CD, as suggested **in** Section 1, then the examples should run as is.*

### 8.1 InitialBlood.sas

The example InitialBlood.sas uses **only** the dataset for Form 02b ("Initial Visit **Form,** Part B"), which has one record each for 1126 subjects. It finds routine descriptives on WBC count, hemoglobin, and hematocrit values at initial visit. Input 1a/b gives the code used to set the data and format libraries. Active lines are for Windows; the UNIX Solaris lines are commented out. The code in Input 1b improves the variable names and specifies the analysis. Output 1a shows part of the text file in the InitialBlood.lst file. Output 1b gives the histogram of the hematocrit values with the superimposed fitted density functions based on the Normal distribution and using SAS's default nonparametric kernel density estimator. (The labeling of those curves was applied using Adobe Illustrator). This is a basic example only; the plot here could certainly be improved by a SAS/GRAPH aficionado.

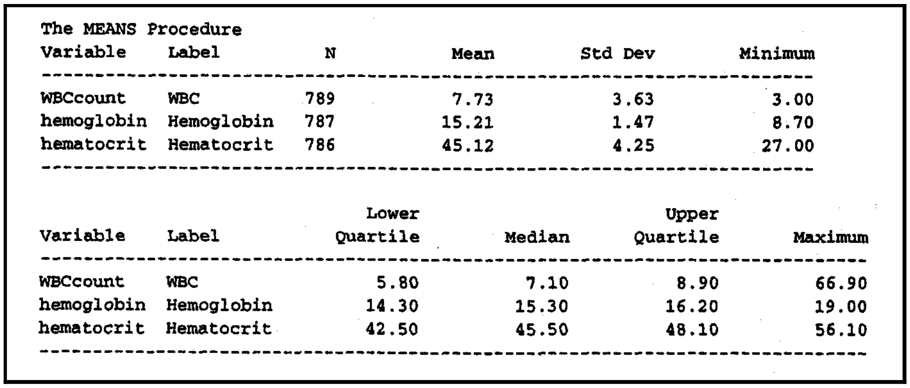
Input la:
InitialBlood.sas
(Part **a)**

```
************************************************************************
InitialBlood.sas
----------------

Using initial visit data (Form 02b, 'Initial Visit Form, Part B."),
compute summary statistics on WBC count (f2bq06c), hemoglobin
(f2bq06d), and hematocrit (f2bq06e).

For hematocrit only, compute additional summary statistics and
produce a relative frequency histogram with overlying curves
showing estimates of the density function.
************************************************************************
*/

options ls=78 nocenter NoFmtErr nodate formdlim='=';

/*
************************************************************************
This example assumes that the entire contents of the CD have been
copied to a hard disk directory called AlphalCD with the CD's
directory structure preserved.

Then just state the path to AlphalCD, as follows:

      %let AlphalCD = complete\path\to\Alpha1CD;
************************************************************************
*/


/*
The following paths were used by Cleveland Clinic programmers.
Yours may be different.
*/
*Solaris: ;
            *%let AlphalCD = /home/alpha1/Alpha1CD;

*Windows: ;
            %let AlphalCD = C:\Alpha1CD;
```

