MANAGING LONGITUDINAL RESEARCH STUDIES:

DEATH STATISTICAL MASTER FILE

By

Linda L Remy, MSW PhD

Ted Clay, MS

Geraldine Oliva, MD MPH, Director
Jennifer Rienks, PhD, Associate Director
Linda L Remy, MSW PhD, Research Director

UCSF Family Health Outcomes Project
500 Parnassus Ave. Room MU-337
San Francisco, California 94143-0900
Phone: 415-476-5283
Fax: 415-476-6051
Web: http://www.ucsf.edu/fhop

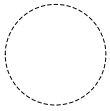
November 2018

TABLE OF CONTENTS

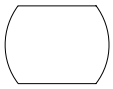
Table of Figures

Table of Tables

Table of Legends

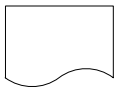


From program creating incoming file or name of next program using file

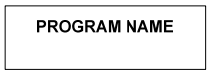


Excel file input or output

SAS file input or output



Flat or text file input or output



SAS program with brief description of steps.

Suggested Citation

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

ACRONYMS

ANSI	American National Standards Institute
CDC	Centers for Disease Control
CADPH	California Department of Public Health
COD	Underlying cause of death
CPHS	Committee for the Protection of Human Subjects
DSMF	Death Statistical Master File
FHOP	Family Health Outcomes Project
FIPS	Federal Information Processing Standards.
LST	Listing file SAS produces to show results of a program
NCHS	National Center for Health Statistics
SSN	Social Security Number
UCSF	University of California, San Francisco
WHO	World Health Organization
VSAC	Vital Statistics Advisory Committee

DEATH STATISTICAL MASTER FILES

This document describes methods the UCSF Family Health Outcomes Project (FHOP) uses to prepare confidential versions of the Death Statistical Master File (DSMF) distributed by the California Department of Public Health (CADPH).

We assume that the user of this document has read other documents describing the foundation of our methodology:

Volume One: The Basic Operating Environment

Volume Two: Standardizing Variables Over Time

Volume Three: Preparing Master Files

These and related documents are available on our website:
https://fhop.ucsf.edu/data-management-methods

OVERVIEW

Confidential versions of DSMF include: detailed demographic information about the decedent; medical data related to the vital event; and confidential personal identifiers such as names, addresses, and other fields that could identify an individual. To have these files users must obtain approval from the California Health and Human Services Agency's Committee for the Protection of Human Subjects (CPHS) and the CADPH Vital Statistics Advisory Committee (VSAC).

This document summarizes steps to import confidential DSMF into SAS, check variables longitudinally, and prepare formats. At the writing of this document, we have files from 1980 through 2017. Over this interval, the DSMF changed from version 9 to version 10 of the International Classification of Disease (ICD) to code cause of death. We discuss the impact of this change as well as various data quality issues in the DSMF. These have important ramifications for longitudinal and/or linkage-based studies using the DSMF.

Although FHOP uses confidential DSMF, processes and resources described here will help users of other population health files. We are making this basic methodology and its associated software public to help researchers understand the nature of data management for complex longitudinal research. This also will provide a background to users of our longitudinal DataBook products and readers of FHOP studies using the DSMF.

Contracts with various departments in the State of California require us to provide funding agencies with an annual backup of all programs, logs, listings, and output files. This creates an audit trail of our work. Since we do not know where the programs and/or resulting files will be used, we try to write code that will run in any environment and provide as much internal documentation as possible. All work is in SAS, assisted by Microsoft Excel and Visio. Because public funding supports development of these programs, they are in the public domain. This is why we are making them available.

STEPS TO CREATE MASTER FILES

Document Incoming Data

Incoming DSMF are stored on the Confidential Drive in password-protected ZIP files, documented in Excel files on the Master Drive, and read into SAS on the Working Drive using a macro program. Volume 1, The Basic Operating Environment, defines the Confidential, Master, and Working Environments. We described these steps in Volume 3, Preparing Master Files.

Figure 1. Document Incoming Death Statistics Master File

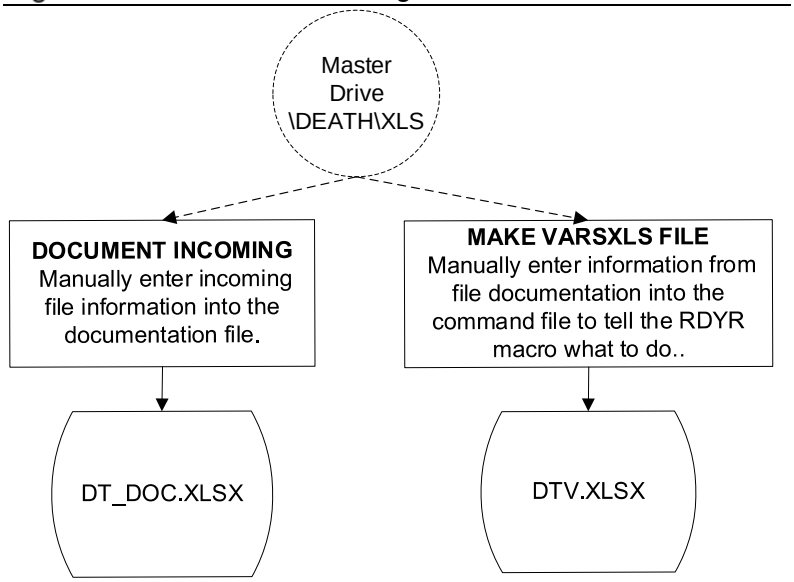


Figure 1 shows a Visio diagram summarizing the first steps to document incoming confidential DSMF into a longitudinally consistent set of files structured per FHOP standards. For details on setting up the VARSXLS file, refer to Volume Three: Preparing Master Files.

Table 1 shows a few lines of the initial DSMF documentation file (DT_DOC.XLSX). DSMF arrive as flat text files with the DAT extension or since 2014 as comma-separated with the CSV extension. Beyond the year designation, notice that names for incoming flat files and files documenting their contents differ over time. Other information we document about incoming DSMF files is the number of bytes per record (range 413 to 1527 from 2000 to 2017), whether the file includes names (D = Decedent (First, Middle, Last) S = Spouse (First, Middle, Maiden), I = Informant (First, Middle, Last), MS = Mother surname, FS = Father Surname), addresses, and

decedent Social Security Number (SSN). If the decedent used the father's surname, father's surname is blank. SSN has been available in the DSMF since 1989. This is a significant asset for longitudinal research. We do not show the notes column. It has sensitive information such as the name of the person sending the file, location of passwords, etc.

