

MANAGING LONGITUDINAL RESEARCH STUDIES:

STANDARDIZING VARIABLES OVER TIME

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/>

March 2021

TABLE OF CONTENTS

Common Issues	1
Time Variables.....	2
Dates	3
Age	4
Day	5
Cohort.....	6
Demographic Variables	10
Sex.....	10
Race/Ethnicity	10
Health Insurance	13
Confidential Data Elements.....	14
Social Security Number.....	14
Medical Record Number.....	15
Record Linkage Number	15
Name	15
Resources	16
Endnotes	17

TABLE OF TABLES

Table 1.	Population Data Files	2
Table 2.	Available numerator and denominator options for race/ethnic reporting	12

TABLE OF FIGURES

Figure 1.	Time, place, events, and age location in a community history	7
Figure 2.	Schaie's sequential designs adapted by Hageaars.....	9

Suggested Citation

Remy L, Clay T. (2021) Managing Longitudinal Research Studies: Standardizing Variables Over Time. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: <http://fhop.ucsf.edu/data-management-methods>.

ACRONYMS

ACCLAIMS	Automated Certification and Licensing Administrative Information and Management System
AHDR	Annual Hospital Disclosure Report
AIAN	American Indian/American Native
API	Asian/Pacific Islander
ASC	Ambulatory Surgery Center Data
BSMF	Birth Statistical Master File
CDC	Centers for Disease Control
CHS	Center for Health Statistics
DOF	Department of Finance
DSMF	Death Statistical Master File
DX	Diagnosis
ED	Emergency Department Data
FDSMF	Fetal Death Statistical Master File
FHOP	Family Health Outcomes Project
LTC	Long Term Care Financial Data
NCHS	National Center for Health Statistics
NHOPI	Native Hawaiian/Other Pacific Islander
OOR	Out-of range or fully missing
PDD	Patient Discharge Data
OMB	Office of Management and Budget
OSHPD	Office of Statewide Health Planning and Development
PX	Procedure
UCSF	University of California, San Francisco

STANDARDIZING VARIABLES OVER TIME

This is the second in a series of documents describing basic methods the Family Health Outcomes Project (FHOP) uses to manage its longitudinal data [1]. Analysts in local health jurisdictions and researchers interested in longitudinal research may find this and the related volumes helpful.

Here we introduce both a general philosophy and methods to maintain the longitudinal integrity of variables in population-based files, within and across datasets. We first describe general issues across datasets for core variable sets. Then we introduce a few SAS macros we developed to standardize variables over time. Data elements addressed here are related to time, demographic, and confidential variables.

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. This also should provide a background to users of our longitudinal DataBook products. We hope this will help people better understand how we preprocess master files to make DataBooks and do our longitudinal research studies.

We learned the hard way that these methods enable us to produce more work, be more confident that the work we produce is accurate, and do more challenging studies with less staff than otherwise would be possible. These methods produce master files with contents that are consistent within and across datasets, over time, and that address source file idiosyncrasies and content changes. All work is in SAS, helped by Microsoft Excel and Visio.

COMMON ISSUES

We start by addressing common issues across multiple datasets and identify some macros we use to create certain types of variables. Anyone who downloaded the file TOOLS.ZIP from our website already has the macros we discuss [2].

Researchers who work with longitudinal population-based administrative datasets (vital statistics (birth (BSMF), death (DSMF), fetal death (FDSMF) statistical master files), patient abstracts (hospital discharge (PDD), ambulatory surgery center (ASC), emergency department (ED)), claims (Medicare, MediCal), disclosure reports (hospital, long term care, patient safety)) face common issues that fundamentally influence research methodology. Maintaining consistency over time, within and across datasets, is the most basic and crucial issue facing the integrity of longitudinal research. Master source file problems arise in three major areas:

- **File format.** Some datasets arrive as text files or Excel files. Others arrive as SAS or SPSS files. Because incoming data differs from year to year, we developed methods to standardize these differences longitudinally when we read data into SAS. Other documents in this series describe how we handle changing incoming data [1].
- **Variable definition.** Over time, variables are added to or removed from datasets. The same content may arrive with different variable names. Variables with the same names may contain different definitions. Content may change, as new categories are added or subtracted. Variables with the same content may be defined as character in one year and numeric in another.
- **Confidentiality.** Providers of each dataset may have different confidentiality requirements. Different datasets may include different types of confidential variables. These variables may be present in some years and absent in others.

Table 1 identifies datasets covered by this document, the acronym we use, and the years we have available as we are writing this.

Table 1. Population Data Files

Agency	Data Type	From	To
<i>California Department of Health and Human Services</i> Office of Statewide Health Planning and Development (OSHDP)	Patient Discharge Data (PDD)	1983	2019
	Emergency Department Data (ED)	2005	2019
	Ambulatory Surgery Center Data (ASC)	2005	2019
	Annual Hospital Disclosure Report (AHDR)	YR08	YR44
	Long Term Care Financial Data (LTC)	1983	2019
<i>California Department of Public Health (CDPH)</i> Vital Records	Death Statistical Master File (DSMF)	1980	2019
	Birth Statistical Master File (BSMF)	1983	2019
	Fetal Death Statistical Master File (FDSMF)	1983	2019
Licensing and Certification	Automated Certification and Licensing Administrative Information and Management System (ACLAIMS)	1986	2003
<i>Department of Finance (DOF)</i>	County-level population files	1975	2021
<i>US Census Bureau</i>	Population and population characteristics at the following levels: County, ZIP/ZCTA, Census Tract, Group, Block	1970	2020
<i>National Center for Health Statistics (NCHS)</i>	Population with bridged race, ethnicity, sex, and continuous age at the County level	1990	2019

TIME VARIABLES

"Tomorrow, and tomorrow, and tomorrow, creeps in this petty pace from day to day [3]."

Time is the constant companion of longitudinal researchers. We count days from last menstruation to birth, last live delivery, last mammogram or prostate screening; from hospital admission to discharge, discharge to readmission, discharge to death, injury to death; age at admission, at placement, at incarceration, at discharge, at death; disability-adjusted life years, years of life lost; day, hour and minute of birth or death.

Longitudinal research uses time to evaluate the population effect of major life events or changes in public policy and tries to grasp their meaning and impact. On this existential theme, this section summarizes methods FHOP uses to manage time-related variables, the heart of longitudinal research.