**Input Ib:**
InitialBlood.sas
(Part b)

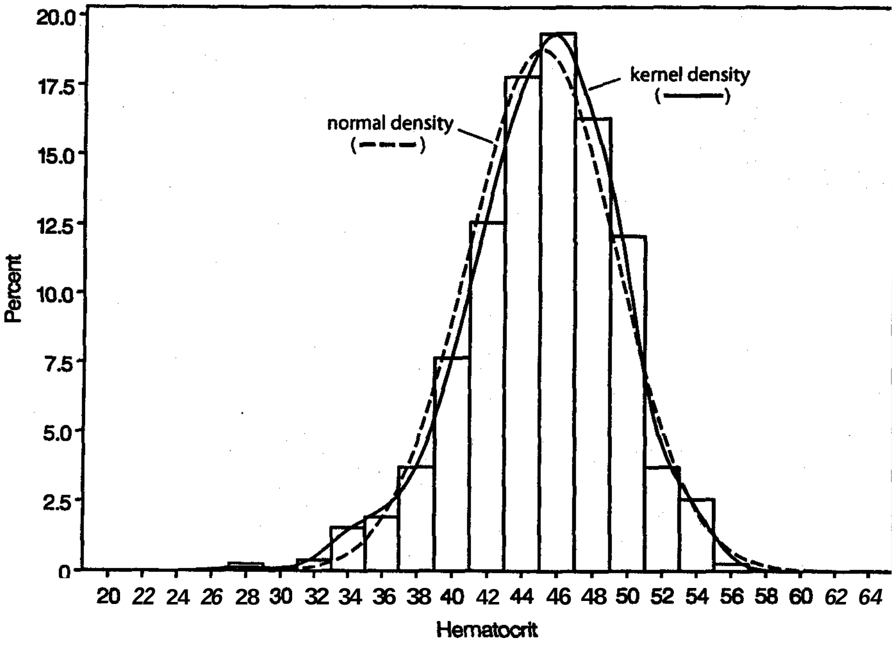
```
/*
Libname specifying the format library.
*/
*libname library "&Alpha1CD./SolarisFormats";
libname library "&Alpha1CD.\WindowsFormats";
/*
Libname specifying where the datasets are stored.
*/
*libname data "&Alpha1CD./Public/SolarisDatasets";
libname data "&Alpha1CD.\Public\WindowsDatasets";

/*
**********************************************************************
Create working (temporary)database copy for data of
Form 02b. For ease of reference, rename the variables, e.g.
f2bq06c to WBCcount.
**********************************************************************
*/
data form02b;
  set data.form02b;
  rename f2bq06c=WBCcount
         f2bq06d=hemoglobin
         f2bq06e=hematocrit;
run;


proc means data=form02b n mean std min q1 median q3 max maxdec=2;
  var WBCcount hemoglobin hematocrit;
run;

/*
**********************************************************************
The density estimates are based on normal ("3" = dashed line) and
kernel ("1" = solid line) methods.
**********************************************************************
*/
goptions rotate=landscape;
proc univariate data=form02b noprint;
   var hematocrit;
   histogram hematocrit /
              normal(color=black l=3)
              kernel(color=black l=1)
              font=swiss height=3
              midpoints-20 to 64 by 2;
run;
```

**Output 1a:**
**Selected text from InitialBlood.lst**



```
The MEANS Procedure
Variable     Label            N            Mean          Std Dev          Minimum
---------------------------------------------------------------------------------
WBCcount     WBC            789            7.73             3.63             3.00
hemoglobin   Hemoglobin     787           15.21             1.47             8.70
hematocrit   Hematocrit     786           45.12             4.25            27.00
---------------------------------------------------------------------------------

                                 Lower                           Upper
Variable     Label             Quartile          Median        Quartile          Maximum
---------------------------------------------------------------------------------------
WBCcount     WBC                   5.80            7.10            8.90            66.90
hemoglobin   Hemoglobin           14.30           15.30           16.20            19.00
hematocrit   Hematocrit           42.50           45.50           48.10            56.10
---------------------------------------------------------------------------------------
```

**Output 1b:**
**Distribution of hematocrit values at initial visit**



## 8.2 SmokingDropDeath.sas

This example demonstrates how to match merge `two` datasets using the primary key, `newID`. It is a rough analysis associating patients' smoking status at the initial visit (Form 02a, ”Initial Visit Form, Part A”) with whether they dropped out or died at some point during the study (Form 06a, “Dropout/Death Notification”). Input 2a sets the data and format libraries **and** is functionally identical to Input 1a. Input 2b merges the `two` datasets **using** newID to match the cases. It defines the `Smoking` values to be either ”smoking” or ”not smoking” at the time of the initial visit, and the `DroppedOrDied` values to be either ”dropped out,“ ”died,” or ”stayed?”. **A** contingency table is specified using **PROC** FREQ. Output 2 shows part of SmokingDropDeath.lst

**and indicates a significant ($p$ = 0.018) association between the two variables using the common chi-square statistic.**