Table 1. Sample of documentation for incoming files

			OTHER INFORMATION			
Zip file	TEXT	RECORD LAYOUT SOURCE	BYTES	NAMES	ADDR	SSN
DSMF2000.ZIP	DSMF2000.DAT	2000DCVR.doc	413	D MFL	N	Y
DSMF2001.ZIP	DSMF2001.DAT	2001DCVR.doc	413	D MFL	N	Y
DSMF2002.ZIP	DSMF2002.DAT	2002DCVR.doc	413	D MFL	N	Y
DSMF2003.ZIP	DSMF2003.DAT	2003DCVR.doc	413	D MFL	N	Y
DSMF2004.ZIP	2004ConfDeath.DAT	2004DCVR.doc	413	D MFL	N	Y
DSMF2005.ZIP	DNOOS05.DAT	05DCVR-ReordLayout.pdf	413	D MFL	Y	Y
DSMF2006.ZIP	DNOOS06.DAT	06noOOSDSMF-ReordLayout.doc	550	D MFL	Y	Y
DSMF2007.ZIP	DNOOS07.DAT	07DSMFall-ReordLayout.doc	550	D MFL	Y	Y
DSMF2008.ZIP	DNOOS08.DAT	DSMF08ca_ids-Layout.doc	550	D MFL	Y	Y
DSMF2009.ZIP	DSMF09.DAT	DSMF09ca_ids-Layout.doc	550	D MFL	Y	Y
DSMF2010.ZIP	DNOOS10.DAT	DSMF10ca_ids-Layout.doc	400	D MFL	Y	Y
DSMF2011.ZIP	DSMF11_SSN.DAT	DSMF11ca_ids-Layout.doc	400	D MFL	Y	Y
DSMF2012.ZIP	DSMF12_SSN.DAT	DSMF12_ADD-Layout.doc	400	D MFL	Y	Y
DSMF2013.ZIP	DSMF13_SSN.DAT	DSMF13_ADD-Layout.doc	400	D MFL	Y	Y
DSMF2014.ZIP	CCDMF2014_Oliva.CSV	CCMDF_2014_Data Dictionary2017_102417.xlsx	1473	DSI MFL	Y	Y
DSMF2015.ZIP	2015Deaths-with-OOS.CSV	CCMDF2015_Data Dictionary.xlsx	1527	DSI MFL	Y	Y
DSMF2016.ZIP	2016Deaths-with-OOS.CSV	CCMDF2016_Data Dictionary.XLSX	1498	DSI MFL	Y	Y
DSMF2017.ZIP	Deaths2017.csv	CCMDF2017_Data Dictionary.xlsx	1462	DSI MFL	Y	Y

Define Data Structure

Table 2 shows a few lines of the Excel file defining DSMF contents (DTV.XLSX) to import the flat files into SAS. The example focuses on a few variables in the AGE set for 2005-2009, when the DSMF had the same layout. Columns not relevant for this example are closed.

Table 2. Sample of file defining structure

VARNAME	LABEL	TYPE	LENGTH	FORMAT	CODE	SASCODE	STATS	_2005_2009
BTHDATEO	Date of birth (CCYYMMDD)	C	8				M	069-076
BTHDATEC	Date of birth (CCYYMMDD)	C	8		1	if &yyyy lt 1999 and BTHDATEO gt ' ' then BTHDATEC = '1'||BTHDATEO; else BTHDATEC = BTHDATEO;	M	CALC
BTHDATE	Date of birth	N	4	date9.	4	%DATEVAR(BTHDATEC,BTHDATE,125,0);	MI	CALC
DTHDATEO	Date of death (CCYYMMDD	C	8				M	058-065
DTHDATEC	Date of death (CCYYMMDD	C	8		2	if &yyyy lt 1999 and DTHDATEO gt ' ' then DTHDATEC = '1'||DTHDATEO; else DTHDATEC = DTHDATEO;	M	CALC
DTHDATE	Date of death	N	4	date9.	5	%DATEVAR(DTHDATEC,DTHDATE,0,0);	MI	
DTHDATE	Date of death	N	4	date9.	5	%DATEVAR(DTHDATEC,DTHDATE,0,0);	MI	CALC
AGECUU	Age of decedent (CUU)	C	3				F	077-079
AGEDTH	Age at death	N	3	agedth.		%AGE(birth=BTHDATE,current=DTHDATE,age=AGEDTH);	FU	CALC
AGEDTHD	Age at death if under 3 year	N	3	agedthd.		if BTHDATE gt . and 0 le AGEDTH lt 3 then AGEDTHD = DTHDATE-BTHDATE;	FU	CALC
INJDATEO	Date of injury (CCYYMMDD	C	8					143-150
INJDATEC	Date of injury (CCYYMMDD	C	8		3	if &yyyy lt 1999 and INJDATEO gt ' ' then INJDATEC = '19'||INJDATEO; else INJDATEC = INJDATEO;	M	CALC
INJDATE	Date of injury	N	4	date9.	6	%DATEVAR(INJDATEC,INJDATE,20,0);	MD	CALC

Many longitudinal date variables arrive in different lengths, not structured to the full 8-byte length CCYYMMDD, where CC is century, YY is year, MM is month, and DD is day. DSMF dates had 7-byte structure for 1989-1998 (CYYMMDD), and the full 8-byte thereafter.

We show how we standardize dates by focusing on steps to make the decedent's birth date variable BTHDATE. We first read in BTHDATEO, the original date. Column _2005_2009 shows the incoming flat file has this is in columns 069-076. RDYR, the controlling macro that reads the data into SAS looks to the SASCODE column and carries out the instruction, to import this trimmed and left aligned. The LENGTH column shows we import it as 8 bytes long, regardless of its incoming length. The TYPE column shows we import it as a character variable (C).

Next, we calculate BTHDATEC. Column _2005_2009 contains CALC. The controlling macro to read the data into SAS looks to the SASCODE column and executes the code to structure the variable appropriately.

To make BTHDATE, column _2005_2009 again contains CALC. The controlling macro looks to the SASCODE column and calls the DATEVAR macro to structure BTHDATE as a length 4 numeric variable, formatted with date9. The OUT1 column indicates that all date variables are output. The STATS column indicates the listing file (LST) will show BTHDATEO and BTHDATEC frequencies based on our special format for missing values (M), and BTHDATE will have M frequencies plus a listing of invalid dates (I) that the DATEVAR macro corrected.

BTHDATEC_ Missing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	52	0.02	52	0.02
Present	234568	99.98	234620	100.00

BTHDATE_ Missing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	52	0.02	52	0.02
Present	234568	99.98	234620	100.00

In the DSMF, we rarely encounter invalid dates so the LST file rarely shows records where the date was changed.

We calculate two variables for age at death. AGEDTH is the decedent's age in years. AGEDTHD is age at death in days for children who die before their third birthday. These two variables have defined formats. Our policy is to give the format the same name as the variable. The LST will show standard frequency counts (PROC FREQ) and a PROC UNIVARIATE (U) for the calculated variables. The DSMF also has a 3-byte field (AGECUU) reporting age at death in years, and to days or minutes for infants.

