MANAGING LONGITUDINAL RESEARCH STUDIES:

THE BASIC COMPUTING ENVIRONMENT

By

Linda L Remy, MSW PhD

Ted Clay, MS

UCSF Family Health Outcomes Project
Geraldine Oliva, MD MPH, Director
Jennifer Rienks, PhD, Associate Director
Linda L Remy, MSW PhD, Research Director

500 Parnassus Ave. Room MU-337
San Francisco, California 94143-0900
Phone: 415-476-5283
Fax: 415-476-6051
Web: https://fhop.ucsf.edu/

November 2018

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Suggested Citation

Remy L, Clay T. (2018) Managing Longitudinal Research Studies: The Basic Computing Environment. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

ACRONYMS

AHRQ	Agency for Healthcare Research and Quality
ASC	Ambulatory Surgery Center
BC	Birth certificate file
DOF	Department of Finance
DT or DTH	Death certificate file
FD or FDTH	Fetal death certificate file
ED	Emergency Department
FHOP	Family Health Outcomes Project
HADR	Hospital Annual Disclosure Report
LTC	Long-term care
NCHS	National Center for Health Statistics
OSHPD	Office of Statewide Health Planning and Development
PD or PDD	Patient Discharge (Data)
UCSF	University of California, San Francisco

THE BASIC COMPUTING ENVIRONMENT

INTRODUCTION

This is the first in a series of documents intended to describe basic methods the Family Health Outcomes Project (FHOP) uses to manage its longitudinal research studies [1]. Data analysts working in local health jurisdictions and researchers interested in longitudinal research may find this series helpful. Here we detail our standards for directory structure, naming conventions, directing SAS, and documenting our work. Example programs highlight important materials. All programs and macros are available on request.

We are making this basic methodology and its associated software public to help population health researchers understand the nature of data management for complex longitudinal research studies. In overcoming sometimes-devastating problems, we learned the hard way that following these methods enables us to do more work, be more confident that our work product is accurate, and do studies that are more challenging and need fewer staff than would be possible otherwise. If conventions outlined in this document are in place, SAS programs developed by FHOP should run on most computers. Table 1 identifies datasets now covered by the series.

Table 1. Population data files

Agency	Data Type	From	To
California Department of Health Office of Statewide Health Planning and Development (OSHPD)	Patient Discharge Data (PD or PDD) Emergency Department Data (ED) Ambulatory Surgery Center Data (ASC) Hospital Annual Disclosure Report (HADR) Long Term Care Financial Data (LTC0	1983 2005 2005 YR08 1983	2017 2017 2017 YR42 2017
Center for Health Statistics	Death Statistical Master File (DT or DTH) Birth Statistical Master File (BC or BTH) Fetal Death Statistical Master File (FD or FDTH)	1980 1989 1989	2017 2017 2017
California Department of Finance (DOF)	County-level Population Files and related	1975	2020
US Census Bureau/National Center for Health Statistics (NCHS)	States submitting data to National Cancer Registry: County, ZIP/ZCTA, Census Tract, Block Group, bridged race population	1970	2017
Various sources	Commercial and other products to manage longitudinal changes in geographic data	1960	2017

We have managed other large datasets such as Medicaid or Medicare files, or insurance claims files following the same general processes we describe in this series.

We begin by describing how we organize our computers with a combination of internal and external drives. The section “The TOOLS environment” explains how to set up certain environmental tools that are the same for all projects. The section “The Working Environment” explains the recommended working environment for a specific data analysis project. We then discuss issues associated with the Master File and the Confidential File environments.

COMPUTERS

FHOP's programmer/statistician staff work on stand-alone computers not connected to the UCSF intranet. They use the most confidential versions of administrative datasets. Working on standalone computers helps protect the data.

FHOP was a pioneer in moving enormous population health files from the mainframe to the workstation environment to do longitudinal research. When we first started, agencies sent bulky reel-to-reel tapes that we mounted at the UCSF “Supercomputer”. One year of hospital discharge data would be on seven or eight tapes. We began then to develop many data management processes described in this series of documents.

Recognizing the rapidly growing power of microcomputers, FHOP moved the population data to workstations in the mid-1990s. That was an ugly conversion, but we were “free at last” from mainframe restrictions. The conversion forced us to develop yet more methods specifically to overcome speed and storage capacity limits of that day.

Agencies providing data were slow to make the transition. For a number of years, they continued to send reel-to-reel tapes that we had to convert to the workstation environment. Only in the last ten years did agencies begin to send data on CDs and now DVDs, or more recently by confidential emails, and making non-confidential files publicly available on the web.

Most files FHOP uses are enormous. Working with them efficiently requires fast computers with large storage capacity. We have learned to apportion different drives to different purposes. To protect the data, all drives are BitLocker encrypted, with encryption keys stored in multiple sites for safety. The following describes the drive layout.

Internal drives

- **C DRIVE**. This drive has all general-purpose software, SAS software, and a PERMWORK directory for temporary SAS files. In addition to the usual software and SAS, this drive also has ESRI mapping software and JoinPoint software.

- **D DRIVE**. We refer to this as the Project Drive. Each project has its own directory. This drive also has a directory SASUSER, which stores all macros and documents describing our methods to prepare each major data set.

- **E DRIVE**. Confidential encrypted master population files prepared for analysis are stored here. Each set of master files has its own directory. We refer to this as the Master Drive. We store annual documentation and other key excel files in master subdirectories.

- **F DRIVE**. Non-confidential population and geography-related files, as received, are stored here. This drive also has a directory of papers from SAS/SUGI and related statistical or data management issues.

External drives

- **CONFIDENTIAL DRIVE.** This drive contains zipped and password-protected copies of confidential files sent by the providing agency. Original DVDs and this drive are stored in a secure location when not in use.