Dates

Most of FHOP's date variable names have two parts, prefix and suffix. The prefix signifies the major life event the date commemorates: birth (BTH), death (DTH), admission (ADM), discharge (DIS), injury (INJ), procedure (PX). An exception is in the BSMF and FDSMF. There we prefix variables pertaining to the infant with I, birth mother with M, and birth father with F (e.g., IBTHDATE, MBTHDATE, FBTHDATE). Other date variables suggest longer intervals, with names such as FROM, THRU, or YEAR.

Date suffixes vary depending on internal structure. All numeric date variables have the name suffix DATE or DT or YEAR. Some files arrive with date variables as character strings. In this case, the variable name includes a second suffix C, e.g., BTHDATEC. Some date variables add numeric suffixes. For example, patient abstracts can contain up to 20 procedure (PX) dates. The variable PXDT00 identifies the principal procedure date.

We use the DATE9 (DDMMMYYYY) format on numeric SAS date variables. We have found that other formats can cause needless confusion. For example, is the string 09/10/99 interpreted as September 10 or October 9? Does 99 mean year is missing or does it mean 1999? The DATE9 format unambiguously returns 10-Sep-1999.

Over time, within and across datasets, character date variables arrive with different structures. The following are a few examples of internal structure for character date variables: MMY, MMDDYY, CYYMMDD, CCYYMMDD, where MM = month, DD = day, YY = year, and C = century. Our task is to give all character date variables a standard internal structure (CCYYMMDD). We then convert that standardized string to a numeric SAS date.

- **Convert character date string to standard character structure.** When date variables do not have the standard internal character string structure, the macro FIXDATE is invoked. FIXDATE requires the name of the input character string variable, and the incoming pattern. The following is an example to convert incoming original character birthdate (BTHDATEO) to outgoing standard character structure (BTHDATEC), with the macro replacing the input variable contents.

```
%FIXDATE(var = BTHDATEO, out = BTHDATEC, pattern = CYYMMDD);
```

Character date variables arrive with different lengths. Before invoking FIXDATE, we set the outgoing variable to character, length 8 (\$8.). Labels for incoming variables are generic because their internal characteristics can differ from year to year, while labels for outgoing date variables unambiguously describe the final standardized structure:

```
label  
BTHDATEO = 'Birthdate (Orig)'  
BTHDATEC = 'Birthdate (CCYYMMDD)';
```

- **Convert standardized character date to numeric date.** In creating numeric date variables, the macro DATEVAR incorporates decision rules developed by the US Division of Vital Statistics, NCHS, CDC [4,5]. It requires the name of the incoming

character string variable, the name of the outgoing numeric date variable, the maximum number of prior years that will enable the date to be valid, and the minimum number of years to be valid.

```
%DATEVAR(BTHDATEC, BTHDATE, 125, 0);  
label BTHDATE = 'Birthdate';
```

In evaluating dates, it is important to know that parts can be out-of-range (OOR). For example, months should be in the range 01 to 12 and days in the range 01 to 28, 30, or 31. Date parts with values such as 00, 01, --, or fully missing (blank) are always OOR. Month values in the range 13 to 99 are always OOR, while OOR day values are contingent on month (30 days hath September, April, June, and November, etc.). Per Federal standards, we assign the value 15 when days are OOR, and the value 06 when months are OOR [4]. If both month and day are OOR, month is given the value 07 and day the value 01.

Between 1980 and 2009, California registered only one death for a person age 123 years. For birth dates, we set the maximum allowable prior years at 125. Any birth year more than 125 years before the current year is set to missing. For example, if someone discharged from hospital in 2005 had a recorded birth year of 1874, returning an age of 131 years, numeric birth date and calculated age is set to missing. The minimum value (0) indicates that we will not accept dates with year greater than the file year, for example, 2005.

Some masters arrive as SAS files, with character dates absent and numeric dates already calculated. These dates are sometimes missing and we do not know why. One possibility is that the program that read the data into SAS did not include date element checks, and when elements were invalid, SAS automatically returned the date as missing. Thus, wherever possible, read date variables into SAS first as character strings, check date elements for validity, and, if needed, correct before converting to SAS date variables.

The California Department of Vital Statistics validates infant birth dates in birth certificates and death dates in death certificates [6]. Other dates are more problematic. For example, Aunt Mary's niece may recall that she was born a few days before the San Francisco earthquake (1906), but because of her advanced age when she dies, may not know the exact date.

Assigning values to OOR date variable parts allows us to keep a "date" for the record. When we link records temporally using other information, this may result in an out-of-sequence record. However, where we had date parts, instead of missing the entire date, we probably will be able to link the record correctly in the end, using other information on the partially dated record plus data on related records. Without date parts, we would lose all information for the variable and possibly the entire record.

Age

If we have date variables, we calculate our own age variables invoking the macro AGE.SAS, which uses the floor function [7]. For example, subtracting the date of interest (death) from

the date of birth might result in an age of 18 years, 11 months, and 25 days. Using the floor function, the patient would be age 18 until she was 19 years, 0 months, 0 days.

```
%AGE(BTHDATE, DTHDATE, AGEDTH)
```

In the case of Aunt Mary, we only know she was born in 1906. We earlier filled in July 1 to estimate her birthdate. If she died before July 1 in 2005, her age would be 98 years.

Age variables have the prefix AGE, and a suffix associated with the event (e.g., BTH, INJ, ADM, DIS, DTH). For example, AGEADM is age at hospital admission and AGEDIS is age at discharge. Again, we have exceptions to this rule in BSMF and FDSMF, where we prefix age variables for mother and father (e.g., MAGE and FAGE).

Sometimes discrepancies occur between reported and calculated age. The CDC/NCHS has decision rules for when to use calculated or reported age for vital statistics indicators, and when to impute if age is missing or implausible. For example, we would not trust the reported or calculated age of a 15-year old mother with nine prior live births. The following macro will evaluate mother's calculated and reported age and decide which to use:

```
%MAGEIMP(MAGE, MAGECALC, MAGECOMP, MAGEIMP)
```

Day

If the person is younger than three years of age, we calculate selected variables in days. AGEADM is age at admission in days for a child younger than 3 years. Newborns are age 0 days on their date of birth, and at least 1 day old on discharge. These variables are useful for monitoring, for example, perinatal deaths among infants less than 28 days of age. Another example is AGEINJD, days from injury to death, which conveys another sorrow.