Make the Master File

Figure 2. Prepare master files through 2013

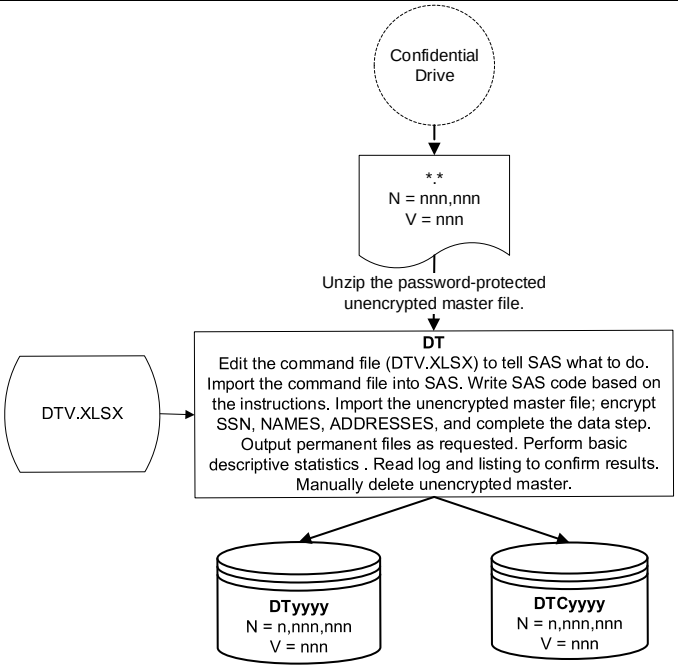
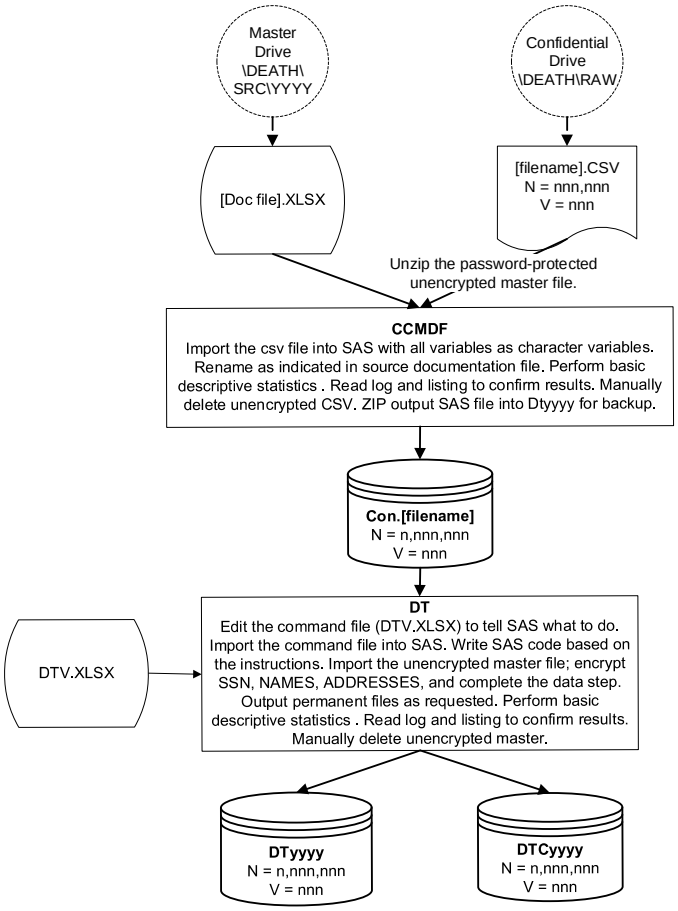


Figure 2 summarizes steps we used to standardize death files until the major structural change in 2014. The analyst modifies the program DT.SAS to add the new year of incoming data. DT.SAS calls the updated DTV.XLSX file, which has instructions for SAS about how to manage the incoming data. Instructions include the descriptive statistics to calculate in the program LST. Instructions also identify that names and addresses are output to a separate file (DTCyyyy) accessed only occasionally for specific studies.

Figure 3. Prepare master death files 2014 forward



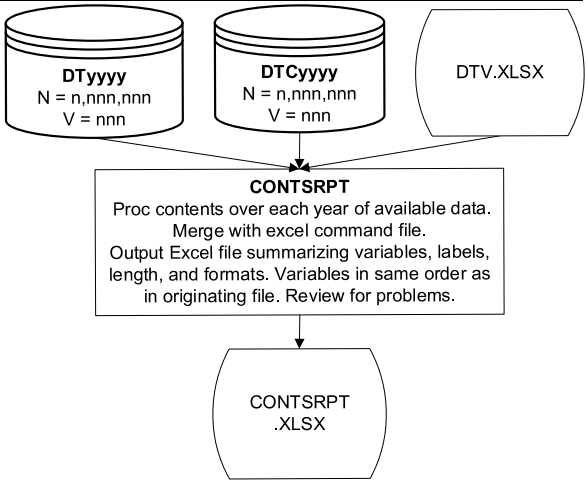
With the 2014 modification, life got more complicated. We wrote a macro program CCMDF that uses the provided excel-structured documentation and the CSV file to make a SAS file with the same variable names as in the documentation file. We obtain basic descriptive statistics from this step. Then we use the new SAS file as input to our standard DT.SAS program and proceed as before.

CHECKING LONGITUDINAL CONSISTENCY

Contents Report

After making all years of all files in a given set, we run CONTSRPT.SAS to verify that the internal structure is consistent longitudinally. This program does a PROC CONTENTS over all available years, and outputs files identifying variable names, labels, type (character or numeric), length, and formats. These are merged by year, then merged with a subset of variables (STORDER, GROUP, and VARNAME) from the VARSXLS file that created the masters. The final step outputs the information to an excel file for review. Figure 4 shows the data flow for the contents report.

Figure 4. Contents report data flow



Review of the CONTSRPT focuses on whether the same variable has the same label, type, length, and format every year it is present. Over time, some variables appear or disappear, and we check for such discontinuities. Early experience with these reports led us to develop the system we now use, so data can be consistent longitudinally. This report looked pretty sad some years ago.

Table 3 shows a cross-section of the confidential data tab (DTC) in CONTSRPT.XLS, for the DSMF decade 1991-2000. Bear in mind that we have these files from 1980 forward. Variables are grouped by STORDER, as specified in the VARSXLS file that was the source for the master files, in this case DTV.XLS. Each yearly column shows the type ((N)umeric, (C)haracter, (D)ate) and format associated with a given variable. Confidential DSMF variables do not have formats.

DSMF names have two lengths. Surnames of the decedent's mother and father have a length 15, while the decedent's last name has a length 20. Address became available in 2005. We put it in the confidential file, but keep city and ZIP in the main file. In confidential files, the E suffix indicates variables are encrypted.