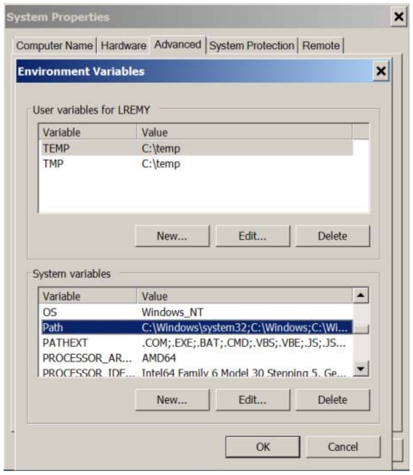
- **SOFTWARE DRIVE.** Another drive contains sub-directories with software packages copied from the original CD/DVD. Installing software is faster from the drive than from the CD. If the program becomes corrupted and must be reinstalled, the dreaded search for the source CD/DVD is avoided. A text or PDF file contains installation key and/or password-related information for each software package. This drive is turned off when not needed.

- **BACKUP DRIVES.** These drives have backup copies of everything on the internal and external drives. We backup our work at least daily to a backup drive. Once a month, we take the external backup to FHOP on the UCSF campus, and return with the drive from the previous month. Because the authors work in tandem, we usually both have most of the files that the other has. In this way, we essentially maintain a triple backup system.

THE TOOLS ENVIRONMENT

The TOOLS environment encompasses the computer environment variables, software, and a variety of macros FHOP has developed to manage longitudinal data. This section describes how to modify environment variables set up the command prompt, and set up FHOP's tools. The section closes by describing two programs that are crucial to our documentation system.

Modify environment variables



Variable	Value
TEMP	C:\temp
TMP	C:\temp

Variable	Value
OS	Windows_NT
Path	C:\Windows\system32;C:\Windows;C:\Wi...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JS...
PROCESSOR_AR...	AMD64
PROCESSOR_IDE...	Intel64 Family 6 Model 30 Stepping 5, Ge...

The following works under 64-bit Windows 7: Select Start | Control Panel | System and Security | System | Advanced System Settings, which opens up a window "System Properties" with the Advanced tab selected. Click the "Environment Variables" button in the lower right. In the Environment Variables window, System Variables list, scroll down to find the Path variable. Click on it to highlight it, then click the "Edit" button. In the "Edit System Variable" window, press the End key to move to the end of the existing value, and add text strings according to the table below. The appropriate strings already may be in your environment variables. Exact strings depend on your operating environment and software and some may be there already from installation. The example is based on 64-bit programs for Windows 7.

Category	**Strings to add (separated by ;)**	**Comment**
SAS	c:\Program Files\SASHome\SASFoundation\9.4;	Required to run SAS programs
Office	C:\Program Files\Microsoft Office\Office16;	Required by SAS programs that call Office products
WinZip (compression utility)	c:\Program Files\Winzip;	Required by SAS programs that zip or unzip
Notepad++	C:\Program Files\Notepad++;	We use Notepad++ to edit SAS programs. We find it more useful than Notepad.
DOS utilities	C:\TOOLS\DOS;	Required to call batch files

PATHEXT is just below PATH. Check that this variable includes the string BAT. If it does not, add the string the same way you added strings to the PATH variable. Remember to end with the semicolon (;).

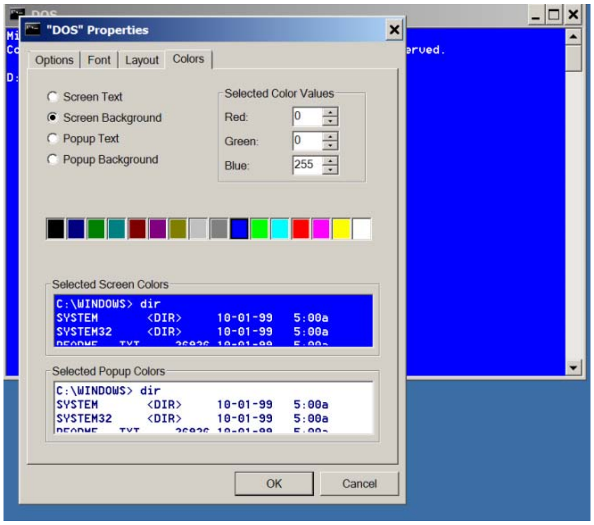
After modifying the PATH and PATHEXT, click OK to implement the changes and exit the control panel. Test the PATH changes you made by opening a DOS window and giving the following commands: “sas”, “excel”, “winword”, and “wzzip”.

Modifying environment variables can cause <<*serious*>> problems if done incorrectly. Unless you are completely comfortable doing this, we recommend asking your computer support specialist to do this.

Set up the Command Prompt

If the Command Prompt is not on your taskbar, add it, or ask your computer support person to do this for you. Start by putting a shortcut for C:\Windows\SysWOW64\cmd.exe on the desktop. Honoring our early computing roots, we renamed this icon DOS and moved it to our taskbar.

You can make this window more “eye friendly”. Click on the Command Prompt (or DOS) icon on the taskbar. Then right click on top of the window, selecting properties. These are a matter of preference. Here is what LR uses.



Red: 0
Green: 0
Blue: 255

```
C:\WINDOWS> dir
SYSTEM        <DIR>      10-01-99    5:00a
SYSTEM32      <DIR>      10-01-99    5:00a
```

- **Options tab.** Cursor Size = Medium; Command History, Buffer Size 50, Number of Buffers = 4; Edit Options, click on Quick Edit Mode and Insert Mode.
- **Font tab.** Lucida Console, Size 18. Save your eyes.
- **Layout tab.** Screen Buffer Size = 80 wide x 300 high; Window Size = 80 wide x 25 high. Click on Let System Position Window.
- **Colors tab.** Screen text - white (box all the way to the right Red = 255, Green = 255, Blue = 255), Screen background - blue (box middle right, Red = 0, Green = 0, Blue = 255), Popup text and Popup background - same as screen text.