Length of hospital stay (LOS) is calculated by subtracting DISDATE from ADMDATE. When a patient enters and leaves the hospital on the same day ("drive-thru" delivery or mastectomy), the result is a value of zero. From the view of the patient and family, the patient spent a day in hospital. In analyses, we adjust LOS by 1 day when the value is 0. Failing to increment LOS when it is zero will undercount total days of hospitalization when LOS is summarized to the facility level to calculate something like Average Length of Stay (ALOS) [8]. On the other hand, in the DSMF, we do not adjust days from injury to death, as someone injured could die before reaching the hospital. Thus, adjusting days depends on the purposes of the longitudinal analyses.

"Days to", another class of day variable, calculate time between one event and another, for example, from admission to the date a given procedure was performed. Sometimes calculating "days to" requires data to be sorted by a unique identifier ("person", "hospital"), date, and perhaps other variables such as disposition (transferred, died). This is an example of a "linked" analysis, perhaps calculating days from admission to discharge based on sequential records to establish an episode of care. To set up these analyses in advance, we presort master datasets at creation by a unique identifier, for example, encrypted social

security number, and other time variables, for example, admission date, discharge date, and disposition, where transfer to another hospital precedes death. Then we use the lag function to calculate "days to".

Another example is days from hospital closure to reopening, using the AHDR or LTC datasets. Here the question of interest might be the length of time during which no facilities are available in a given county to care for the pediatric or mentally ill population, which could result in sharp increases in out-of-county care. In the ACLAIMS database, it might reflect days between patient complaints or findings of violations of patient safety laws in a given hospital, where clustering of days within a short period might reflect serious patient safety conditions. To facilitate these analyses, we presort datasets at creation by facility identifier and various dates and/or temporally sequential outcome variables. When we are doing analyses, we use the lag function to calculate elapsed days.

Cohort

A growing literature reflects the interplay between time, place, events, and health. Research on the long-term impact of exposure to various environmental risks is increasingly well developed [9-12].

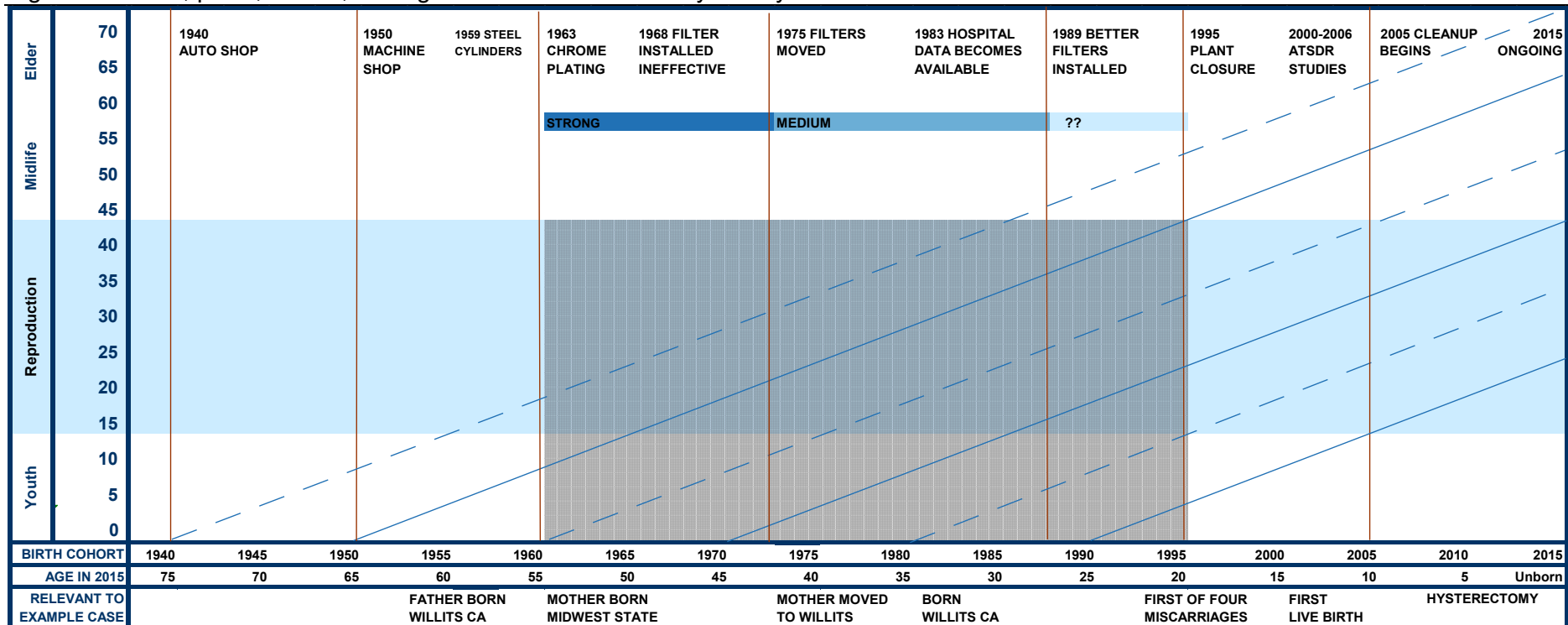
Strauss and Howe were among the first to define, locate, and name the sequence of American generations and describe how major events of each generation's time shape them [13]. FHOP adapted their timeline model to visualize longitudinal relationships between birth cohorts, developmental stage (age), and life events, for example, changes in health insurance coverage eligibility [14].

Figure 1. shows relationships between time (horizontal bar), age and developmental stage (vertical bar), and generation (diagonal bar) for Willits, California, a rural community exposed to hexavalent chromium [15,16]. Willits had one of California's most stable populations. Federal investigators determined that exposure had been enough to impact population health [17-22].

Willits is an example of an enduring industrial presence, with gradual changes in potential health impact. Vertical lines highlight key events in the exposure history. In 1950, owners converted a small auto shop in Willits to a machine shop and introduced chemicals to harden metals. By 1960, they began chrome plating and expanded for large-scale production of consumer and military products. Efforts to contain emissions were unsuccessful until 1990. Ownership changed several times over the decades, and with the winding down of military expenditures (the "peace dividend"), the plant closed in 1995 after years of investigations.

Figure 1. shows that cohorts born 1940 through 1995 were exposed, particularly during the reproductive period and childhood (grayed area). Given the history of changing exposure risk, our research goal was to assess if the Plant negatively affected community health, and if risk was constant across cohorts.

Figure 1. Time, place, events, and age location in a community history



Solid diagonal lines signify 20-year generations; dashed lines signify 10-year cohorts. The first horizontal bar at the bottom shows successive 5-year periods when people were born. The age bar shows how old someone would be at period end, relative to 2015. The 1950 cohort is just beginning to enter Elderhood, when cancers begin to manifest.