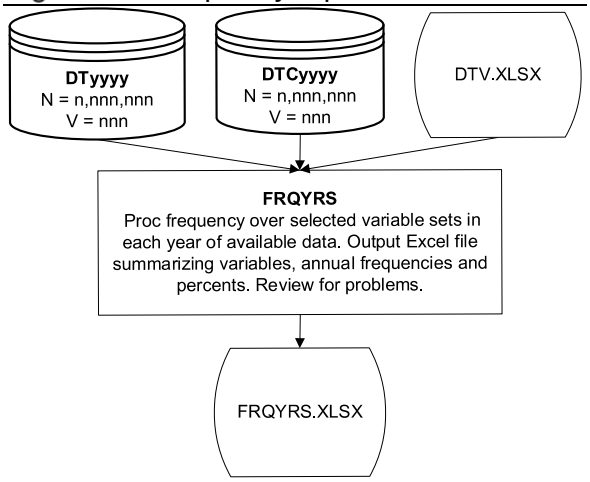
Table 3. Sample contents report of DSMF confidential variables

GROUP	NAME	LABEL	_2000	_2001	_2002	_2003	_2004	_2005	_2006	_2007	_2008	_2009
LINKAGE	YR_OBS	OBSERVATION NUMBER	n7	n7	n7	n7	n7	n7	n7	n7	n7	n7
IDENTIFY	LASTE	LAST NAME OF DECEDENT	c20	c20	c20	c20	c20	c20	c20	c20	c20	c20
IDENTIFY	FIRSTE	FIRST NAME OF DECEDENT	c12	c12	c12	c12	c12	c12	c12	c12	c12	c12
IDENTIFY	MIDDLEE	MIDDLE NAME OF DECECENT	c12	c12	c12	c12	c12	c12	c12	c12	c12	c12
IDENTIFY	FLASTE	SURNAME OF DECEDENT FATHER	c15	c15	c15	c15	c15	c15	c15	c15	c15	c15
IDENTIFY	MLASTE	BIRTH SURNAME OF DECEDENT MOTHER	c15	c15	c15	c15	c15	c15	c15	c15	c15	c15
IDENTIFY	INITSPE	INITIAL OF DECEDENT SPOUSE	c2	c2	c2	c2	c2	c2	c2	c2	c2	c2
IDENTIFY	ADDRESSE	RESIDENCE ADDRESS						c50	c50	c50	c50	c50

Frequency Reports

The program FRQYRS.SAS outputs an excel file of the same name, with tabs for categorical variables in a given group. This includes formatted continuous variables, for example, age at admission. Here, we are looking for unformatted values, or sharp changes in distributions that might indicate variables read into SAS incorrectly.

Figure 5. Frequency report data flow



FRQYRS.SAS outputs both the number of cases and the percent of cases in that year. When definitions change (for example, education (EDU to EDUN)), we look for changes in the number of cases in a given category that might indicate definitional issues.

Table 4 shows the number of cases for decedent's education variables, again focusing on the 2000-2009 cross-section. Notice that formatted values (FMTVALUE) show the underlying number. This allows programmers to make next generation variables without having to refer to the original codebook.

Table 4 also highlights how variables measuring the same construct have different values and definitions. This is an example of a variable we will have to bridge for consistency if we use it in a longitudinal study that includes the changeover years. Bridging is a method to reduce the number of categories for time trends. The purpose is to make groups comparable, across indicators and time, between numerators and denominators. Bridging continues until numerators and denominators are available over the period of interest in the current format.

Table 4. Sample frequencies report for decedent's education variables

VARIABLE	LABEL	FMTVALUE	_2000	_2001	_2002	_2003	_2004	_2005	_2006	_2007	_2008	_2009
EDU	Highest grade completed	00 No Education	6,396	6,509	6,532							
EDU	Highest grade completed	01-12 Elementary or Secondary	141,281	142,951	142,494							
EDU	Highest grade completed	13 One Yr of College	8,219	8,529	8,712							
EDU	Highest grade completed	14 Two Yr of College	25,797	26,874	27,317							
EDU	Highest grade completed	15 3 Yr of College	5,374	5,310	5,337							
EDU	Highest grade completed	16 4 Yr of College	23,322	23,867	24,016							
EDU	Highest grade completed	17 5+ Yr of College	14,584	15,030	15,407							
EDU	Highest grade completed	99 Unknown or Not Available	6,555	6,735	6,366							
EDUN	Education completed	1 < 8th grade				38,582	36,634	36,894	36,665	35,106	34,547	33,186
EDUN	Education completed	2 9-12 grade, no grad, age 9+				38,132	32,322	29,204	26,189	23,879	21,612	20,634
EDUN	Education completed	3 HS grad/GED, age 16+				74,717	76,994	80,022	83,084	83,649	85,330	83,766
EDUN	Education completed	4 Some coll, no deg, age 17+				36,465	37,509	39,127	38,777	38,206	38,426	38,654
EDUN	Education completed	5 AA deg, age 18+				9,746	9,100	9,578	10,009	10,407	10,727	10,869
EDUN	Education completed	6 Bach, age 20+				23,533	22,999	24,157	24,722	24,984	25,882	26,280
EDUN	Education completed	7 Master, age 21+				8,248	8,127	8,427	8,710	8,898	9,411	9,515
EDUN	Education completed	8 Post-grad degree, age 23+				4,237	4,627	4,940	5,192	5,501	5,581	5,841
EDUN	Education completed	9 Unknown				8,641	6,988	6,879	6,069	5,747	5,688	5,875

UNDERLYING CAUSE OF DEATH

Correctly analyzing the underlying cause of death is perhaps the most difficult issue in working with the DSMF. The cause of death is based on the International Classification of Disease

(ICD), which changes about every ten years. Correctly bridging diagnosis codes over time is a great challenge. In this section, we review some history of death registries and disease classifications, then show how we make formats to address some of the longitudinal issues.

History

Death registration began in mid-fifteenth century Italy, where medical education and social administration was more advanced than the rest of Europe at that time [1]. Italy set up boards of health to consider how to address plague pandemics that had killed more than one-third of Europe's population. A primary power was to require death registration -- including the name of the person, age, and cause of death certified by a physician -- before issuing a certificate authorizing burial. Extending from this were quarantine regulations, and powers to address the quality of food and water, and the disposal of refuse and sewage.

In 1532, London began to publish Bills of Mortality, the earliest, systematic collection of data on causes of death. These weekly lists of burials included the name of the deceased, the parish in which the burial took place, and the cause of death, with particular reference to the plague.

Diagnostic coding dates back to seventeenth-century England. Making ingenious use of the Bills of Mortality data, John Gaunt attempted to estimate the proportion of liveborn children who died, and their causes of death. In 1662, he estimated that about 36% of liveborn children died by age six [2]. Gaunt, considered the first demographer, introduced the idea that vital statistics could be used not only to keep records, but also to construct life tables for the entire population.

At its inception in 1837, the General Register Office of England and Wales appointed William Farr to head the agency. The General Register Office is responsible for vital records. Farr, considered the first medical statistician, did the first study of occupation-based mortality, first documented that male children died more often than female children, and did many of the earliest studies on health inequities [3]. He not only made the best possible use of the imperfect disease classifications available at the time, but labored to secure better classifications and international uniformity in their use.