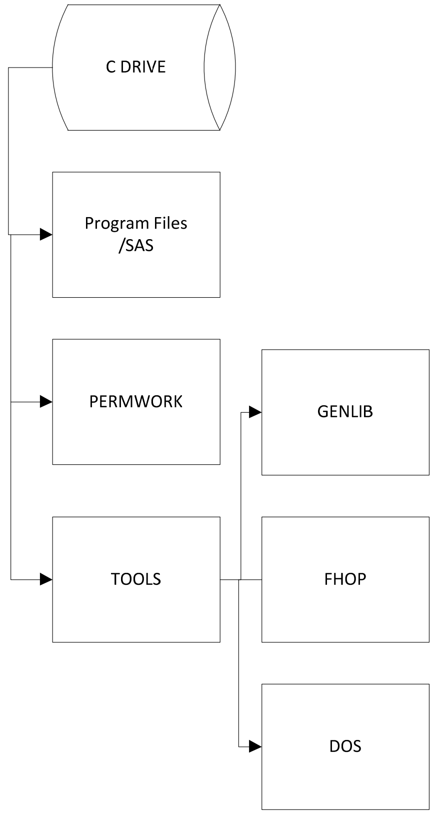
Set up FHOP Tools

Download the file TOOLS.ZIP from the FHOP website to your computer. We recommend unzipping it C:\. This will make a directory TOOLS with several subdirectories. This document assumes you have chosen this location.

The TOOLS environment

Figure 1 shows the required structure for C Drive. We standardize directory conventions for SAS, PERMWORK, and our macro library TOOLS.

Figure 1. Tools Environment



SAS is in default directories that SAS creates at installation.

PERMWORK is where we temporarily store output data while we debug a program. We use this instead of WORK. With PERMWORK, we can continue a program that bombs without starting over.

TOOLS has subdirectories containing macros and DOS batch files that are the core of our system.

GENLIB contains general-purpose macros that are dataset independent.

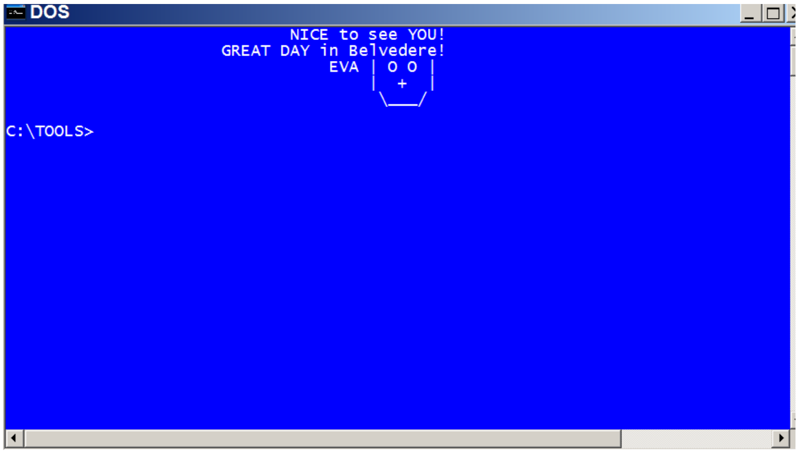
FHOP contains macros for specific tasks, primarily data file construction, classification, summary, and reporting.

The **DOS** directory contains batch files that primarily enable us to maneuver around the computer without having to leave the DOS window. It also contains batch backup files.

For example, upon submitting the string TOOLS from any DOS window, the directory changes to the window at the right. Here is the contents of TOOLS.BAT

```
@C:
@cd C:\TOOLS
@cls.bat
@au.bat
```

CLS.BAT clears the screen.
AU.BAT calls the spirit of
LR’s computer, ***Eva Curie*****.**



```
                                NICE to see YOU!
                         GREAT DAY in Belvedere!
                                    EVA | O O |
                                        |  +  |
                                         \___/

C:\TOOLS>
```

The analyst is working in Belvedere, on a computer named Eva Curie, on the C Drive in the TOOLS directory.

Programs required for the TOOLS environment

FHOP writes SAS programs to run in “batch mode”. That means the programmer submits a SAS program file to SAS, which starts up, executes the program, produces a .LOG and .LST file, and terminates. Work is organized into “projects”, each with its own parent folder and sub-folders. The PGMS sub-folder stores all programs, logs and listings.

The PGMS sub-folder contains a file AUTOEXEC.SAS. This defines settings that tailor the SAS environment to the needs of the particular project. Among other things, AUTOEXEC.SAS defines the location of standard tools and options, and contains all LIBNAME statements to identify data locations. Normally, AUTOEXEC.SAS is the only file that needs to be modified before running FHOP SAS programs.

To run a SAS program, open Command Prompt (DOS), change to the project programs directory, and give the DOS command to submit a program to SAS. Because the full syntax for submitting a program is somewhat long, the FHOP tools include a utility command batch file to abbreviate the syntax of running a program.

STARTUP.SAS is called in the second line of every FHOP program, where the first line identifies the program name. STARTUP first verifies that the internal program name equals the external program name. If they are not identical, the program terminates. Next, STARTUP defines a standard listing footnote. The third line of every FHOP program contains the line CARDS4 followed by text lines through a sequence of four semicolons “;;;;”. Text lines between CARDS4 and “;;;;” are read in as variable lines. Additional observations with standard information are added: program name, date, time, location, analyst, operating system, and SAS version. Program name, date, and time are added to each observation. The resulting information is written out to a temporary file and echoed to the SAS log.