The bottom bar summarizes relevant events for one plaintiff in now-ended litigation. Born in 1981, Diane Doe (name changed to protect identity) was 34 years old in 2015. She is a member of the 1980 cohort, a potential bearer of the long-term effects of particular events [13]. Similarly, her parents (1950 cohort) bear long-term effects of events that occurred to them at developmentally specific periods. Depending on when someone moved to Willits, diagonal lines in Figure 1 show that the 1950-1969 “parent” generation has a different risk exposure history than Diane’s 1970-1989 “child” generation. The story of Diane’s family parallels the Plant and helps to understand generational effects.

Diane’s grandfather worked in the Plant, and conceived her father in Willits. Her father, born 1957 and exposed from childhood forward, lived in Willits his entire life. Her mother, born 1959 in a mid-west state, moved to Willits at age 18, a few years before Diane’s birth. Her mother was unexposed during infancy, childhood, or adolescence, but was exposed during her reproductive period. A brother born a year before Diane has significant congenital birth defects. After separating from Diane’s father in 2000, her mother moved elsewhere in the County.

Diane lived in Willits through the developmentally vulnerable stages of infancy, childhood, and early adolescence. When the plant closed in late 1995, Diane was about 14 years old, just entering her reproductive period. She remained in Willits until 1999, the early years of conceiving yet another vulnerable generation. She had three miscarriages, followed by a period of infertility, followed by four difficult pregnancies where she was able to deliver four live infants, each preterm and/or low birthweight. In 2008, at about age 29, physicians surgically removed her uterus because of pre-cancerous tumors and they removed polyps from her colon in 2010. She still lives in the County but not in Willits.

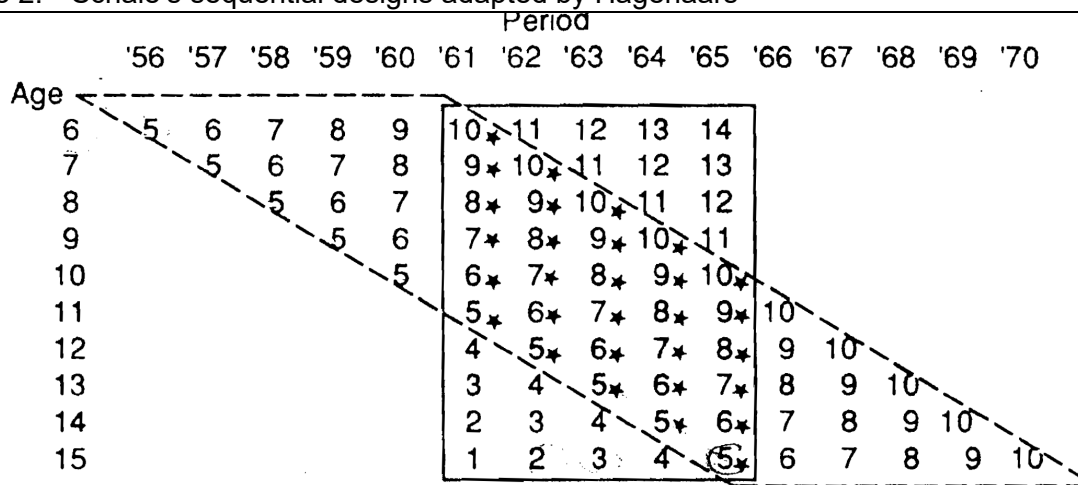
Age, birth cohort (year of birth), and time are collinear [23]. Knowing two of these, one can calculate the third. The status of cohort is the same as variables such as sex or race except for the inherent time component [24]. Cohort provides a framework to examine data as an interaction between age and period, namely, the result of aging during specific historical events.

As we are writing this, we sadly remember the Japanese in Nagasaki and Hiroshima at the end of World War II and the health problems this cohort faced through their lives. Today, their history echoes to another Japanese cohort, residents of Fukushima in 2011. The entire nation struggled with the aftermath of one of the largest earthquakes in world history, plus tsunami, plus nuclear power plant meltdown, with Fukushima residents most affected, and nuclear trash circling the globe. This is the current, frightful example of the intersection of age, place, and time that the cohort construct attempts to understand. Cohorts also may be defined for other types of




periods, for example, periods when people worked at a facility with changing worker safety standards evaluated relative to the period when they died and the cause of death [25], or if men working in a hexavalent chromium processing plant developed cancers [26].

In longitudinal research, omitting birth cohort is justified only if it is unrelated to the dependent variable. Figure 2 highlights the problem. In longitudinal research, the choice of model depends on the question asked. The objective of an analysis based on age-period (time sequential) is descriptive: a person is in different groups at different times. This is similar to the objective of the DataBooks FHOP produces. One uses the cohort-period when the group is the unit of analysis and the hypothesis is that events over an extended period affected the outcome [23].

Figure 2. Schaie's sequential designs adapted by Hagenaars



in the cells: cohort i

-  : age x period (time-sequential) design
-  : age x cohort (cohort-sequential) design
-  : cohort x period (cross-sequential) design

If one has a dataset with age but not cohort (e.g., population numbers to use as a denominator), cohort can be calculated by subtracting age from the year of the dataset. For example, the population age 20 in the year 2000 was born in 1980. Depending on the number of cases available for analysis, cohorts typically are grouped into 5, 10, or even 20-year intervals as needed to minimize small number problems.

DEMOGRAPHIC VARIABLES

Sex

The word "sex" refers to biological and physiological characteristics that define male and female, while gender reflects role-based social expectations [27]. Some datasets define sex with alpha characters (M/F), others use numeric (1, 2). Sometimes 1 is defined as male, other times 1 is defined as female. Recognizing that biological and physiological characteristics defining male and female can be ambiguous, some datasets allow ambiguous sex codes ((U)known, (I)ndeterminate). When the incoming sex variable is in the form M/F/I/U, the SEXC macro converts it to our standard numeric form (1 = male, 2 = female), where FROM is the incoming character variable and TO is the outgoing numeric variable.

```
%SEXC(from = , to =);  
%SEXC(SEXC,SEX);
```

Race/Ethnicity

There is general agreement that the terms "race" and "ethnicity" are social-political constructs and should not be interpreted as genetic, biological, or anthropological in nature. With a focus on California, Yanow showed how definitions of these social constructs have changed over time, and, although the terms are used to refer to different things, how they are used interchangeably [28]. FHOP and Yanow concur that the term "race/ethnicity" is the preferred referent to describe such classifications of complex human relationships.