International Classification of Disease

In 1891, the International Statistical Institute appointed a committee headed by Jacques Bertillon, Chief of Statistical Services of the City of Paris, to prepare a classification of causes of death. The resulting International Classification of Disease (ICD) was based on Farr's principle of distinguishing between general diseases and those localized to a particular organ or anatomical site [4]. In 1898, the American Public Health Association recommended that Canada, Mexico, and the United States adopt the ICD and that it be updated every ten years

[1]. The first revision (ICD1) was approved in 1899, and was used from 1900 to 1909. The first parallel classification of diseases for use in statistics of sickness also was adopted at this time.

The 1946 International Health Conference gave responsibility for maintaining international classifications of diseases and mortality to the nascent World Health Organization (WHO) in collaboration with twelve international organizations including the United States National Center for Health Statistics (NCHS). In 1948, WHO released a classification for both morbidity and mortality. This led to the International Classification of Diseases, 9th Revision (ICD-9) [4]. The United States (and California) classified deaths using this revision for the period 1979-1998.

In addition to being a classification system for causes of death, the ICD serves a number of other purposes for mortality [5]. The ICD:

- Includes rules for coding causes of death. These allow a coder to identify the single condition, the underlying cause of death, on the death certificate that is considered most informative from a public health point of view.
- Standardizes definitions such as underlying cause of death, live birth, maternal death, and many others.
- Includes tabulation lists, with recommended cause-of-death groupings, that countries use to present and compare mortality data among countries.
- Prescribes the format of the medical certification of death.
- Includes regulations on compiling and publishing statistics on diseases and causes of death for compiling mortality and morbidity statistics.

New International Classification of Disease

The great expansion in ICD use required a thorough rethinking of its structure and an effort to devise a stable and flexible classification that should not require fundamental revision for many years [4]. Effective with deaths occurring in 1999 forward, tabulating causes of death worldwide uses the tenth revision (ICD-10). The tenth revision was a drastic break from previous ICDs in a number of respects [6]:

- ICD-10 has about 8,000 categories compared with 4,000 categories in the ICD-9. The expansion mainly provides more clinical detail for morbidity applications.
- ICD-10 uses alphanumeric codes rather than numeric codes in ICD-9.
- Three chapters have been added and some chapters rearranged

- Cause-of-death titles have been changed and conditions regrouped.
- Some coding rules have been changed.

Each ICD revision creates major discontinuities in mortality trend data. The comparability ratio is used to measure the extent of the discontinuity. NCHS developed various software and other materials to permit longitudinal uses across the break between the ICD-9 and the ICD-10. Even with these resources, problems using the DSMF for longitudinal research are significant. FHOP relies extensively on these resources, and we describe later in this document how we use them.

Clinical Modification

By 1977, the ICD-9 had attained wide international recognition. This prompted the NCHS to modify the ICD-9 with clinical information. These clinical modifications provided a way to classify morbidity data for medical records, medical case reviews, and ambulatory and other medical care programs, as well as for basic health statistics. The result was the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). This version delineates the clinical picture of each patient, and provides exact information beyond that needed for statistical groupings and analysis of healthcare trends.

There are many differences between the ICD and ICD-CM even though the basic structure and design of the two classifications is similar [7]. The ICD used for mortality in the United States adheres closely to the ICD promulgated by WHO. Changes made to the ICD in the United States are generally sanctioned by WHO and occur infrequently, usually only to accommodate new medical conditions widely recognized as public health threats that are not in the existing classification. The United States implemented the ICD-10-CM in Oct-2015.

Clinical Classification Software

AHRQ is the health services research arm of the U.S. Department of Health and Human Services. Its research centers specialize in major areas of health care research such as quality improvement and patient safety, outcomes and effectiveness of care, clinical practice and technology assessment, and health care organization and delivery systems. AHRQ's Healthcare Cost and Utilization Project (HCUP) is a family of health care databases and related tools for research and decision-making [8], developed in coordination with WHO.

The most relevant are the free tools developed to use with administrative databases such as the hospital or DSMF data. These tools include Clinical Classification Software (CCS) modules for ICD-9 DX [9], ICD-9 PX [10] and PX classes [11], and the Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding (HCPC) system [12] for hospital and related services. The CCS-10 module is specific to mortality [13]. It collapses ICD-10 codes and classifies them consistently from 1999 forward.

Reducing the enormous number of codes in the various classification systems to a small number of clinically meaningful groups consistent across systems and time is its most important contribution. This "clinical grouper" makes it easier to understand illness and treatment patterns so local health jurisdictions, health plans, policy makers, and researchers can analyze costs, utilization, and outcomes associated with particular illnesses and treatments. It enables direct and uniform outcome comparisons to other regions, the state, or nation. The CCS developers engaged very high-level thinking. They brought clarity to near chaos.

ICD-9 Formats

FHOP has confidential DSMF from 1980 forward, or one year after ICD-9 started. The ICD-9 assigns a 3- to 4-digit numeric code to each underlying cause of death (COD) diagnosis (CODNINE), a 2-digit group (CODNINEG), and a 2-digit chapter (CODNINEC). CODNINE comes into SAS as a character variable, in order not to lose the leading zeros. The group aggregates roughly similar conditions to 91 classifications. For example, fourteen tuberculosis diagnoses are under group 2 (Tuberculosis), which in turn is under chapter 1 (Infectious and parasitic diseases). Chapters are organized approximately as Farr first envisioned them [1,4]. Early years of the DSMF sometimes had a dash if the fourth digit was missing. When we would attach formats, those records would show the original value rather than the formatted value. We solved this by removing the dash when we import the data into SAS, as follows:

```
if 1980 le &yyyy le 1998 then do;
   CODNINE = left(compress(CODNINE,"-"));
end;
```

The ICD-9 format has been discontinued and is no longer publicly available. The CDC provided us with a copy of the format they used for a study to evaluate ICD-9 with ICD-10 for longitudinal comparability [14]. We used this as the basis for the CODNINE format ($codnine.). As part of their standard ICD-9 documentation, CADPH provides an excel file identifying the values and labels to format CODNINEG. We make that format the same way, except that CODNINEG is numeric, so the format name is CODNINEG.

The DSMF do not arrive with chapter assigned. We made an excel file containing the chapter range values and labels for the ICD-9 and ICD-10 [15]. Table 5 shows a few lines of the chapter numbers, ICD-9 ranges assigned to those chapters, and the chapter text label.