STARTUP finishes by reading a check list into the program SAS log. The purpose is to focus the analyst on reviewing the log to verify that SAS has not reported a variety of potential problems. In the “old days”, when we used to maintain paper binders logging our studies, the analysts had to check and sign the CHECKLIST page.

STOREDOC.SAS is called in every FHOP program, in the line following the sequence of four asterisks “;;;;”. STOREDOC begins by writing the temporary file STARTUP made to the permanent file PGMLOG. This overwrites previous documentation for a given program with the current version. Next STOREDOC writes a single line describing the job to the cumulative RUNS.DTA in C:\TOOLS and to RUNS.DTA in the study SAS directory.

THE WORKING ENVIRONMENT

Copy and modify the “generic project” structure

We have learned from grim experience that people working on the same study on different computers in different locations must use the same directory structure. We have developed directory structures based on function: basic tools, master files, and working environments. Optimally, for safety, each environment will be on physically different drives.

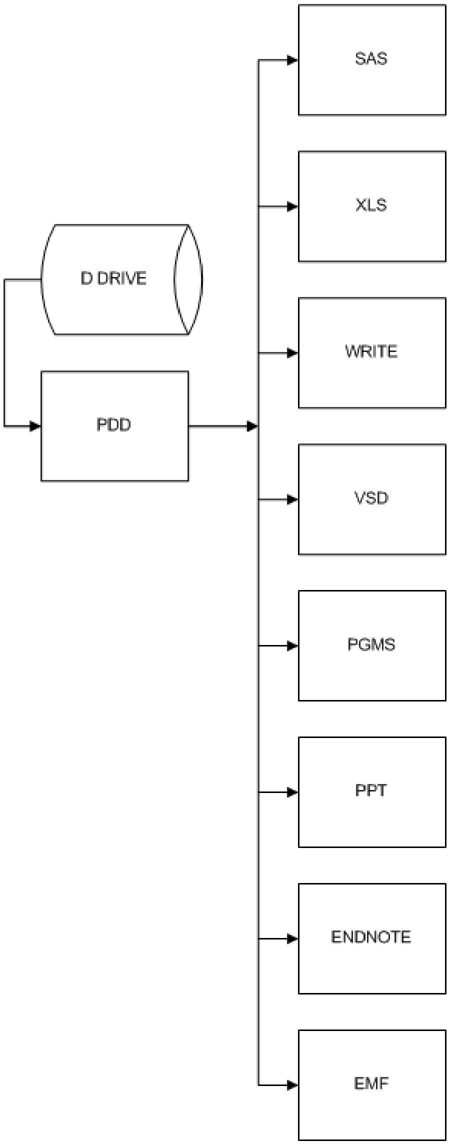
Under the TOOLS directory, find the folder “Generic_Project”. Copy this folder and paste it into the desired working environment drive. Change the name of “Generic_Project” to the study’s short name. For example, a directory devoted to prepare longitudinal patient discharge data for analysis would have the short root directory name PDD.

In the PGMS directory of renamed “Generic_Project”, modify AUTOEXEC.SAS to contain the correct location of the project. Modify LIBNAME statements to reflect the true locations of data on your system. These are discussed in the next section.

Run the program TEST.SAS to verify that your environment is working correctly. If you are batch submitting from a DOS window, modify T.BAT with the correct path so it will work.

Figure 2 shows the structure of the working environment. The example is the PDD directory on the D Drive (Diana Prince) on computer Eva Curie. We submit SAS programs to make master files from this drive and do all subsequent work calling those files, for example, to make format libraries for this dataset. Staff working on a given study agree on the root directory name, and maintain similar subdirectory structures. The following identifies the types of files stored in the subdirectories.

Figure 2. Working Environment



SAS. Files SAS creates for a given study are stored here. The sole exception is that master files are stored on a different drive in the MASTER environment. This directory also stores study-specific files RUNS.DTA and PGMLOG created and maintained by STOREDOC.SAS, described earlier. *We require that permanent files include the name of the program that created them as the first part of their name*.

XLS or more recently XLSX. Excel files are stored here. Excel files output by a SAS program must have as their prefix the name of the program that created them. When we copy information in these files to another file, for example Word, we embed the full path and name of the file source in the document as hidden text.

WRITE. Working documents created during the course of the study, e.g., memoranda, correspondence, reports.

VSD or more recently VSDX. VISIO diagram files made to document data flow are stored here. When we copy information in these files to another file, for example Word, we embed the full path and name of the file source in the document as hidden text.

PGMS. SAS programs, logs, and listings are stored here.

PPT or more recently PPTX. PowerPoint files related to this body of work are stored here.

ENDNOTE. This directory stores PDF journal articles pertinent to the given study. When we reference a PDF in writing a report, we embed the full path and name of the file source in the document as hidden text. When a study results in multiple papers, each can have subdirectories.

EMF. This directory stores EMF or TIFF files (graphs, maps) created for the study. Graphic files are required to have as a prefix the name of the program that created them. When we import them into another file, for example Word, we embed the full path and name of the graphic source file as hidden text.

The working environment can have other subdirectories. Everyone working on the given project agrees on subdirectory names.

Programs for the working environment

After making a working environment, we make basic management files. These are located in the PGMS directory.

AUTOEXEC.SAS. When SAS starts to execute a program, it reads the file AUTOEXEC.SAS. This sets default configuration options for a working directory. We customize AUTOEXEC.SAS for each project. For example, some projects have their own format and macro libraries. You also may need standard libnames, formats and macro libraries. The following is an example for the working directory PDD, where we prepare OSHPD patient discharge datasets for analysis.