Since 1977, through Office of Management and Budget (OMB) Policy Directive No. 15, the United States has used five "standard" race and/or ethnic categories -- White, Black, Hispanic, Asian, American Indian/American Native (AIAN) (and their variants) -- to organize, summarize, and describe human experience [29]. The 1997 revision of Directive No. 15 changed minimum categories to AIAN, Asian, Black or African American, Native Hawaiian or Other Pacific Islander (NHOP), and White, adding two categories for ethnicity: "Hispanic or Latino" and "Not Hispanic or Latino" [30]. The revised standards also added a requirement that respondents be allowed to select one or more race categories when responding to a query on their racial identity. This provision means there are potentially 31 race groups, depending on whether an individual selects one, two, three, four, or all five of the race categories [31]. In 2000, the OMB issued guidance on using these categories, including methods to bridge race/ethnicity for longitudinal studies [32,33].

Acknowledging the diversity of California's population, the state legislation adopted a series of regulations specifying that the following additional categories should be collected: Asian Indian,

Cambodian, Chinese, Filipino, Hmong, Japanese, Korean, Laotian, Thai, Vietnamese, Other Asian, Native Hawaiian, Guamanian, Samoan, and Other Pacific Islander [34].

In 2003, FHOP) and the California Center for Health Statistics (CHS) issued guidelines on race/ethnicity [35]. Their purpose was to guide compliance with the new national racial/ethnic data collection standards while also fulfilling California's need for consistent and more specific data given the unparalleled complexity of its population.

The federal government subsequently issued bridging guidelines to address variation in race/ethnic definitions over time [32]. These guidelines suggest that longitudinal investigations use their recommended groupings until enough years are available to permit more detailed analyses. The researcher must evaluate this recommendation in the context of the proposed analysis. Detailed race/ethnic comparisons are more reliable at higher levels (e.g., nation, state) than at lower levels (county, city). If numbers become too small, statistical issues limit the usefulness of the analysis. Small numbers for a given race/ethnic group in a small community play directly into concerns about protecting confidentiality. Thus, assigning detailed race codes to larger race groups requires thinking through many more issues than the categories available.

Officially, race bridging is defined as “making data collected using one set of race categories consistent with data collected using a different set of race categories, to permit estimation and comparison of race-specific statistics at a point in time or over time. More specifically, race bridging is a method used to make multiple-race and single-race data collection systems sufficiently comparable to permit estimation and analysis of race-specific statistics [36].” The goal of bridging is to approximate the size of single-race groups rather than to approximate how each individual would have responded to the traditional single-race question [37].

Table 2 shows race groups available for numerator and denominator datasets. Vital Statistics categories are pre-and post-2000. Different groups have been available at different times in data sets we use for numerators (OSHPD, Vital Statistics). OSHPD pre-1995 covers the period 1985 forward. They had yet another definition before 1985. A similar problem exists for denominators (DOF or NCHS). California's DOF changed race groups after the 2000 census. NCHS annual race bridging files use the same groups from 1990 forward.

Table 2. Available numerator and denominator options for race/ethnic reporting

Group	Numerator								Denominator							
	Vital Statistics				OSHPD				DOF				NCHS			
	Pre-2000		Post-2000		Pre-1995		Post-1995		Pre-2000		Post-2000		1990-1999		Post-2000	
	RACE	HISP	RACE	HISP	RACE	HISP	RACE	HISP	RACE	HISP	RACE	HISP	RACE	HISP	RACE	HISP
AIAN	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
API	C	Y	C	Y	Y		Y	Y	Y	Y	C	Y	Y	Y	Y	Y
Asian	C	Y	C	Y							Y					
Black	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
NHOPI	C	Y	C	Y							Y					
White	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Multi		Y	C*	Y							Y*					
Other	Y*	Y	Y*	Y	Y*		Y*	Y								
Unknown	Y*	Y	Y*	Y	Y*		Y*	Y								
Hispanic	C		C		Y		C		C		Y		C		C	

The RACE column shows whether the group is defined (Y) or can be calculated (C). In Vital Statistics data, API, Asian, NHOPI, and AIAN groups are aggregates of multiple categories. Blank cells indicate the option is not available. While Vital Statistics introduced different codes at different times, all codes aggregate to these groups.

The HISP column reflects whether the agency separately reports ethnicity (Hispanic and Non-Hispanic), which Federal rules prefer. Table 2 highlights that DOF and OSHPD groups limit longitudinal options for classifying race.

Note that multi-race is not in Vital Statistics files before 2000 and is not in OSHPD files. The asterisk (*) indicates the group must be bridged to provide consistent longitudinal race classifications. Asterisked groups are multi-race, other and unknown (which includes declined to state), and missing (OUM). In this document, we refer to multi-race and OUM collectively as the Bridge Group.

NCHS publishes population numbers for Hispanic and non-Hispanic bridged race groups (American Indian/Alaska Native (AIAN), Asian/Pacific Islander (API), Black, White) [38] from 1990 forward. It uses the most sophisticated algorithms to produce estimates. NCHS calculates population to the county level by Hispanic, bridged race, and sex in 1-year age intervals. Federal agencies use these files as denominators to calculate national bridged statistics, including those that monitor vital statistics and progress toward Healthy People objectives. We have come to prefer bridging race/ethnicity using Federal rules, and using the NCHS population estimate files for denominators, if that is at all possible.

California requires state-funded researchers to use Department of Finance (DOF) population estimates [39]. The DOF provides county-level estimates by sex and race/ethnicity, with age in 1-year intervals. Through 1999, DOF files categorized race/ethnicity as White, Black, Hispanic (all races), API, and AIAN [40]. To allow us to calculate population rates involving longitudinal analyses beginning before 2000, we first assign cases to Hispanic all-race. If multi-race variables are available, we use the first variable, assigning the remainder to White, Black, API,

and AIAN. Pre-2000, DOF assigned the remainder to White race/ethnicity. By using DOF categories, it is possible to calculate longitudinal population rates using DOF population estimates where the study period begins before 2000.

For 2000 and later, DOF classifies race as White, Black, Hispanic, Asian, Pacific Islander, AIAN, and Multi-race [41]. To make classifications compatible longitudinally, we combine the 2000- and-later categories as follows. First, we combine Asian with Pacific Islander. Then we reassign Multi-race proportionately, within each county, closely adopting NCHS decision rules. Note that the DOF Hispanic category has no multi-race allocation because DOF assigns this category first before assigning other single- or multi-race categories.

