CalWORKs Home Visiting Program Evaluation Legislative Report

Submitted: January 7, 2022

Appendix C: Quantitative Methods (Secondary Data)

PREPARED FOR:

The California Department of Social Services

PREPARED BY:

Family Health Outcomes Project, Dept. of Family and Community of Medicine, University of California, San Francisco

Linda Remy, MSW PhD; Ted Clay, MS; Rita Shiau, MPH; Michael Clay, BS; Louise Kaseff, MS; Jennifer Rienks, PhD, MS

School of Nursing, University of California, San Francisco

Linda Franck, RN, PhD

Suggested Citation

Remy LL, Clay T, Shiau R, Clay M, Kaseff L, Rienks J, Franck L. (2021) CalWORKs Home Visiting Program Evaluation: Secondary Data Methods. University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

This work was supported by Contract 18-3262 with the California Department of Social Services. Jennifer Rienks, PhD, Principal Investigator; Linda L Remy, MSW PhD, Co-Principal Investigator Quantitative Methods; Linda Franck, RN PhD, Qualitative Methods.

Contents

List of Abbreviations

AFDC	Aid to Families with Dependent Children
AS	Ambulatory Surgery Center admissions, OSHPD
BT	Birth Statistical Master File, California Vital Statistics
CDSS	California Department of Social Services
CEDD	California Employment Development Department
CalWORKs	California Work Opportunities and Responsibility to Kids
CDPH	California Department of Public Health
CDSS	California Department of Social Services
CONP	Condition pull
CSV	Comma Separated Values
DRG	Diagnosis Related Group
DT	Death Statistical Master File, California Vital Statistics
DX(T)	ICD-9 or ICD-10 diagnosis code
ED	Emergency Department admissions, OSHPD
EDD	Employment Development Department data
EOS	Episode of service
FD	Fetal Death Statistical Master File, California Vital Statistics
FHOP	Family Health Outcomes Project
GSDP	Gender-specific diagnoses or procedures
HVP	Home Visiting Program
IP	In-patient hospital admissions, OSHPD
MDC	Major Diagnostic Category
MMEF	Medi-Cal Monthly Extract File
OSHPD	Office of Statewide Health Planning and Development
PX	Procedure code, type depending on IP/ED/AS setting
SSNC	Social Security Number, encrypted
SSNCBTH	Concatenated SSNC, sex, birth year
UCSF	University of California, San Francisco
VS	Vital Statistics

Background

The California Department of Social Services (CDSS) awarded the Family Health Outcomes Project (FHOP) a contract to evaluate the Home Visiting Program (HVP) operating under the auspices of CDSS CalWORKs (California Work Opportunities and Responsibility to Kids). CalWORKs is the state's implementation of the federal welfare-to-work Temporary Assistance for Needy Families (TANF) program that provides cash aid and services to eligible needy California families.

The quantitative (secondary data) part of the evaluation used hospital data from the Office of Statewide Health Planning and Development (OSHPD, inpatient (IP), emergency department (ED), and ambulatory surgery (AS)), vital statistics (VS) data (birth (BT), death (DT), and fetal death (FD)) from the California Department of Public Health (CDPH); employment data (EDD) from the California Employment Development Department (CEDD), data from various CDSS subdivisions, and client service data from county-level HVP. These various resources enabled FHOP to assess if outcomes improved for HVP clients and children.

The purpose of this document is to summarize the methods to prepare the various types of administrative data to support the CalWORKs HVP evaluation and to provide an overview of the methods to analyze those data.

OSHPD Files

OSHPD DATA

OSHPD distributes IP files from 1983 forward and ED and AS files from 2005 forward, with SSNC available in IP files from 01-Jul-1990 and in ED and AS files from the outset. The FHOP protocol allows us to have all years of the most confidential versions of these files which include dates (birth, admission, discharge, procedure) and Social Security Number (SSN, which we encrypt, SSNC). Developed for OSHPD's first hospital outcome reports, the encryption method was adopted by others [1-5].