%let studyttl = OSHPD Patient Discharge Data;
%let study = OSHPD PDD;
%let insti1 = University of California, San Francisco;
%let insti2 = Family Health Outcomes Project;
%let invest1 = Gerry Oliva MD MPH;
%let invest2 = Linda Remy PhD;
%let analyst = LREMY;
%let location = MYTOWN EVA CURIE D:\PDD\;
%let TOOLS = C:\TOOLS;
title1 "&STUDYTTL";
%PUT ;
	These lines define the study, institution, investigators, analyst, and location for OSHPD's patient discharge data (PDD).
	We reserve title1 as the first line in every program listing.

%PUT [[STUDY]];
%PUT [[STUDY: &STUDYTTL;
%PUT [[INSTITUTION: &INSTI1;
%PUT [[&INSTI2;
%PUT [[INVESTIGATOR: &INVEST1;
%PUT [[&INVEST2;
%PUT [[LOCATION: &LOCATION;
%PUT [[TOOLS: &TOOLS;
%PUT [[DATE: &sysdate &systime;
%PUT ;
	STARTUP.SAS echoes this information to the program log through these put statements.

%PUT [[WORKING ENVIRONMENT]];
%let TOOLS   = C:\TOOLS;
%LET GENLIB   = &TOOLS\GENLIB;
filename GENLIB   "&GENLIB";
libname GENLIB   "&GENLIB";
%let FHOP   = &TOOLS\FHOP;
filename FHOP   "&FHOP";
libname FHOP   "&FHOP";
libname user   "C:\PERMWORK";
%PUT ;
	This section defines the Working Environment. We identify the location of the tools directory, and the macro libraries, and the project-specific file documenting each submission of a SAS job. This section does not change for any study
	Notice that we use both %let and libname statements to define the same locations. Let statements are useful in the macro environment.
	We store temporary files in PERMWORK. This is helpful when developing programs, because we can restart from the last datastep before the program bombed.

%PUT [[REQUIRED PROJECT LIBRARIES]];
%let PROJECT   = D:\PDD;
%let PROJPATH   = "&PROJECT";
%let SAS   = &project\SAS;
%let REPORT   = &project\REPORT;
%let PDD   = &project\SAS;
%let XLS   = &project\XLS;
%let EXLS   = E:\PDD\XLS;
%let RAW   = X:\PDD\RAW;
libname SAS   "&PROJECT\SAS";
libname PDD   "&PROJECT\SAS";
libname LIBRARY   "&PROJECT\SAS";
libname STUDYLIB   "&PROJECT\SAS";
%PUT ;
	This section defines the project environment.
	This section can change for most studies.

%PUT [[MASTER ENVIRONMENT]];
%let ESAS   = E:\PDD\SAS;
%let MASTER   = &ESAS;
%let EXLS   = E:\PDD\XLS;
%let RAW   = X:\PDD\RAW;
libname ESAS   "&ESAS";
libname MASTER   "&ESAS";
libname RAW   "&RAW";
%PUT ;
	This section defines the Master Environment.
	This is where we store master files with variables already standardized for longitudinal research. If needed, certain confidential variables have been encrypted.
	RAW defines the removable encrypted external drive with incoming confidential unencrypted data. This drive is disconnected and stored in a safe place when not in use. When RAW is disconnected the log will say "Library RAW does not exist."

%PUT [[OTHER NEEDED LIBRARIES]];
libname AHDR   "D:\AHDR\SAS";
libname AHRQ   "D:\AHRQ\SAS";
libname ASC   "D:\ASC\SAS";
libname ED   "D:\ED\SAS";
libname EASC   "E:\ASC\SAS";
libname EED   "E:\ED\SAS";
%PUT ;
	This section defines other libnames for a given study.
	The AHRQ directory contains data prepared to be read into our format library. The format library is defined as PDD. We store formats for all OSHPD patient files here: discharge, emergency department, ambulatory surgery center.

%PUT ;
%PUT [[POPULATION LIBRARIES]];
libname POP   "D:\POP\SAS";
libname CEN1970 "D:\CEN1970\SAS";
libname CEN1980 "D:\CEN1980\SAS";
libname CEN1990 "D:\CEN1990\SAS";
libname CEN2000 "D:\CEN2000\SAS";
libname CEN2010 "D:\CEN2010\SAS";
%PUT ;
	The population libraries are standardly defined for each study. POP contains population data from California's Department of Finance, already prepared to make population files for a given study. Census files also have population and other data that we can use as needed.

%PUT [[GEOGRAPHY LIBRARIES]];
libname GEOG   "D:\GEOG\SAS";
libname CALMAPS   "D:\CALMAPS\SAS";
%PUT ;
	The directory GEOG contains information describing various levels of geography. CALMAPS has files needed for mapping.

* General options; options nocenter missing = ' ' ls = 200 ps = 64 errorabend macrogen mprint nodate noovp noxwait source source2;	We set standard general starting options that assure our macros will run. However, the programmer can change options in a given program as needed.

* Location-specific options;
options   sasautos = ("&GENLIB" "&FHOP" "&PROJECT\PGMS" SASAUTOS) fmtsearch = (USER WORK PDD GEOG CALMAPS CEN2000 CEN2010);
	Sasautos location are standard across projects. Fmtsearch may need to change from study to study.

* Define DOS commands which compress or extract zip files on this system. ;
%LET ZIPCMD = wzzip;
%LET UNZIP = wzunzip;
	Some confidential raw files are stored as ZIPS. They have to be unencrypted before programs read them into SAS. Some production macros require us to zip the products at the end, so we can distribute them.
	The programmer must modify the PATH environment to show the path where called programs are located. SAS programs can use these macro variables to be transportable between systems that may have a different zip/unzip commands.

* Other settings;
%let PGMLOG = SAS.PGMLOG; * SAS documentation data set;
%let RUNSLOG = &sas\RUNS.DTA &tools\RUNS.DTA;
%let TABS   = delimiter = "09"x;
%let KEY   = xxxxxxxxx;
	The KEY is blocked out. It is confidential and cannot be distributed.