**Input 2a:**
**SmokingDropDeath.sas**
**(Part a)**

```
***********************************************************************
SmokingDropDeath.sas
--------------------

This rough analysis associates patients' smoking status at the
initial visit (Form 02a, Q#19a & Q#19b) with whether they dropped
out or died at some point in the study (Form 06a, Q#5a & Q#6a).

It demonstrates how to match merge two datasets using the
primary key, newID.
***********************************************************************
*/

options ls=78 nocenter NoFmtErr nodate formdlim='=';

/*
***********************************************************************
This example assumes that the entire contents of the CD have been
copied to a hard disk directory called Alpha1CD with the CD's
directory structure preserved.

Then just state the path to Alpha1CD, as follows:

    %let Alpha1CD = complete\path\to\Alpha1CD;
***********************************************************************
*/


/*
The following paths were used by Cleveland Clinic programmers.
Yours will likely be different.
*/
*Solaris: ;
        *%let Alpha1CD = /home/alpha1/Alpha1CD;

*Windows: ;
        %let Alpha1CD = C:\Alpha1CD;


/*
Libname specifying the format library.
*/
*libname library "&Alpha1CD./SolarisFormats";
libname library "&Alpha1CD.\WindowsFormats";

/*
Libname specifying where the datasets are stored.
*/
*libname data "&Alpha1CD./Public/SolarisDatasets";
libname data "&Alpha1CD.\Public\WindowsDatasets";
```

Input 2b:
SmokingDropDeath.sas
(Part b)

```
data smoking (keep=newID Smoking DroppedOrDied);
/*
f6aq05a: Has patient dropped out? 1=yes 2=no
f6aq06a: Has patient died? 1=yes 2=no
f2aq19b: Do you now smoke cigarettes? 1=yes 2=no
*/
  merge data.form06a data.form02a; by newID;
  length Smoking $11;
  Smoking = "";
  if f2aq19b=1 then Smoking="smoking";
  if (f2aq19a=2 or f2aq19b=2) then Smoking="not smoking';
  length DroppedOrDied $13;

  if f6aq06a = 1 then DroppedOrDied = 'died";
  else if f6aq05a = 1 then DroppedOrDied = 'dropped out";
  else DroppedOrDied = 'stayed?';

  label Smoking = 'Smoking at Initial Visit"
        DroppedOrDied = "Dropped Out or Died During Study';
run;


/*
Assess association between Smoking and DroppedOrDied.
*/
proc freq data=Smoking;
  tables Smoking*DroppedOrDied / NoCol NoPercent Chisq;
run;
```

output 2:
Selected text from
SmokingDropDeath.lst

```
Table of Smoking by DroppedOrDied

Smoking(Smoking at Initial Visit)
             DroppedOrDied(Dropped Out or Died During Study)

Frequency   |
Row Pct     |died    |dropped |stayed? |  Total
            |        |out     |        |
------------+--------+--------+--------+
not smoking |    193 |     31 |    811 |   1035
            |  18.65 |   3.00 |  78.36 |
------------+--------+--------+--------+
smoking     |     15 |      8 |     71 |     94
            |  15.96 |   8.51 |  75.53 |
------------+--------+--------+--------+
Total            208       39      882     1129


Statistics for Table of Smoking by DroppedOrDied

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     2      8.0152    0.0182

<Other test statistics are not shown.>

Sample Size = 1129
```

### 8.3 VisitsPerYear.sas

This example was created to demonstrate how to:

- use **SQL** statements to take datasets having multiple and varying numbers of records per subject (newID) and create a new dataset with one record per subject. The **SQL** language, here embodied in **PROC** SQL, is a powerful tool to manipulate such datasets.
- compute a difference between *two* dates. Because **those** dates have been identically fuzzed within each subject, all true differences have been preserved.

Input 3a gives mostly comments; all active statements are identical to those given in Inputs l a and ***2a.*** Input 3b shows the code that operates on **Form** 05b ("Follow-up Visit Form, **Part** B"), which was completed at each follow-up visit, so there are varying numbers of records **per** subject. These statements serve to:

- count the number of records (visits) per patient (NumFollowUps).
- find the (fuzzed) date (F5BQ05_fzd) of each subject's final visit in the database (DateFinalVisit_fzd).
- compute YearsFollowed, the number **of** years from the initial visit date (F2BQ05_fzd) until DateFinalVisit_fzd.
- compute the ratio

  ```
  VisitsPerYear = NumFollowUps/YearsFollowed.
  ```

- summarize VisitsPerYear **with PROC** UNIVARIATE and plot its distribution using a relative histogram.