OSHPD files have patient's sex, race/ethnicity, ZIP code and county of residence, expected source of payment, disposition, principal and up to 24 secondary diagnoses and procedures, and up to 12 external causes of injury. The PDD also includes admission source, type of admission, type of unit to which admitted, total charges, Major Diagnostic Group (MDC) and Diagnosis Related Group (DRG or MS-DRG).

Until 30-Sep-2015, all datasets described diagnoses using International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes. After that date, the datasets describe diagnoses using the ICD-10-CM.

The PDD use ICD procedure codes, while the ED and AS files use the CPT and Healthcare Common Procedure Coding System (HCPCS) to classify procedures. CPT is a proprietary coding system developed by the American Medi-Cal Association.

Documents describing our methods to standardize OSHPD files longitudinally are available on our website [6]. Unless specifically mentioned, all work was done in SAS.

CONDITION PULL

Approaching this task, we defined the reproductive-age population as age 0 to 49 at admission. Within this developmental age range, we pulled OSHPD records that indicated the patient lived in California (county of residence code in the range of 1 (Alameda County) to 58 (Yuba County) and not in a Pacific Armed Forces ZIP-code [7]). We also required that the principal diagnosis (PDX), patient sex and birth date be recorded; that birth date be on or before admission date; and if inpatient, that the stay was less than one year. Within each file (PD, ED, AS) we applied these basic criteria to select records, then flagged groups of interest based on age at admission:

- Newborn — Born in a hospital unit, e.g. not transferred in from another location or admitted through the ED, and birth date equaled admission date
- Infant — Not newborn and age less than one year
- Child — Age 1 to 14
- Reproductive age — 15 to 49 years old, and reproductive age female.

THE LONGITUDINAL FRAME – 1990 TO 2020

Our sets of VS files also start in 1983, with SSNC available in BC from 1997-2008 and in DT from 1989 forward. Names are available in VS files from 1983 forward, with BC addresses introduced in 1997, DT addresses in 2014, and FD addresses in 2019. When standardizing VS data, we encrypt SSNC, names and addresses, which we use for linkage. We use ESRI software to geocode VS addresses for mapping, and for the HVP evaluation, to select comparison families living near families that received HVP services.

Unlike VS files, OSHPD files do not have names or addresses. Our plan was to look back five years from when a family entered the HVP to understand their risk profile before entry and outcomes after entry. With the HVP starting in January 2019, we had to look back to January 2014 for the first participant families. The Family Registry, a companion resource also developed for the CalWORKs HVP evaluation, longitudinally connects siblings in birth order to their mother as identified in BC and FD files. The Family Registry calculates variables that set up linkage to OSHPD files to provide a better sense of health conditions and outcomes than is possible with VS files.

To learn how far back to look to construct a family, we turned to the 2014 BC file and retrieved the year of last live birth for mothers age 15 to 49 . To our great surprise, that file had 103 last live births between 1984 and 1990. Given this, we decided to look forward from 01-Jul-1990 (when OSHPD introduced SSNC) to make the Health History Registry. Not knowing yet who were cases or comparisons, or their reproductive or Medi-Cal histories, this kept the most options open. It took no more programming time, just more machine cycles and storage space. And it took a lot.

Over the years, both within and between OSHPD and VS files, race/ethnicity classification changed from simple to increasingly complex. We identified similar complexities in HVP case management and CDSS files we received. In this context, to facilitate data linkage and race/ethnicity reporting both within and across the various files, we calculated a longitudinally

consistent race/ethnicity variable that applies to data from all sources: White, Black, Hispanic all race, Asian/Hawaiian/Other Pacific Island (API), American Indian/Native American (AIAN), and Multi-racial, other, unknown (Other). The Federal Government uses these groups to report national performance measures such as Healthy People objectives [8].