Table 5. Layout for ICD-9 chapters

CODNINEC	FROM	THRU	LABEL
1	001	139	Infectious and parasitic diseases
2	140	239	Neoplasms
3	240	279	Endocrine, nutritional and metabolic diseases
4	280	289	Blood and blood-forming organs
5	290	319	Mental and behavioral disorders
6	320	389	Nervous system diseases system
7	390	459	Circulatory system diseases
8	460	519	Respiratory system diseases
9	520	579	Digestive system diseases
10	580	629	Genitourinary system diseases
11	630	676	Pregnancy, childbirth, puerperium
12	680	709	Skin, subcatenous tissue
13	710	739	Musculoskeletal/ connective tissue
14	740	759	Congenital anomalies
15	760	779	Perinatal period origin conditions
16	780	799	Symptoms/signs/ill-defined conditions
17	800	999	Injury and poisoning

The following is SAS code calling our MAKEFMT macro, to import this into a format library.

```
%makefmt(data = temp2,
   values    = FROM,
   values_hi = THRU,
   labels    = CODNINECL,
   library   = studylib,
   fmtname   = $CODNINC);
```

Notice in this example we use values = FROM and values_hi = THRU. ICD-9 codes between the range 001 to 139 are assigned to chapter 1 (Infectious and parasitic diseases). On the THRU side of the range, makefmt extends to include values such as 1390 or 1399. We make a character format ($codninc) because we are crosswalking from character diagnosis codes to numeric chapter codes. In assigning a format name, we remove E from CODNINEC, because format names cannot have a length greater than 8 characters and CODNINE is used. Format $CODNINC makes the variable CODNINEC, formatted with CODENINE, as follows:

```
CODNINEC = put(CODNINE,$codninc.);
format CODNINEC codninec.;
```

The following code is an example of making a label for the chapter variable:

```
CODNINECL = put(CODNINEC,codninec.);
```

After making these formats, we summarized all ICD-9 codes and groups found over the period 1980 through 1998, the last year for ICD-9, adding the number of records found for a given code in a given year and calculating a total for the number of times the code occurred over the period. We assigned the chapter, and made label variables for the ICD-9 codes, groups, and chapters. Table 6 is a snapshot of the summary listing

Table 6. Sample listing summary of ICD-9 diagnosis codes

COD	CODL	CODCL	CODGL	TOTAL	N1995	N1996	N1997	N1998
410	410 Acute myocardial infarct	07 Circulatory system d	34 Acute ischemic heart dise	375,696	18,081	17,564	17,149	17,490
411	411 Other acute and subacute forms of isc	07 Circulatory system d	34 Acute ischemic heart dise	3,281	52	61	100	86
412	412 Old myocardial infarction	07 Circulatory system d	35 Old myocardial infraction	1,637	66	57	53	83
413	413 Angina pectoris	07 Circulatory system d	36 Angina pectoris	1,036	49	48	47	46
4140	4140 Coronary atherosclerosis	07 Circulatory system d	37 Other chronic ischemic he	356,449	16,459	15,673	15,438	14,937
4141	4141 Aneurysm of heart	07 Circulatory system d	37 Other chronic ischemic he	543	14	18	21	14
4148	4148 Chronic ischemic heart disease other	07 Circulatory system d	37 Other chronic ischemic he	26,590	1,942	2,077	2,019	2,095
4149	4149 Chronic ischemic heart disease unspe	07 Circulatory system d	37 Other chronic ischemic he	139,467	10,060	10,828	11,502	11,985
4240	4240 Mitral valve disorders	07 Circulatory system d	38 Other disease endocardiu	4,201	287	227	303	282
4241	4241 Aortic valve disorders	07 Circulatory system d	38 Other disease endocardiu	17,004	1,085	1,133	1,165	1,204

NCHS published code to crosswalk ICD-9 injury classifications consistent with ICD-10 [16]. Comparability ratios for total injury, unintentional injury, suicide, and assault were close to 1.0 with standard error ranges from .0006-.0045 [15]. We incorporated the CDC SAS code, calculating injury mechanism (MECH) and intent (INTENT), labels for those variables, and from these, the formats. The following is SAS code to classify intent and mechanism:

```
length INTENT MECH 3;
MECH = put(ICDNINE,$mechnin.);
INTENT = put(ICDNINE,$intnin.);
format MECH mech. INTENT intent.;
```

We output the result of this work to an excel file for manual editing. Edited cells are highlighted with yellow. We used this excel file as input to make the final version of the ICD-9 formats. It is available upon request.

NCHS and WHO continue their research on crosswalking the ICD9 and ICD10. Other than injury, many conditions of interest for mortality surveillance have inadequate comparability ratios. For example, ambulatory care sensitive conditions such as bronchiolitis, pneumonia, and influenza have comparability ratios ranging from .39 to .74. On the other hand, malignant neoplasm deaths were stable across ICD9 and ICD10. Thus, before proceeding with longitudinal research to bridge causes of death other than injury or malignant neoplasms, it is important to evaluate published comparability tables and make the appropriate adjustments to smooth the bridging period.

ICD-10 Formats

We use the list of ICD10 codes and labels that NCHS publishes to make formats to label cause of death ($CODTEN) [17]. Vital Statistics provides the list of groups as part of its standard documentation. We use this to create a crosswalk from CODTEN to the group ($CODTENG) and to make the group format (CODTENG). The format crosswalking CODTEN to the chapter ($CODTENC) is made the same way we made $CODNINC [15]. Finally, we make cross-classification summaries of CODTEN, CODTENG, CODTENC and use this to assign injury classifications MECH and INTENT adapting SAS code from NCHS [18]. We export this to excel, hand edit as needed, and use the Excel file as input to make formats ($MECHTEN, $INTTEN) to classify ICD10 injuries.

CCS has not updated ICD-10 classifications for several years, and new codes have been added in the interval. We applied the CCS format information we had to the summary list of ICD-10 cause of death, and output them in an excel file with the full list of COD, injury classifications, and CCS classifications. We edited the excel file and then used it as the input file to make the final formats.

There are important differences between the group variable in the DSMF and the CCS. For example, the CCS classifies A162 to A199 in CCS group 1 (Tuberculosis). The ICD-10 assigns A162 to A169 to COD group 9 (Respiratory tuberculosis) and A170 to A199 to COD group 10 (Other tuberculosis). Thus, the purpose of the analysis should determine whether to use the CCS or the COD group. In linking patient data with the DSMF, we recommend using CCS, as codes will be grouped similarly regardless of differences in underlying coding system. Table 7 shows a few rows of the COD10.XLSX file. Columns show COD and label (COD(L)), chapter and label (CODC(L)), group(CODG(L)), and CCS sub-grouper and label (CODCCSL(L)), the total for all years and the number in each year.