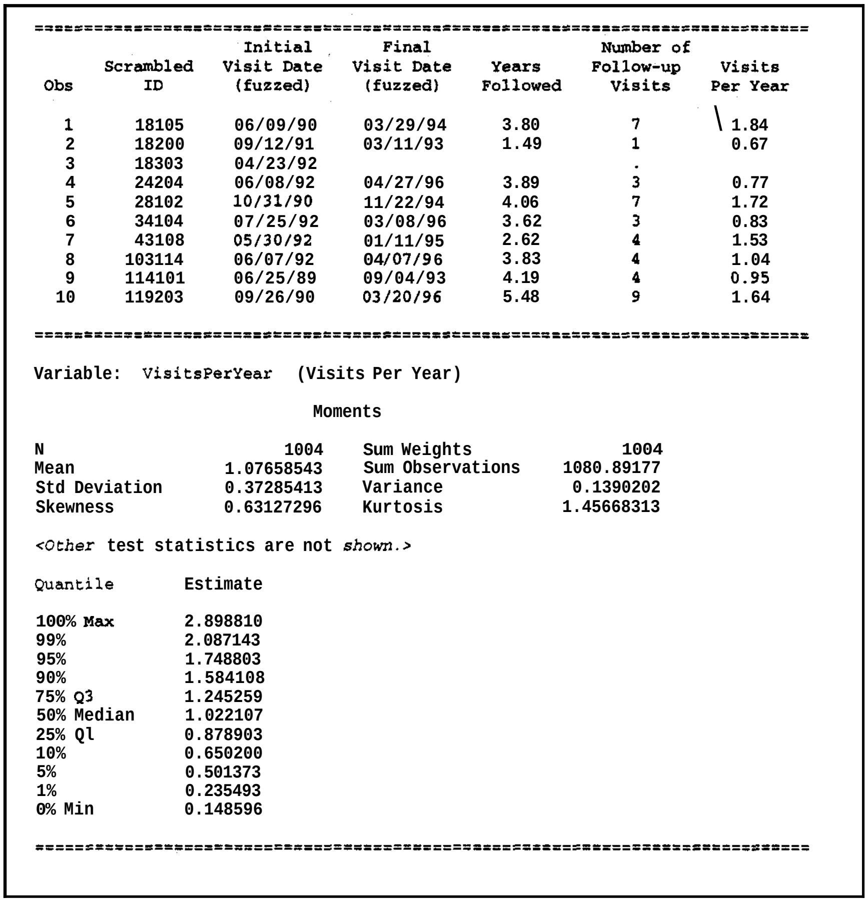
### input 3a: VisitsPerYear.sas (Part a)

```
*****************************************************************
VisitsPerYear.sas
----------- -----
        -

This examines subjects' number of follow-up visits per year, in
order to demonstrate:

  o  using SQL to manage datasets that have multiple records
     (and varying numbers of records) per subject (newID), in
     order to create a new dataset with one record per subject.

  o  computing a difference between two dates that have been
     identically fuzzed so that their true difference is
     preserved.

  o  match merging datasets, using newID as the primary key.

Form 05b (Follow-up Visit Form, Part B) was completed at each
follow-up visit, so there are varying numbers of records per
subject. The SQL language, here embodied in PROC SQL, is a
powerful tool to manipulate such datasets. This demonstration

  o  counts the number of records (visits)per patient
     (NumFollowUps).

  o  finds DateFinalVisit_fzd, the (fuzzed) date (F5BQ05_fzd
     on Form 05b) of each subject's final visit in the database.

  o  computes YearsFollowed, the number of years from the
     initial visit date (F2BQ05_fzd on Form 02b) until
     DateFinalVisit_fzd.

The ratio

          VisitsPerYear = NumFollowUps/YearsFollowed

is summarized with PROC UNIVARIATE and plotted using a relative
histogram.
*****************************************************************
*/

options ls=78 nocenter NoFmtErr nodate formdlim='=';

<Other statements not shown are identical to those in Inputs 1a and
2a..>

libname data "&AlphaCD.\Public\WindowsDatasets";
```