CLASSIFYING HEALTH CONDITIONS

Clinical Classification System

FHOP uses the Clinical Classification System (CCS) [9] developed by AHRQ to classify health conditions and treatment. The CCS clusters patient diagnoses and procedures into a manageable number of clinically meaningful categories [10]. It offers the ability to group conditions and procedures without having to sort through thousands of codes. This clinical grouper makes it easier to understand patterns of diagnoses and procedures to analyze costs, utilization, and outcomes. Since first adopting the CCS, FHOP staff have published several studies using it [11-16].

We use both the multi- and single-level CCS, searching over the array of diagnoses and procedures [17] to classify all records as to affected body systems and diagnoses groups within them. The CCS also groups procedures into major and minor diagnostic and therapeutic procedures, and we classify all records accordingly. For specific groups, we added other indicators.

Given coding changes with introduction of the ICD-10 in October 2015, we used reverse (also known as backward) mapping [18,19,20]. Excel spreadsheets summarizing CCS diagnosis and procedure cross-classifications for ICD-9, ICD-10, and GEMS reverse mapping are available on our website [6].

CCS-Services and Procedures

ED and AS files use the CPT and Healthcare Common Procedure Coding System (HCPCS) to classify procedures. The CCS Services and Procedures (CCS-SP) software classifies these codes into clinically meaningful procedure categories [21]. It also classifies procedures into major or minor diagnostic or therapeutic procedures. These categories are parallel to the CCS ICD procedure classes with specific categories added that are unique to professional service and supply codes in CPT/HCPCS. CCS-SP is used with data that includes CPT or HCPCS procedure codes, as in the present instance, the ED/AS data. Excel spreadsheets summarizing CCS-SP cross-classifications are available on our website [6].

CLASSIFYING ACCESS AND OUTCOME INDICATORS

Every IP record was classified as to whether the patient was admitted to care from outside their county of residence, was admitted from jail or prison, was homeless at admission, transferred in from another hospital, or entered through the ED. AS and ED files do not have information on settings from which patients were admitted before seeking care.

In IP, ED, and AS, we identified admissions where the patient had adverse Medi-Cal events or died. In IP, we identified patients who had lengthy stays, transferred to another hospital, returned to jail or prison, or to some other non-residential location.

IDENTIFYING A "PERSON"

With almost 187 million records in the IP, ED, and AS condition pull files, 130 million (69.7%) had a recorded SSNC. In this section, we describe the work to identify a "Person" using this variable as the base. First, we assessed the availability of SSNC across several groups of variables. Next, we report work to identify the quality of data available to identify a "Person" in addition to SSNC, specifically birth date, sex, and race/ethnicity. We finish this section with a summary of the decision rules we arrived at for SSNC-based linkage in OSHPD files.

Availability of SSNC

To assess availability of the SSNC, we classified major demographic, access, disposition, type of care, and risk measures as described. Then we stratified the data by the presence or absence of SSNC

Due to the large number of newborns, infants, and children admitted IP, that setting is overall less likely to have the SSNC to facilitate linkage (60.5%). SSNC is more available for ED encounters (71%), and most available for AS (79%). In all settings, females are more likely to have the SSNC than males, and Blacks are more likely while Hispanic and Other (which includes multi-race) are less likely to have SSNC. Detailed tables (SSNC present, yes or no) by setting are available elsewhere [6].

Across settings, 86% to 88% of homeless have SSNC. It is less likely to be available for IP with payment from a public source including Medi-Cal (52%) than for ED (67%) or AS (68%) care. Only 52% of patients dying in the ED have SSNC compared with 59% of AS patients and 65% of IP patients. For patients with conditions of concern (mental health, substance use, injury), the SSNC is available for 84% to 87% across settings.

The 130,040,702 records with SSNC summed to 27,499,364 unique SSNC, an average of 4.7 admissions per SSNC. As expected, many of the same SSNC appeared in the various files. For example, of the 15,960,578 IP admissions with SSNC, 8,252,871 (51.7%) and 3,397,106 (21.3%) appeared respectively at least once in ED and AS units.