Thus, if one needs to analyze race separately from ethnicity, use the NCHS bridged files with detailed age, sex, bridged race, and ethnicity [38]. If analysts have the flexibility, we recommend using those files for the denominator and bridging race in the numerator using the NCHS bridging file [42].

Variables describing race and ethnicity vary within and across datasets and over time. When we read data into SAS, we only do the simplest correction of data quality for these variables. For example, we convert missing or out-of-range values to the code representing "unknown" in that dataset and year. We keep all original race/ethnic variables. The type of race/ethnic classification we do depends on definitions available over the research period to be studied, and whether data will be merged with other datasets, for example, hospital discharge with death or birth certificates.

Given longitudinal issues in race code availability, and taking small numbers issues into account, we have come to prefer a longitudinally consistent variable with five groups: Hispanic All-Race, and non-Hispanic White, Black, API, and AIAN, created following Federal bridging rules. We have developed various macros that take into account the years covered by an analysis, the data source (birth certificates (BC), patient discharge (PD), death (DT), etc) that needs to be bridged, and the aggregation rules to be followed (NCHS, DOF, etc). These are available in TOOLS.ZIP, on our website [2].

We do not calculate rates for AIAN. Their numbers are small and both DOF and Federal rules diminish their numbers so significantly that numbers are not reliable [43]. For more information on various bridging methods, see our separate document on bridging race/ethnicity [44].

Health Insurance

OSHPD and birth certificate files contain information about health insurance coverage. Again, definitions vary within and across datasets and time. We use insurance as a demographic variable, a proxy for income. For example, people insured by MediCal or County Medical Services Programs are poor by definition as are most uninsured.

Where insurance status is OOR (less than 0.0003% of discharges), we assign insurance to Medicare if the person is age 65 or older. Otherwise, we assign OOR values to MediCal.

We defer standardizing insurance variables longitudinally until we make analysis datasets. How it is standardized depends on definitions available over the research period, the population studied, and datasets used. For example, will the study use one dataset or several with different definitions, for example, PDD, ED or BSMF.

CONFIDENTIAL DATA ELEMENTS

FHOP is privileged that its human subjects protocol allows it to receive confidential files with names, addresses, medical record numbers, and social security numbers. Original files with unencrypted confidential variables are stored on a password-protected external drive, in password-protected ZIP files. The external drive is stored in a locked cabinet when not in use.

When we read datasets with confidential variables into SAS, we capitalize character-based variables (NAME, ADDRESS), encrypt all confidential variables using an algorithm we developed, and put the encrypted variables into a separate confidential database.

The confidential file with the encrypted variables has a name similar to the main file, with a C suffix. For example, BT2005 is the output file containing non-confidential data elements from the 2005 BSMF. BTC2005 is the name of the file containing confidential elements. Both files contain a variable YR_OBS to permit linkage with the main file when needed.

FHOP uses an encryption algorithm originally developed by the authors when they were part of the team that did California's first patient outcome study based on linked data [45]. After moving to FHOP, they improved the algorithm, which is not made public.

Social Security Number

We encrypt the original Social Security Number (SSN) when we read the data into SAS but do not keep SSN in the final output. The macro MAKESSNC checks the SSN for possible errors (e.g., 111111111, or SSN equals birth date or admission date). If we find an error, we output the incoming SSN to a new encrypted variable (SSNCX) and set SSN to missing. The character variable (SSNC) contains the encrypted SSN if it passed the error check or is missing if it failed the error check.

We next invoke the macro SSNCN to make a new numeric variable (SSNCN) which contains the value of SSNC if it is available or a numeric value made by concatenating birth date, sex, and ZIP-code of residence. The purpose of SSNCN is to come as close as we can to make a unique "person" identifier. If we have SSNC, we use it. If we do not, we combine ZIP, birthdate, and sex, which is not unique but can be better than nothing and can be helpful in linkage.

Data are read into a temporary file, then sorted by SSNCN plus other variables to set the stage for data linkage. We store SSN variables in the main file. We create a variable YR_OBS to sequentially number records in the sorted dataset. YR_OBS can link sub-files to the main file.

Medical Record Number

When they are available, we encrypt Medical Record Numbers using the same algorithm. This can be some help in linking admissions for the same person in a given hospital, but is not helpful for constructing episodes of care where people transfer from one facility to another. It is helpful to pull records for medical reabstraction studies. When available, the encrypted Medical Record Number is stored in the confidential sub-file.

Record Linkage Number

OSHPD provides the Record Linkage Number (RLN) as a data linkage tool. It is an encrypted SSN using their proprietary algorithm. Shortly after its introduction, researchers on California's first linked patient outcome study [45] identified that the RLN did not permit soft linkages as well as the SSN. A soft linkage looks for data errors to increase the likelihood of linking records for the same person. The RLN can be used when SSN is not available, but results are not as robust, losing perhaps 10 to 15% of possible linkages. When we have the RLN, we save it on the confidential file. We do not further encrypt it.

Name

As Juliet asked of Romeo, "What's in a name? That which we call a rose by any other name would smell as sweet" [46]. These star-crossed lovers knew well that names convey sex, gender, culture, origin, class, wealth, hopes, and accomplishments.

Relationships between names and gender are fluid and change with time and context. In 1900, John and Mary were the most popular baby names in the United States [47]. In 2000, Jacob and Emily were the most popular, while John dropped to 14 and Mary to 47. The surname Shirley was firmly established as male until Shirley Temple became famous. The nation's baby book becomes increasingly complex with each wave of immigrants. Angel was a male name favored by Hispanic cultures until the last few years when parents began to name their daughters Angel. When parents of girls begin to give previously male names to their daughters, parents of boys tend no longer to give those names to their sons.

Confidential Vital Statistics and cancer datasets are examples of files with names. When names are available, we use standard name parts: TITLE (Mrs, Mr, Dr), LAST, FIRST, MIDDLE, SUFFIX (Jr, Sr, III, PhD, MD, JD, MSW). If a file has name in one segment, we call a macro to apportion the parts to separate variables. The BSMF/FDSMF have infant, mother, and father

names. Here, I is the prefix for infant variables, M for mother variables, and F for father variables.

The encryption macro uses the name of the new variable to be encrypted. We use the suffix E for encrypted name variables, e.g., LASTE. In this macro, we do not define the variable LAST, as logic in the macro knows that. When unencrypting name variables, they revert to the original (LAST).

```
%encrstr(LASTE);
```

Name parts can be unencrypted, merged with the source file, then summarized by SEX or race/ethnicity. At this point names are no longer confidential. We used this to make a first name database to impute gender or race/ethnicity when it is unknown. Name also can be used in linking algorithms or as the basis for unique identifiers [48].