**Input 3b:**
**VisitsPerYear.sas**
**(Part b)**

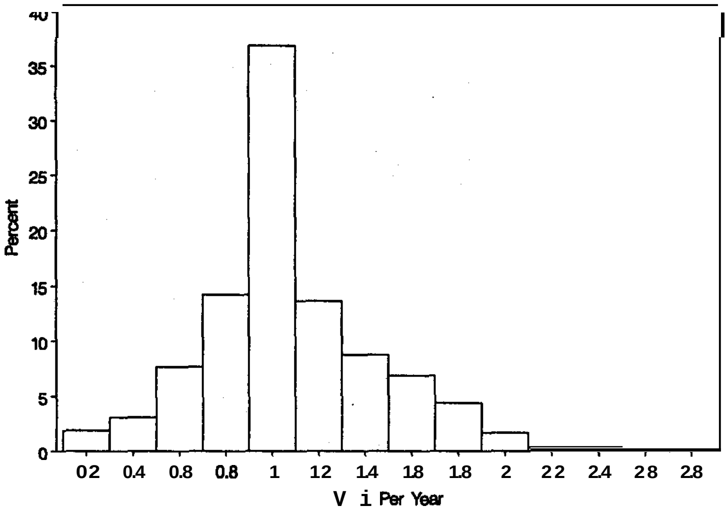
```
data form05b;
  set data.form05b;
  rename F5BQ05_fzd = VisitDate_fzd;
run;

proc sort data=form05b;
  by newID VisitDate_fzd;
run;

proc sql;
   create table FollowUp as
     (select newID, count(*) as NumFollowUps,
                  max(VisitDate_fzd) as DateFinalVisit_fzd
      from form05b
      group by newID);
   quit;


/*
*****************************************************************
This step match merges the two datasets using newID as the
primary key. It also computes VisitsPerYear and specifies labels
and formats for better output.
*****************************************************************
*/
data VisitsPerYearData (keep=newID YearsFollowed DateFinalVisit_fzd
                VisitsPerYear NumFollowUps DateInitialVisit_fzd);
  merge FollowUp data.form02b; by newID;

  DateInitialVisit_fzd = F2BQ05_fzd;
  YearsFollowed = (DateFinalVisit_fzd - DateInitialVisit_fzd)/365.25;
  VisitsPerYear = NumFollowUps/YearsFollowed;

  label NumFollowUps = 'Number of Follow-up visits'
        DateInitialVisit_fzd = 'Initial Visit Date (fuzzed)'
        DateFinalVisit_fzd = 'Final Visit Date (fuzzed)'
        VisitsPerYear = 'Visits Per Year'
        YearsFollowed = 'Years Followed'
        newID = 'Scrambled ID':

  format DateInitialVisit_fzd DateFinalVisit_fzd mmddyy8.
         VisitsPerYear YearsFollowed 4.2;
run;


proc print data=VisitsPerYearData (obs=10) label;
  var newID DateInitialVisit_fzd DateFinalVisit_fzd YearsFollowed
      NumFollowUps VisitsPerYear;
run;


goptions rotate=landscape;
proc univariate data=VisitsPerYearData;
  title1 ' ';
  var VisitsPerYear;
  histogram VisitsPerYear / font=swiss height-3
                            midpoints = .2 to 2.8 by .2;
run;
```

Output **3a:**
Selected text from
VisitsPerYear.lst



```
================================================================================
                    Initial         Final                        Number of
        Scrambled   Visit Date    Visit Date     Years          Follow-up     Visits
 Obs        ID      (fuzzed)      (fuzzed)       Followed        Visits       Per Year

   1      18105     06/09/90      03/29/94        3.80             7        \ 1.84
   2      18200     09/12/91      03/11/93        1.49             1          0.67
   3      18303     04/23/92                                       .
   4      24204     06/08/92      04/27/96        3.89             3          0.77
   5      28102     10/31/90      11/22/94        4.06             7          1.72
   6      34104     07/25/92      03/08/96        3.62             3          0.83
   7      43108     05/30/92      01/11/95        2.62             4          1.53
   8     103114     06/07/92      04/07/96        3.83             4          1.04
   9     114101     06/25/89      09/04/93        4.19             4          0.95
  10     119203     09/26/90      03/20/96        5.48             9          1.64

================================================================================

Variable:  VisitsPerYear   (Visits Per Year)

                           Moments

N                        1004    Sum Weights             1004
Mean               1.07658543    Sum Observations    1080.89177
Std Deviation      0.37285413    Variance             0.1390202
Skewness           0.63127296    Kurtosis            1.45668313

<Other test statistics are not shown.>

Quantile        Estimate

100% Max        2.898810
99%             2.087143
95%             1.748803
90%             1.584108
75% Q3          1.245259
50% Median      1.022107
25% Ql          0.878903
10%             0.650200
5%              0.501373
1%              0.235493
0% Min          0.148596

================================================================================
```

Output **3b:**
Distribution of
Visits Per Year



## 9 Support and ProjectTeam

The Registry, including the production of this public distribution of the database, was supported by contract No. N01-HR-86036 from the National Heart, Lung, and Blood Institute, National Institutes of Health.

The file Misc\AllPersonnel.pdf lists all personnel involved over the entire project period, 1988-2000.