Decision Rules

We begin this discussion by introducing the possibility that the data entry clerk incorrectly entered the SSNC and the record may be reporting information for a different person. That is, the data quality problem may be with SSNC. We have no way to check this, so the core decision is to trust the SSNC. In OSHPD files, basic variables available to link records in addition to SSNC are birthdate, sex, and race/ethnicity.

BIRTHDATE. Across the three data sources, for SSNC with more than one record, differences in birthdate within SSNC ranged about 5% overall. We identified three different types of data errors. This included 1.6% of SSNC sets that disagreed on one date part (month, day, or year), and 0.7% that disagreed on two of the date parts. Another 2.7% had a SSNC record set with three inconsistencies. For example, 06-01-2001 on one record and 01-06-2002 on ten others. However, a good number of the 3-inconsistency sets had much more complex patterns.

For the HVP evaluation, we need all records for a given SSNC to have the same (hopefully correct) birth date to link the person's Medi-Cal history both within a given OSHPD file, across to other OSHPD files, and to merge with other files such as birth records that lack SSNC. In this context, we followed Romano et al [22].

> ***DECISION RULE****. If a birth date part (month, day, or year) is different on a given record from the majority of records with this SSNC, replace that part with the most frequent value across all records for the SSNC. If only two records, use the most recent. What is not Romano et al, because of the much longer period, save up to three of the most often found dates for a given SSNC.*

SEX. Across the three data sources, differences within SSNC in reporting sex was about 2% overall. One possibility is that the patient is female but the data entry clerk entered 1 (male), or vice versa. Yet another possibility is that the patient had a sex-change operation. In this context, we searched for gender-specific diagnoses or procedures (GSDP) that would lead us to select one sex over another.

> ***DECISION RULE****. When a given SSNC has both male and female sex and GSDP data do not conflict, reassign sex accordingly. If GSDP are not available, and less than three usable birthdates are present, assign the majority sex to all records. In either case, make a variable to identify the sex not used.*

RACE/ETHNICITY. Definitions of race and ethnicity changed over time. We standardized it to five consistently available race/ethnic groups: Hispanic all race, white, black, Asian/Pacific Island, and other which includes multi-race, unknown, and missing. We save the original variables, to help resolve a small number of cases.

Across the OSHPD data sources, differences in reporting race/ethnicity for SSNC with multiple records ranged about 19% overall, and even after relaxing criteria, it still remained about 5% overall. Some large part of this is due to the change over time in race/ethnic coding. For many years it was not possible to identify oneself in Medi-Cal records as multi-racial and the patient had to pick one race over another or select "Other". Sometimes a multi-racial person might use one of their racial backgrounds and use another some other time. Sometimes people claimed their Hispanic heritage and sometimes they did not. This becomes more complex when simultaneously considering the variation in coding race/ethnicity in vital statistics files or files specific to the HVP evaluation. And less we forget, the accuracy of data entry.

> ***DECISION RULE****. In this context, save up to four of the most frequently reported race/ethnicities for a given SSNC, with the most frequent first.*

Implementing these decision rules produced 27,499,364 unique SSNC with 55.1% female on primary sex after resolving GSDP. A second sex was identified for 540,570 SSNC (2.0%). Arranged in descending frequency, a second race/ethnicity was found for 5,260,731 SSNC (19.1%), and a third for 846,416 (3.1%). In cases with a tie, priority was given to the most recent. Again arranged in descending frequency and implementing a 99%ile cutoff, the maximum number of residential ZIP-codes for the SSNC was 8 and the maximum number of unique hospitals reporting admissions for the SSNC was 10. At least one IP admission occurred for 58.0% of SSNC, at least one ED admission occurred for 68.2%, and at least one AS admission occurred for 21.8%. The "Person" file was ready.