Table 7. Sample listing summary of ICD-10 cause of death

COD	CODL	CODC	CODCL	CODG	CODGL	CODCCSL	CODCCSLL	TOTAL	N2013	N2014	N2015	N2016	N2017
I10	I10 Essential (primary) hyperte	9	09 Circulat	161	161 Essential	98	098 Essential h	42583	3028	2971	3363	3290	3600
I110	I110 Hypertensive heart disease	9	09 Circulat	162	162 Hyperter	99	099 Hypertens	30394	1667	1548	1882	2061	2351
I119	I119 Hypertensive heart disease	9	09 Circulat	162	162 Hyperter	99	099 Hypertens	47335	2592	2616	2948	2714	2926
I120	I120 Hypertensive renal disease	9	09 Circulat	163	163 Hyperter	99	099 Hypertens	25000	1669	1594	1739	1791	1962
I129	I129 Hypertensive renal disease	9	09 Circulat	163	163 Hyperter	99	099 Hypertens	727	40	24	50	56	70
I130	I130 Hypertensive heart and re	9	09 Circulat	164	164 Hyperter	99	099 Hypertens	349	22	29	23	40	50
I131	I131 Hypertensive heart and re	9	09 Circulat	164	164 Hyperter	99	099 Hypertens	3360	215	235	246	250	292
I132	I132 Hypertensive heart and re	9	09 Circulat	164	164 Hyperter	99	099 Hypertens	6104	378	366	404	447	478
I139	I139 Hypertensive heart and re	9	09 Circulat	164	164 Hyperter	99	099 Hypertens	277	14	11	15	21	19

GEOGRAPHIC CLASSIFICATION

DSMF has a number of geography-associated variables. Nation, state, and county-level examples include where the decedent was born, lived, died, and health jurisdiction filing the record of the vital event. Sub-county variables include geographic characteristics such as city of residence, addresses, ZIP-codes, and Census Place and Tract codes.

We introduce some longitudinal issues for DSMF nation, state, county, and city variables. Detailed information related to these issues and other geography-related variables are in our Geography Master document.

DSMF Nation, State, and County Variables

The fundamental problem with this group of variables is that different codes describe the same geographic entity across variables at the same time and over time. For example, the code for Hawaii is variously HI, 112, and 12. The code for Humboldt County is 12 and 012.

Before the national 2003 DSMF revision, NCHS used unique geographic codes [19]. With the 2003 revision, they adopted Federal FIPS codes [20]. Then the US Census converted from FIPS to ANSI codes for the 2010 Census [21]. We addressed these various longitudinal changes by creating a crosswalk from older geographic variables to 2010 Census standards ($csn.). This is useful for consistently assigning geographic variables over time.

People from different regions of the world have different kinds of health problems. We also developed a format to crosswalk nation of birth to larger regions ($marc).

We discuss development of the DSMF region, nation, and state crosswalks in the document about our Geography Master. These formats enable merging DSMF data with census data.

City Variables

The DSMF city variable is released as written because it appears on the vital records certificate. Thousands of typographic errors exist, which increases geocoding and mapping problems. For example, we found 19 incorrect spellings for Los Angeles and 12 for San Francisco. As part of the ongoing maintenance of our geography master, we summarize all city names found in various files including commercial vendors. We manually correct misspellings and convert them to a format.

Format $CITYC. replaces the incorrectly spelled city with the correct name. If we lack a city name but have a ZIP, we impute city with the format $ZIPCITY. Note our format naming convention: left side of the format name is the variable we have, right side of the format name is the variable we are imputing. The following is code we use to make for the DSMF:

```
* no ZIP before 1989;
if 1989 le &yyyy then do;
   if CITYO = ' ' then CITY = put(DZIPC5,$zipcity.);
   else if CITYO gt ' ' then CITY = put(CITYO,$cityc.);
end;
* Pick up new city spelling errors;
if CITYO gt ' ' and CITY = ' ' then CITY = CITYO;
```

The DSMF collected the 4-byte Census Place code from 1986-2002, and the 5-byte FIPS place code from 2003 forward. We do not use variables for Census Place or Tract. Regardless of the version, the same place code refers to multiple cities. Tract is an optional field with most missing. We have done a number of analyses on relationships between ZIP, City, County, Place, and Tract. Geography-related errors are pervasive in the DSMF, which was part of the motive for developing our geography master series. VS knows of the problems and has undertaken an agency-wide initiative to improve geocoding. In the meantime, ZIP-code has been available since 1989, and while still problematic, returns a better city name.

DATA QUALITY

Data quality issues plague DSMF across the nation. Death certificates are filed as recorded, and reabstraction studies are exceedingly rare. Important known issues are associated with cause of death and geography assignment that impact research methodology.

Cause of Death

Death certificates record the proximal cause of death and greatly underestimate certain conditions as a cause of death. For example, Hoel and colleagues reported that the total death certificate error rate varies considerably by cancer type, time period, and age at death, returning a consistent 18% under-estimation of cancer deaths [22]. Johansson and Westerling found 83% agreement on cause of death between hospital records and death certificates for patients dying in-hospital, but only 47% agreement when the patient died after discharge [23]. Dramatic conditions had the highest agreement. They recommended routine linkage of death and hospital records to improve quality of death records.

In FHOP's study of injured California youth age 10 to 24 [24], we found 10,628 youth in the DSMF where 5,357 reportedly died in a California facility. Of these, 5,336 were recorded as dying in a state-licensed hospital, 21 in all other licensed facilities, and 61 dead on arrival at the hospital. Once having arrived, the DSMF showed 1,972 individuals died as inpatients, 3,350 as outpatients or in the emergency room, 54 dead on arrival, with 15 unknown as to where they died in the facility. However, the discharge data had 2,346 inpatient deaths for this population, a significant discrepancy. We strongly recommended a reabstraction study focusing on inpatient injury deaths and injury data in the DSMF. No such study has been done.

In diagnostic analyses of the DSMF for 1999-2009, we found 145 cases where age at death ranged from age 4 to 79 years for the following newborn or perinatal cause of death groups:

274	Newborn affected by maternal factors and by complications of pregnancy, labor and delivery
275	Disorders related to short gestation and low birth weight, not elsewhere classified
276	Birth trauma
277	Intrauterine hypoxia and birth asphyxia
278	Respiratory distress of newborn
279	Other respiratory conditions originating in the perinatal period
280	Infections specific to the perinatal period
281	Other and unspecified conditions originating in the perinatal period

Geographic Variables

CADPH commissioned FHOP to do a geocoding case study of DSMF in one large California County. The study focused on geographic data quality in 2005 and 2007, before and after CADPH implemented the Electronic Death Registration System [25]. CADPH selected the

county because it was thought to have high quality data. The analysis focused on in-county deaths to county residents, where the local jurisdiction has full control of the recording process. Many DSMF records had a city or alternate or a ZIP or alternate but not both, and we had to remove 465 records obviously not in the county. We encountered thousands of data quality problems. Many other data quality problems in the county's geographic data system exacerbated the geocoding results. We made a number of recommendations intended to improve the quality of geographic variables in the DSMF.

A more recent study focused on Willits in Mendocino County [26]. Part of the analysis examined the relationship between Census place, ZIP, and county in the DSMF during the late 1980s and early 1990s. The place code for Willits (3120) was assigned to ZIP 94940 (Marshall, Marin County), 95428 (Covelo, Mendocino County), 95459 (Manchester, Mendocino County), 95470 (Redwood Valley, Mendocino County), and 95490 (Willits, Mendocino County). The Willits ZIP (95490) was assigned to Lake, Solano, and Mendocino counties, and to the places Clearlake Oaks (Lake County), Covelo (Mendocino County), Vallejo (Solano County), and Willits. Mendocino County was assigned to 54 ZIPS. These included 95446 (Guerneville, Sonoma County), 95485 (Upper Lake, Lake County), 95565 (Scotia, Humboldt County), 95954 (Magalia, Butte County), 94940 (Marshall, Marin County).