CORE DOCUMENTATION. The first lines of every FHOP program provide basic information that forms the core of our documentation system.

%LET PGMNAME=PGMLOG; %STARTUP; cards4; PURPOSE: Create the study program log DESCRIPTION: This is the first program to begin a study. NOTE: Rerunning to convert from SAS V6 INPUT: NONE OUTPUT: D:\PDD\SAS\PGMLOG SOURCE: TED CLAY 13-Feb-1996 started EDITED: LINDA REMY on EVA CURIE on 07-May-2011 ;;;; %STOREDOC	Every program FHOP writes must have the lines between %let pgmname = and %STOREDOC at the top. Each time a SAS job is submitted, STOREDOC writes this information in the study file, PGMLOG. If this is a resubmission, the current version overwrites the previous. Information echoes to the program log, giving basic information on the program. PURPOSE is limited to one line long. DESCRIPTION can be as many lines as needed to document the program, using natural language. We limit line length to 80 characters. If we copy a program from another directory and later find a bug, the SOURCE line tells us we may need to return there and rerun that previous study from that point forward to fix consequences of the bug originating there. EDITED tells us who last worked on the file, the computer, and the date work started on the program.

PGMLOG.SAS. This program maintains basic documentation for each study, automating the creation and maintenance of the study book. After editing for study-specific content, PGMLOG is the first program submitted at the start of a study. It also is the last program submitted at study end. After sorting PGMLOG by date, time, and program name, the program prints the exact order jobs need to run to replicate the results, with a one page per program summary. When we make the final study book, we delete "dead end" programs from the listing.

`proc sort data = SAS.PGMLOG out = pgmlog;`
`  by DATE TIME PGMNAME;`
`run;`
 	
`proc print data=pgmlog noobs;`
`  by DATE TIME PGMNAME;`
`  pageby PGMNAME;`
`  var line;`
`  title2 'Program Documentation';`
`  format date date9. time time5.;`
`run;`
	Each time a SAS job is submitted, STOREDOC writes documentation information in the study file, PGMLOG. If this is a resubmission, the current version overwrites the previous. Information echoes to the program log, giving basic information on the program.
	PGMLOG must be submitted two times at the start of the project.

Submitting programs

We submit programs to SAS using a batch process, either through a DOS window or the Windows desktop. LR adds a line to RUN.BAT with the name of the program, another way to track work chronologically. When we need to rerun a sequence of programs, we only need to delete the @rem from each line. The following shows some helpful batch (BAT) files.

`RUN.BAT`
`@rem @SAS -nodate -sysin D:\PDD\PGMLOG.SAS`
`@SAS -nodate -sysin D:\PDD\FORMATS.SAS`
	@rem comments out commands that do not need to be executed at this time.

`T.BAT`
`@SAS -nodate -sysin D:\PDD\TEST.SAS`
	When we are trying to debug a program, or need to get information about something, we run "side programs" named temp.sas or test.sas. These are invoked using T.BAT.

`CH.BAT or CHECK.BAT`
`@sas -sysin c:\tools\genlib\checklog.sas -sysparm '%1.log'`
`Notepad++ checklog.lst`
`erase checklog.*`
	CH.BAT checks the log after a program runs, and returns a listing of the type and number of issues that arose. This helps the analyst to know quickly if the program ran without problems. Invoke by typing ch <programname>. Do not include log.

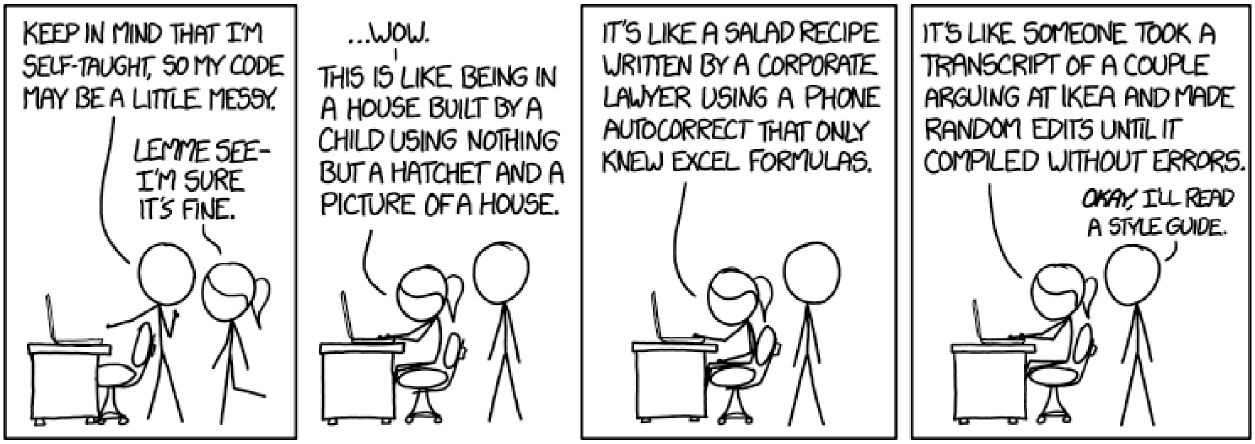
`DIFF.BAT`
`@SAS -sysin C:\TOOLS\GENLIB\DIFF.SAS -sysparm '%1 %2'`
`Notepad++ diff.lst`
	Compares two programs to highlight version differences. Invoke by typing DIFF OLD.SAS NEW.SAS. This is very helpful in understanding program version changes.

Notice that CH.BAT and DIFF.BAT include the string “notepad++”. That is calling the freeware NotePad++ text editor to view the file [2]. We use this instead of Notepad because it allows column block moves. The text editor is the analyst’s choice. Be sure to edit batch files that call other programs, for the program to run successfully.