Frequent Flyers

We define “Frequent Flyer” [23] SSNC as having a number of admissions greater than 99.5% for the unit. The maximum number of times a given SSNC appeared in the IP was 1,059, 4,124 in ED, and 333 in AS, with 4,302 the maximum times overall. We set cut points for IP, ED, and AS respectively at 16, 40, and 5. This identified 451,907 (1.6%) frequent flyers overall, with 395,848 appearing in only one hospital unit, 50,707 in two, and 5,352 in all three units. We will approach with great caution any CDSS data that may appear with these SSNC.

SUMMARY

We broadly described our methods to prepare OSHPD files for the HVP evaluation. This included an overview of how we assign health conditions when diagnosis codes convert from the ICD-9 to the ICD-10. We also identify available access and outcome indicators. The way we identify a “Person” changed much for the better for the HVP evaluation compared to the method used in prior work [24,25].

A detailed description of methods to prepare OSHPD files is available on FHOP’s website [6].

Vital Statistics Files

VITAL STATISTICS DATA

FHOP has confidential CDPH vital statistics (birth, fetal death, and death) files from 1983 through 2019. As they arrive annually, they are read into SAS and standardized to support longitudinal research. Descriptions of methods to standardize incoming data are available on the FHOP website [6]. As with OSHPD files, we encrypt confidential variables when available. Unless specifically mentioned, all work was done in SAS.

CLASSIFYING BIRTH AND FETAL DEATH INDICATORS

To classify indicators in birth and fetal death files, we used code developed to produce the product we call DataBooks. These report 12-year trends for various maternal, child, and infant outcomes, most of which are also Federal Healthy People indicators [8]. We distribute DataBooks to Local Health Jurisdictions (LHJ, California’s 58 counties, Long Beach, Pasadena, and Berkeley) through our contract with the Maternal, Child, and Adolescent (MCAH) branch of CDPH. This code has been validated and approved by MCAH. Information on DataBook coding rules is available on the FHOP website [6].

Variables available to describe education of the mother and father changed four times over the years. This required that we develop a macro to standardize education consistently over time.

Marital status is not available, but when fathers do not claim paternity, their name is not on the certificate [26]. In this context, we first clean the father’s last name of strings such as “refused”, ”unknown”, “withheld” and various spellings of such phrases, then if the father’s last name is present, we make the variable FATHER to indicate that paternity was established. This is part of data standardization when we prepare these files.

Reporting of race and ethnicity changed over time. We classified it several ways to facilitate linkage within these files and across to others. This includes basic race/ethnicity categories

used for OSHPD files, as well as bridging race/ethnicity using tools available from the Federal government [27].

We output the classified data into the following files:

- Mothers N = 19,360,771 65 variables
- Fathers N = 15,234,851 34 variables
- Baby N = 19,360,771 75 variables

We output unclassified data into these files for subsequent work:

- Names N = 19,297,490 52 variables
- Addresses N = 11,821,527 15 variables
- Work N = 11,253,792 21 variables

CLASSIFYING DEATH INDICATORS

We used the same macro to classify race/ethnicity and education that we used in the BC and FD files. In addition to basic variables needed for linkage, we kept the variable reporting if the decedent had been pregnant in the last year, had an operation before death, the cause of death (which converted over time from ICD-9 to ICD-10), and the address where the person died. These variables facilitate linkage with various files.

Confidential death files include decedent names and, since 2014, names of the spouse and decedent's mother and father. Occupation and address variables also became available in 2014. Address variables include where the patient died, which usually is in a hospital. This is an important new linkage resource. As in other files, we encrypt SSNC, names, and addresses. We also output name, address, and occupation files.

We output the following files

- Decedent N = 2,407,976 52 variables.
- Names N = 8,464,128 32 variables
- Addresses N = 3,680,540 11 variables
- Work N = 1,573,969 21 variables.

CLEANING NAMES

With SSNC missing for most years of birth and all fetal death files, linking pregnancies requires the use of names. Additionally, most HVP case management data only use names. Unfortunately, names often are mis-spelled, increasing the linkage challenge.