Names in hospital disclosure reports (Chief Executive Office, Person Completing the Report) are not protected data elements, because hospital-level data is not confidential.

RESOURCES

We have focused on general principles for managing important longitudinal variables. We discussed general naming conventions and introduced a few key macros. The macro libraries we use are available on our website, in the file TOOLS.ZIP [2]. We strongly recommend that people adopting our methods join the FHOP SAS User 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: <http://fhop.ucsf.edu/sas-tools-0>
- 3 Shakespeare, W. (1623). Macbeth, Act 5, Scene 5, lines 18-21.
- 4 Instruction Manual Part 12: Computer Edits for Natality Data. Effective 1993 Vital Statistics Data Preparation USDHHS, PHS, CDCP, NCHS Hyattsville, Maryland, March 1995 Available at: Last accessed 20-Mar-2021 at: <http://www.cdc.gov/nchs/data/dvs/instr12.pdf>
- 5 Xu J (2005). COMP_GEST_NEW_FMT.SAS. US Division of Vital Statistics, National Center for Health Statistics, Centers for Disease Control. Dated 08-Feb-2005. Forwarded by L Sangare, California Department of Health Services, 14-Feb-2005. Last accessed 20-Mar-2021 at: http://mchepi.com/wp-content/uploads/2008/02/comp_gest_new_fmt.sas.
- 6 Salazar M. (2007) California Dept of Public Health Center for Health Statistics Information Technology Services Section. Personal Communication to LR 01-Nov-2007.
- 7 Christensen J. (2004) California Office of Health Information and Research, California Department of Public Health. Personal communication to LR 06-May-2004.
- 8 State of California Office of Statewide Health Planning and Development. (July 2006). Patient Discharge Data File Documentation. January-December 2005 Nonpublic Version, page 16.
- 9 Bronfenbrenner U. (1979). The ecology of human development. Cambridge MA: Harvard University Press.
- 10 Moen P, Elder GH, Luscher K (eds) (1995) Examining Lives in Context: Perspectives on the Ecology of Human Development. Washington DC: American Psychological Association.
- 11 Kawachi I, Berkman LF. (2003) Neighborhoods and Health. Oxford University Press.
- 12 Woodruff TJ, Janssen SJ, Guillette Jr LJ, Giudice LC (eds) (2010) Environmental Impacts on Reproductive Health and Fertility. Cambridge University Press, New York.
13. Strauss W, Howe N. (1991) Generations: The History of America's Future, 1584 to 2069. New York: William Morrow and Company, 1991.
- 14 Oliva G, Remy L, Clay T. (2004) The Impact of Changing Public Policy on California's Hospital Infrastructure and Children's Hospital Outcomes - 1983-2000. Available at: <http://fhop.ucsf.edu/fhop-publications-hospitalizations-trends-and-outcomes>.
- 15 Remy LL, Clay T (2014) Longitudinal analysis of health outcomes after exposure to toxics, Willits California, 1991-2012: application of the cohort-period (cross-sequential) design. Environ Health. 2014 Oct 24;13:88. doi: 10.1186/1476-069X-13-88. Available at: <http://www.ehjournal.net/content/13/1/88>
- 16 Remy LL, Byers V, Clay T (2017) Reproductive outcomes after non-occupational exposure to hexavalent chromium, Willits California, 1983-2014. Environ Health. 2017 Mar 6;16(1):18. doi: 10.1186/s12940-017-0222-8. Available at: <https://ehjournal.biomedcentral.com/articles/10.1186/s12940-017-0222-8>

-
17. James T, Thomasser RG: Final remedial investigation report: former Remco Hydraulics facility. Willits Environmental Remediation Trust. 18-Apr-2002. Montgomery Watson Harza, Last access 20-Mar-2021 at: http://www.willitstrust.org/key_documents/RI.pdf
 18. Underwood MC, Barreau T, Hoshiko S: Evaluation of exposure to historic air releases from the Abex/Remco Hydraulics Facility, Willits, Mendocino County, California. CERCLIS CAD000097287 July 21, 2003. California Department of Health Services under Cooperative Agreement with the US Department of Health and Human Services Agency for Toxic Substances and Disease Registry. Last access 20-Mar-2021 at: <http://www.ehib.org/ehib/www.ehib.org/cma/projects/AbexRemcoFinalAirPHA.pdf>.
 - 19 Barreau T, Underwood MC, Hoshiko S, Rojas T, LaPlante J, Eng G, McRae T, Zarus G: Evaluation of Exposure to Historic Air Releases from the Abex/Remco Hydraulics Facility, Willits, Mendocino County, California. CERCLIS CAD000097287. July 2, 2006. Last access 20-Mar-2021 at: http://www.atsdr.cdc.gov/HAC/pha//AbexRemcoHydraulics_TJ/AbexRemco_PHAfinal07-02-06.pdf
 20. Hoshiko S, Underwood MC: Evaluation of health studies possibilities and limitations at the Abex/Remco Hydraulics Facility Willits, Mendocino County, California Cerclis CAD000097287 July 11, 2006. Department of Department of Health Services under Cooperative Agreement with the U.S. Department of Health and Human Services Agency for Toxic Substances and Disease Registry (ATSDR). Last access 20-Mar-2021 at: <http://www.atsdr.cdc.gov/HAC/pha/AbexRemcoHydraulicsFacility/Abex-RemcoHydraulicsHC091906.pdf>
 - 21 Harrison R Expert Panel Report. Recommendations for Conducting Medical Monitoring for Residents of Willits, California and Workers Exposed to Hexavalent Chromium and Volatile Organic Compounds from the Abex/Remco Hydraulics Facility. University of California, San Francisco. Division of Occupational and Environmental Medicine. November 2006. Last accessed 18-Jul-2016 at: <http://www.ehib.org/ehib/www.ehib.org/cma/projects/MMReportFinal.pdf>
 22. Barreau B, Underwood MC, Hoshiko S: Evaluation of Exposures to Contaminants from the Former Abex/Remco Hydraulics Facility, Willits, Mendocino County, California. CERCLIS No. CAD000097287. Department of Department of Health Services under Cooperative Agreement with the U.S. Department of Health and Human Services Agency for Toxic Substances and Disease Registry. Sep 2006. Last access 20-Mar-2021 at: http://www.atsdr.cdc.gov/HAC/pha//AbexRemcoHydraulics_TJ/AbexRemco_PHAfinal07-02-06.pdf
 - 23 Hagenaaars JA: *Categorical Longitudinal Data: Log-Linear Panel, Trend, and Cohort Analysis*. Newbury Park, Sage Publications, 1990.
 - 24 Schaie KW, Baltes PB. (1975) On sequential strategies in developmental research. *Human Development* 1975, 18(5): 384-390.
 - 25 Taylor FH: The relationship of mortality and duration of employment as reflected by a cohort of chromate workers. *Am J Public Health* 1966, 56(2):218-229.
 - 26 Gibb HJ, Lees PS, Pinsky PF, Rooney BC. Clinical findings of irritation among chromium chemical production workers. *Am J Ind Med*. 2000 Aug;38(2):127-31.
 - 27 World Health Organization. What do we mean by "sex" and "gender"? Last accessed 22-Jul-2016 at: <http://www.who.int/gender/whatisgender/en/index.html>
 - 28 Yanow D. (2003). *Constructing "Race" and Ethnicity" in America: Category-Making in Public Policy and Administration*. Armonk NY: M E Sharpe.