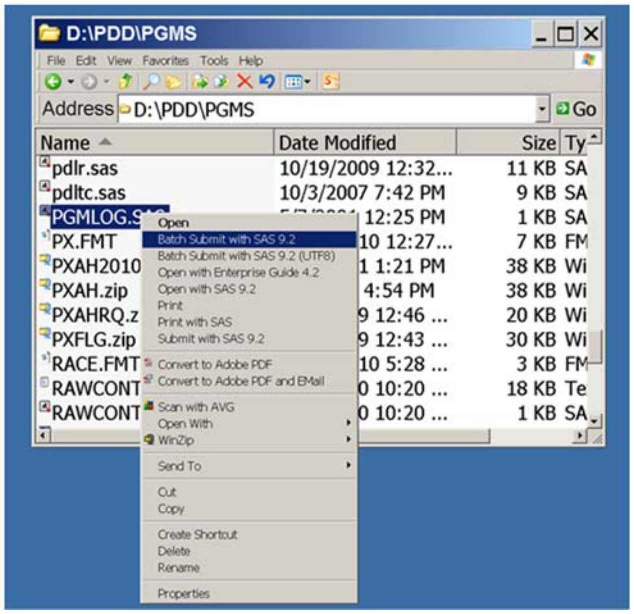
Many people write and submit their programs in the SAS window. This will create a program, log and listing that can be viewed in the SAS window, but many times people forget to save those files. Our programs are in the public domain. The agencies that support FHOP require us to give them our work product.

We hired two programmers who refused to work outside of the SAS window. They rarely saved any work product other than the final files. As each of them left, we had absolute nightmares trying to reconstruct their work. Much of their surviving work product was loaded with errors. We had to redo all their work, because we had no audit trail and could not trust their product. Learning from this, we no longer hire SAS cowboys. Everyone working for us batch submits.

We do not know where we got this unsigned cartoon, but hopefully readers will enjoy it.

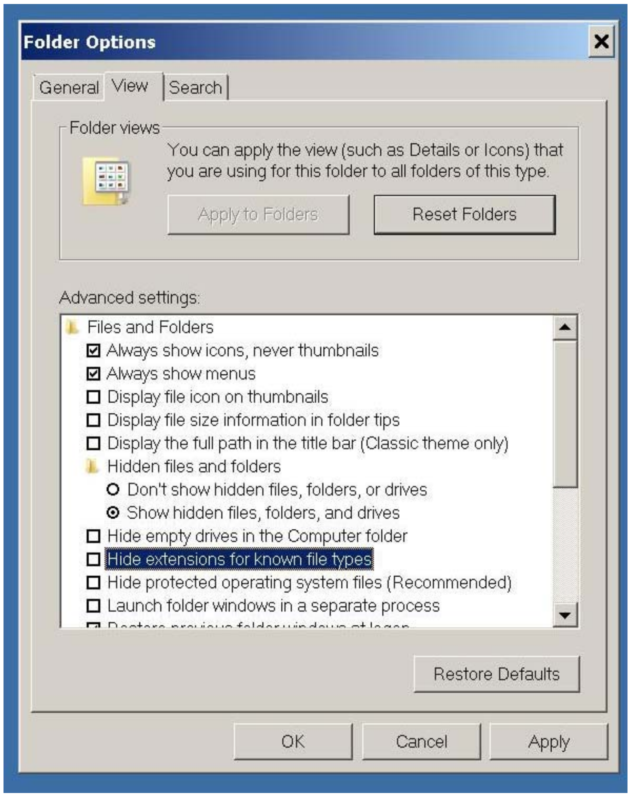

KEEP IN MIND THAT I'M SELF-TAUGHT, SO MY CODE MAY BE A LITTLE MESSY.
LEMME SEE– I'M SURE IT'S FINE.
...WOW.
THIS IS LIKE BEING IN A HOUSE BUILT BY A CHILD USING NOTHING BUT A HATCHET AND A PICTURE OF A HOUSE.
IT'S LIKE A SALAD RECIPE WRITTEN BY A CORPORATE LAWYER USING A PHONE AUTOCORRECT THAT ONLY KNEW EXCEL FORMULAS.
IT'S LIKE SOMEONE TOOK A TRANSCRIPT OF A COUPLE ARGUING AT IKEA AND MADE RANDOM EDITS UNTIL IT COMPILED WITHOUT ERRORS.
OKAY, I'LL READ A STYLE GUIDE.

FHOP is very serious about structuring our programs for readability and "style". Our processes ensure that we keep the program, log, and listing generated by SAS. These files document each step in a study. This enables us to maintain an audit trail for our work. In our view, not saving these files is akin to a crime laboratory technician testifying in court as to the results of a blood test and then failing to provide documentation of the work.



Name	Date Modified	Size	Ty
pdlr.sas	10/19/2009 12:32...	11 KB	SA
pdltc.sas	10/3/2007 7:42 PM	9 KB	SA
PGMLOG.S	12:25 PM	1 KB	SA
PX.FMT	10 12:27...	7 KB	FM
PXAH2010	1 1:21 PM	38 KB	Wi
PXAH.zip	4:54 PM	38 KB	Wi
PXAHRQ.z	9 12:46 ...	20 KB	Wi
PXFLG.zip	9 12:43 ...	30 KB	Wi
RACE.FMT	10 5:28 ...	3 KB	FM
RAWCONT	0 10:20 ...	18 KB	Te
RAWCONT	0 10:20 ...	1 KB	SA

It also is possible to batch submit from the desktop. To do this, right click on the program name, and click on "Batch submit with SAS". This will work, although it will be slightly more difficult to maintain the history of program chronology.



You can apply the view (such as Details or Icons) that you are using for this folder to all folders of this type.