This public distribution of the database was produced by:

| | |
|---|---|
| Principal Investigator | Ralph O'Brien, PhD (robrien@bio.ri.ccf.org, 216.445.9451) |
| Study Coordinator | Susan Sherer, BS (ssherer@bio.ri.ccf.org, 216.444.9935) |
| Systems Analysts | William Crasser, MS<br>Gerald Roberson, MBA<br>Paul Sartori, AD |
| Statistical Programmers | Shannon Neeley<br>Ben Neibaur<br>Jana Shepherd<br>Alvin Van Orden<br>Aimee Wahle |
| Secretary | Lucy Giaimo |

The statistical programmers were summer interns from the Department of Statistics at Brigham Young University.

**Updated information and current contacts for technical support:**
**www.bio.ri.ccf.org/Alpha1CD**

# APPENDIX

How to install The Alphal xpt datasets..
Form00.xpt will be an example for all xpt datasets.

The export file, form00.xpt, is a copy of the Alphal formoo data that is designed to be able to reside on any computer's file system, or to be communicated through any electronic connection between computers, via e-mail, modem, or ftp. Although it is in a very general, very transportable format, the export file needs to be converted into a SAS system file on a local computer before use. We are including instructions on how to install the data on a PC type system with Windows capability. These instructions can easily be modified for other systems.

Installation Guidelines

System requirements

1) A CD-ROM drive with these 17 xport data sets, contents, coding manuals and pdf files require 81 MB of hard drive space. **SAS** versions of the xport data sets require an additional 39 MB of hard disk space.
2) Access to the Statistical Analysis System **(SAS)** software package for PC or on a mainframe.

In the following instructions, the following is assumed:

1) The CD-ROM drive is assigned the letter D:.
2) The hard drive is assigned the letter **C.**
3) The directory you want to store the data in is called C:\Alpha1.

The following program will generate a SAS system file from the form00 XPORT file, assuming it is located on the CD-ROM.

```
libname in1 xport 'd:\Public\WindowsDatasets\form00.xpt';
libname out1 'c:\Alpha1\';
proc copy in=in1 out=out1;                    /*  Create a permanent file    */
```

The following SAS statement will create output which can be compared to the output included after these instructions.

```
proc freq data=out1.form00; tables f00q06 f00q12 f00q13;

run;
```

At the conclusion of this operation point, you will have copied and translated 17 files onto your hard drive to a **SAS** format.

| File Name | No. of Variables | No. of Observations |
|---|---|---|
| form00 | 27 | 1384 |
| form01 | 36 | 1129 |
| form02an | 282 | 1129 |
| form02b | 74 | 1126 |
| form03nu | 107 | 5604 |
| form04 | 13 | 5637 |
| form05a | 311 | 4515 |
| form05b | 81 | 4515 |
| form6a | 21 | 247 |
| form6b | 25 | 204 |
| form7 | 18 | 989 |
| form8a | 16 | 1326 |
| form09 | 14 | 730 |
| form10 | 15 | 32070 |
| form1 1 | 87 | 730 |
| form20 | 27 | 169 |
| icd9data | 3 | 19536 |

## Questions about the Alpha1 files

Please direct any questions or problems to the Division of Epidemiology and Clinical Applications, Epidemiology and Biometry Program, Two Rockledge Centre, 6701 Rockledge Drive, MSC 7934, Bethesda, Maryland 20892-7934, (301) 435-0707 (phone), (301) 480-1667 (fax).

```
                          The SAS System           11:04 Friday, March 7, 2003   2

                         The FREQ Procedure

                                Sex

                                                   Cumulative    Cumulative
             F00Q06       Frequency     Percent     Frequency      Percent
       ---------------------------------------------------------------------
       Male                    765       55.27           765        55.27
       Female                  617       44.58          1382        99.86
       Pregnant Female           2        0.14          1384       100.00


                              Smoking

                                              Cumulative    Cumulative
           f00q12    Frequency     Percent     Frequency      Percent
           ---------------------------------------------------------
                1        1023       74.08          1023        74.08
                2         358       25.92          1381       100.00

                         Frequency Missing = 3


                            Lung Disease

                                              Cumulative    Cumulative
           f00q13    Frequency     Percent     Frequency      Percent
           ---------------------------------------------------------
                1        1084       78.55          1084        78.55
                2         296       21.45          1380       100.00

                         Frequency Missing = 4
```