-
- 29 OMB Directive No. 15: Race and Ethnic Standards for Federal Statistics and Administrative Reporting As adopted 12-May-1977. Last accessed 20-Mar-2021 at: <https://wonder.cdc.gov/wonder/help/populations/bridged-race/Directive15.html>
 - 30 OMB Directive No. 15: Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. As adopted 30-Oct-1997. Last accessed 03-Mar-2021 at: <https://www.whitehouse.gov/wp-content/uploads/2017/11/Revisions-to-the-Standards-for-the-Classification-of-Federal-Data-on-Race-and-Ethnicity-October30-1997.pdf>
 - 31 Ingram DD, Parker JD, Schenker N, Weed JA, Hamilton B, Arias E, Madans JH. (2003) United States Census 2000 population with bridged race categories. National Center for Health Statistics. Vital Health Stat 2(135).
 - 32 Office of Management and Budget, Washington, D.C. Provisional Guidance on the Implementation on the 1997 Standards for Federal Data on Race and Ethnicity, December 15, 2000. Last accessed 20-Mar-2021 at: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/omb/fedreg/r&e_guidance_notice.pdf
 - 33 Office of Management and Budget, Washington, D.C. Appendix C, The Bridge Report: Tabulation Options for Trend Analysis, Provisional Guidance on the Implementation on the 1997 Standards for Federal Data on Race and Ethnicity, December 15, 2000. Last accessed 20-Mar-2021 at: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/omb/fedreg/r&e_guidance_notice.pdf
 - 34 Government Code Sections 8310.5, 11092 and 11092.5. Government Code Sections 8310.5, Last accessed 20-Mar-2021 at: <https://leginfo.legislature.ca.gov/faces/codes.xhtml>
 - 35 FHOP/CHS (2003) Guidelines on Race/Ethnicity Data Collection, Coding and Reporting. UCSF/FHOP. Last accessed 20-Mar-2021 at: <http://fhop.ucsf.edu/fhop-publications-data-analysis-methods-guidelines>.
 - 36 See National Vital Statistics System, last accessed 20-Mar-2021 at: http://www.cdc.gov/nchs/nvss/bridged_race.htm.
 - 37 Parker JD, Schenker N, Ingram DJ, Weed JA, Heck KE, Madans JH (2004) Bridging between two standards for collecting information on race and ethnicity: An application to Census 2000 and vital rates. Public Health Reports, 119: Mar-Apr, 192-205. Last accessed 20-Mar-2021 at: <http://wonder.cdc.gov/wonder/help/populations/bridged-race/PublicHealthReports119-2-p192.pdf>.
 - 38 National Center for Health Statistics (2010). Postcensal estimates of the resident population of the United States for July 1, 2000-July 1, 2009, by year, county, age, bridged race, Hispanic origin, and sex (Vintage 2009). Prepared under a collaborative arrangement with the U.S. Census Bureau; released June 20, 2010. Last accessed 20-Mar-2021: www.cdc.gov/nchs/nvss/bridged_race.htm as of July 23, 2010.
 - 39 See: <http://www.dof.ca.gov/Forecasting/Demographics/Estimates/>. Last accessed 20-Mar-2021.
 - 40 State of California, Department of Finance, Race/Ethnic Population with Age and Sex Detail, 1990–1999. Sacramento, CA, Revised May 2009. Last accessed 20-Mar-2021 at: <http://www.dof.ca.gov/Forecasting/Demographics/Estimates/Race-Ethnic/1990-99/index.html>
 - 41 State of California, Department of Finance, Race/Hispanics Population with Age and Gender Detail, 2000–2010. Sacramento, California, September 2012. Last accessed 20-Mar-2021 at: <http://www.dof.ca.gov/Forecasting/Demographics/Estimates/Race-Ethnic/2000-2010/>

-
- 42 Johnson DP (2008). Using the Multiple Race Bridging Matrix. Documentation and NCHS race bridging file emailed to Linda Remy on 27-May-2011 from Dave Johnson, Survey Statistician, DHHS/PHS/CDC/CCHIS/NCHS/DVS/SPSRB. CDC/NCHS uses this file to bridge multi-race for all Federal vital statistics publications. Per Dave Johnson, CDC revisited whether to update the file in 2010, determined it did not need to be updated.
 - 43 Korenbrot CC, Crouch JA (2003) Disparities in hospitalizations of rural American Indians. *Medical Care*, 41(5): 626–636.
 - 44 Remy L, Clay T, Oliva G. (2011) Issues and Decisions to be made on Collecting, Coding and Reporting Race and Ethnicity for Public Health Indicators. Last accessed 20-Mar-2021 at: <http://fhop.ucsf.edu/data-management-methods>.
 - 45 Luft HS, Romano P, Rainwater J, Remy L. (1996) Annual Report of the California Hospital Outcomes Project: Acute Myocardial Infarction. OSHPD.
 - 46 Shakespeare W. (1594?) *Romeo and Juliet* Act 2, Scene 2.
 - 47 Social Security Administration. Popular Baby Names. Last accessed 20-Mar-2021 at: <http://www.ssa.gov/OACT/babynames/>
 - 48 Family Health Outcomes Project (1995). Unique Identifiers, Discussion, Recommendations, and Testing. Last accessed 20-Mar-2021 at: <http://fhop.ucsf.edu/public-health-data>