Advanced settings:

- Files and Folders
  - ☑ Always show icons, never thumbnails
  - ☑ Always show menus
  - □ Display file icon on thumbnails
  - □ Display file size information in folder tips
  - □ Display the full path in the title bar (Classic theme only)
  - Hidden files and folders
    - ○ Don't show hidden files, folders, or drives
    - ● Show hidden files, folders, and drives
  - □ Hide empty drives in the Computer folder
  - □ Hide extensions for known file types
  - □ Hide protected operating system files (Recommended)
  - □ Launch folder windows in a separate process
  - ☑ Restore previous folder windows at logon

If you decide to batch submit from the desktop, it is helpful to reset Folder Options to display the file extensions (e.g., log, lst, sas7bdat, etc). To do this, click on Start | Control Panel | Folder Options. On the View tab, unclick “Hide extensions for known file types. This will give you a display to identify the differences between the SAS program and the log and lst files it produces.

THE MASTER FILE ENVIRONMENT

Master files are prepared for case selection. We require all variables to be labeled and, where appropriate, formats applied. File names, file locations, variable names, and variable formats are longitudinally consistent. We store master files on a physically different drive from study-specific files. We think of the master file environment as our "well". This is where we store basic documentation that arrives with the files, define how to preprocess master files, summarize basic results to verify that programs did what we intended, and store results as SAS files.

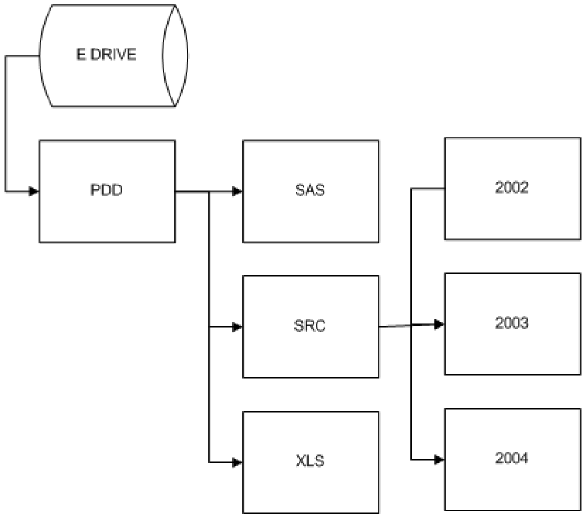
Table 2 summarizes information about the master files. In a given cell, “Y” means materials are in the public domain and available upon request. DOC identifies if source documentation is available. EXCEL identifies if the excel file to read the data into SAS is available. PGMS means the SAS programs are available. DATA means source data files are available. SAS indicates if the final SAS files are available for distribution.

Table 2. Master file environment for confidential and non-confidential population files

Directory	Contents	Doc	Excel	Pgms	Data	SAS
AHDR	Office of Statewide Health Planning and Development Annual Hospital Disclosure Reports	Y	Y	Y	Y	Y
AHRQ	Agency for Healthcare Research and Quality (Clinical Classification System and related software and documentation)	Y	Y	Y	Y	Y
ASC	Office of Statewide Health Planning and Development Ambulatory Surgery Center	Y	Y	Y	N	N
BC	California Department of Vital Statistics Birth Certificates Data	Y	Y	Y	N	N
CEN1970-CEN2020	1970-2010 US Census, at state, county, place, tract, and ZCTA levels (ZCTA 1990 forward)	Y	Y	Y	Y	Y
DEATH	California Vital Statistics Death Files	Y	Y	Y	N	N
DOF	California Department of Finance Population Estimates	Y	Y	Y	Y	Y
ED	Office of Statewide Health Planning and Development Emergency Department Data	Y	Y	Y	N	N
FDTH	California Vital Statistics Fetal Death Files	Y	Y	Y	N	N
LTC	Office of Statewide Health Planning and Development Long Term Care Data	Y	Y	Y	Y	Y
PDD	Office of Statewide Health Planning and Development Patient Discharge Data	Y	Y	Y	N	N
NCHS Population	National Center for Health Statistics longitudinal population estimates to county-level by detailed bridged race, ethnicity, age and sex	Y	N	Y	Y	Y
USPS	US Postal Service ZIP-code data	Y	N	Y	Y	Y
ZIPCODES	ZIP-Code changes through 2016	Y	N	Y	Y	Y
ZP4	Data to make CASS-certified Addresses	Y	N	Y	N	N

For example, Figure 3 shows the standard directory structure in the Master File Environment. The example is the PDD directory on the E Drive (Eudemon) on computer Eva Curie.

Figure 3. Master File Environment



SAS has master files after they have been read into SAS. We standardize all variables longitudinally and encrypt confidential variables.

SRC has the original documentation, stored by file year, in the case of PDD, 1983 forward.

XLS has the Excel file SAS calls to structure incoming and outgoing files. It also contains Excel files documenting basic information about received files, their internal structure, and longitudinal frequencies for selected variables.

We use this directory structure for every master data source and limit subdirectories to 3 levels.

THE CONFIDENTIAL FILE ENVIRONMENT

We receive confidential master files (hospital, birth, death, fetal death, etc.) on CDs or increasingly via confidential web download, prepared according to rules of the agency sending the data. These are stored on an external hard drive, which is password-protected and stored in a locked file when not in use. We describe details of the confidential file environment in Volume 3 of this series: Preparing Master Files [1].

RESOURCES

We have focused on basic computer techniques FHOP uses to manage its longitudinal studies. All programs described here are available on our website. A link is available on the website to join our SAS Users Group. FHOP has only two people who can provide a limited amount of handholding to learn how to use these resources. Users will have to contract for more than one hour of support.

ENDNOTES

[1] See: http://fhop.ucsf.edu/data-management-methods.

[2] See: https://notepad-plus-plus.org/