RESOURCES

We have focused on the preparation of DSMF for analysis per FHOP standards and highlighted some limitations of this important data resource. Approach work with the DSMF cautiously. Understand the limitations of recorded cause of death. Review published studies to understand known data quality problems. Be aware of the limitations of geographic data for small area studies. Do not push the design, analysis, or interpretation of DSMF data beyond what is supportable given known data quality problems.

All programs are available upon request. The current format library for California deaths is available on our website (https://fhop.ucsf.edu/data-management-methods). (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 Moriyama IM, Loy RM, Robb-Smith AHT. (2011) History of the statistical classification of diseases and causes of death. Rosenberg HM, Hoyert DL, eds. Hyattsville, MD: National Center for Health Statistics. Last accessed 01-Jul-2011 at: http://www.cdc.gov/nchs/data/misc/classification_diseases2011.pdf.

2 Graunt J. (1662) Natural and political observations in a following index and made upon the bills of mortality. London. 1662. Last accessed 01-Jul-2011 at: http://www.edstephan.org/Graunt/graunt.html

3 Whitehead M. (2000) William Farr's legacy to the study of inequalities in health. Bulletin of the World health Organization, 78(1). Last accessed 02-Jul-2011 at: http://www.who.int/bulletin/archives/78%281%2986.pdf.

4 World Health Organization (2004). History of the development of the ICD. Last accessed 02-July-2011 at: http://www.who.int/entity/classifications/icd/en/HistoryOfICD.pdf. .

5 New International Classification of Diseases (ICD-10): The History and Impact. Health Statistics Section, Colorado Department of Public Health and Environment. March 2001, No. 41. Last accessed 02-July-2011 at: http://www.cdphe.state.co.us/hs/briefs/icd10brief.pdf.

6 National Center for Health Statistics (2000). A guide to state implementation of ICD-10 for mortality. Part II: Applying Comparability Ratios. National Center for Health Statistics, Centers for Disease Control and Prevention, December 4, 2000. Last accessed 02-Jul-2011 at: http://www.cdc.gov/nchs/data/statab/Document%20for%20the%20States.pdf.

7 National Center for Health Statistics (1998). A guide to state implementation of ICD-10 for mortality. National Center for Health Statistics, Centers for Disease Control and Prevention. 16-Jul-1998. Last accessed 02-Jul-2011 at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD-9_10Con/let2.doc.

8 Databases and Related Tools from the Healthcare Cost and Utilization Project (HCUP) (2011). Last accessed 01-Jun-2011 at: http://www.ahrq.gov/data/hcup/datahcup.htm#Tools.

9 http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

10 http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp.

11 http://www.hcup-us.ahrq.gov/toolssoftware/procedure/procedure.jsp.

12 http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp

13 Clinical Classifications Software for ICD-10 Data: 2003 Software and User's Guide. January 2003. Agency for Healthcare Research and Quality, Rockville, MD. Downloaded 14-Mar-2007 at: http://www.ahrq.gov/data/hcup/icd10usrgd.htm. CCS labels for icd10: http://www.hcup-us.ahrq.gov/toolssoftware/icd_10/ccs_icd_10.jsp. Checked web 11-Jul-2011. File downloaded in 2007 is still the most recent.

14 David P. Johnson (2011) Email to Linda Remy 21-Jul-2011;

15 Anderson RN, Miniño AM, Hoyert DL, Rosenberg HM (2001). Comparability of cause of death between ICD–9 and ICD–10: Preliminary estimates. National Vital Statistics Reports; 49(2). Hyattsville, Maryland: National Center for Health Statistics. 2001. Last accessed 02-Jul-2011 at: http://www.cdc.gov/nchs/data/nvsr/nvsr49/nvsr49_02.pdf. D:\DEATH\ENDNOTE\ICD0910\ANDERSON_2001_COMP_ICD9_ICD10_RATIOSnvsr49_02.pdf

16 Miniño AM, Anderson RN, Fingerhut LA, Boudreault MA, Warner M. Deaths: Injuries, 2002. National vital statistics reports; vol 54 no 10. Hyattsville, Maryland: National Center for Health Statistics. 2006. Last accessed 24-Jun-2011 at: http://www.cdc.gov/nchs/data/nvsr/nvsr54/nvsr54_10.pdf.

17 International Classification of Diseases, Tenth Revision (ICD-10). Last accessed 23-Jun-2011 at ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10/. See particularly ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10/allvalid2009(detailed%20titles%20headings).xls E:\DEATH\SRC\ICD0910|

18 See: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/injury/sascodes/icd10_external.sas. Last accessed 24-Jun-2011.

19 National Center for Health Statistics (2003) NCHS Geographic Coding 2003. Division of Vital Statistics, CDC/NCHS. Last accessed 13-Jun-2011 at: http://www.cdc.gov/nchs/data/dvs/Geocode_pres.pdf. .

20 National Center for Health Statistics (2004). Specifications for collecting and editing the United States standard certificates of birth and death and the report of fetal death -- 2003 Revision. Dated Apr-2004. Updated 18-Mar-2005. Last accessed 12-Jun-2011 at: http://www.cdc.gov/nchs/data/dvs/FinalBirthSpecs3-24-2005.pdf. See also: http://www.cdc.gov/nchs/nvss/vital_certificate_revisions.htm.

21 American National Standards Institute (ANSI) Codes (2010). US Census Bureau. Last accessed 13-Jun-2011 at: http://www.census.gov/geo/www/ansi/ansi.html. Site last revised 27-Dec-2010.

22 Hoel DG, Ron E, Carter R, Mabuchi K: Influence of death certificate errors on cancer mortality trends. J Natl Cancer Inst. 1993 Jul 7;85(13):1063-8.

23 Johansson LA, Westerling R: Comparing Swedish hospital discharge records with death certificates: implications for mortality statistics. Int J Epidemiol. 2000 Jun;29(3):495-502.

24 Oliva G, Remy L, Clay T. (Aug 2000). The California Child and Youth Injury Hot Spot Project Report for the Period 1995 to 1997. Sacramento, CA: California Department of Health Services. See: http://fhop.ucsf.edu/fhop/htm/publications/index.htm.

25 Remy L (2009). A geocoding case study: California Death Certificates 2005 and 2007. Funding by the California Department of Public Health, Center for Health Statistics under CDPH Contract No. 07-65212.

26 Remy L (2011). Harry Whitlock, et al, Plaintiffs, v. Pepsi Americas, et al., Case No.: C-08-2742 SI. United States District Court, Northern District of California.