In this context, from the BC and FD files we output all first, last, and middle names for mothers, fathers, and infants. From the death file we output name parts for the decedent and spouse. Concatenation and transposition produced a long, skinny file of 199,375,185 million records, flagging whether name parts were from mother, father, infant, or decedent and whether first, last, or middle name part. Now separated from the source files, we unencrypted name parts

then summarized the data to 5,324,664 records. We wrote some macros to clean these and made a format to clean name parts found in the files we call during linkage.

This process identified records in the name summary file with issues we could correct, specifically name prefixes (e.g., Mr., Mrs., Sra., Dr., Lt., Capt., Rev.) and suffixes (Jr., Sr., II, III). We applied our format to the original list of records and found we could clean 2.5% of first or last names in the BC file and 1.13% in the DT file. This translated into cleaning the first name for 614,941 females, 614,766 males, 541,540 newly born infants and 229,142 spouses of decedents.

CLEANING ADDRESSES

In BC files, mother's street address has been available since 1997 with mailing address introduced in 2007. The DT file introduced the decedent's residence address in 2005 and the address where the decedent died in 2014. The FD file does not have addresses. In the end, we had 21,971,309 address records available to use as part of our linkage algorithm.

For data linkage, we want to compare two records and decide whether they should be linked or not. For records with address variables, we use two versions, the original and the "best". When comparing records we look for variables whose values match exactly and approximately, also known as "deterministic" and "probabilistic" linkage.

To prepare address data for linkage, the first step involves cleaning and standardizing address parts. For example, consider the many ways street pre- and post-directions (East, West, North, South) and street types (e.g., Avenue, Street, Boulevard, Highway) might be entered. To give a sense of this task, we identified 136 spellings of "Avenue". We developed macros to clean these parts of addresses.

Next, we identified records that were not home addresses. Addresses that do not represent street locations may be PO Box records, extremely short street addresses, or problems of other sorts. In the BC file, 6,469,243 records had a mailing address as well as a street address, with 150,396 identified as PO Boxes.

Another issue is potentially high-risk residence situations. We discovered that 405 women gave the same address when their baby was born. Thinking initially that this would be a data quality problem, we output to a spreadsheet all addresses with 10 or more births (N = 1,446) then began an on-line search. To our surprise, a good many are the location of a non-hospital-based drug or alcohol recovery program, domestic violence shelter, homeless shelter, or pregnancy shelter for unwed mothers. Following up on this, we located a California Department of Health Services list of non-hospital-based licensed drug, alcohol, and mental health treatment programs [28]. We converted these into a format to flag these addresses in any file where they may appear. These provide another resource complementary to OSHPD's in-patient admissions to mental health and substance treatment units.

We also made a format to flag other high-risk address strings we identified. This includes several hundred heart-wrenching address strings: "in Ventura River bottom, north end, near highway 33", "bus stop at Market and Van Ness", "homeless walked in", "999 Transient Way", "abandoned warehouse", and simply "homeless". These complement the homeless code in OSHPD files.

Preparing to go to ArcGIS, 189,227 addresses had some kind of preliminarily identifiable problem. For example, 17,520 records with a California county as their residence did not have a California ZIP Code, 14,379 had a street address that was too short to geocode (length of 4), and 9,025 did not have a city name.

GEOCODING ADDRESSES

We flagged PO Box addresses in the BC file, addresses in California counties, and converted the file from SAS to CSV (Comma Separated Values) to import into ArcGIS for geocoding. The output file had 11,821,551 BC records with a street address, 6,469,252 BC records with a mailing address, and 3,680,540 death records.

For every address it is given, ArcGIS produces a record with results of its geocoding process. Sometimes that process is not successful. The first type of geocoding failure is where ArcGIS rejects the geocoding results, specifically the result product did not meet the 60% minimum score threshold built into ArcGIS

The second type of geocoding failure is “FHOP rejects”. Here, ArcGIS produces a match that we reject. Specifically, we rejected matches where state is not California, county is not adjacent to the county on the original record, and possibly the score is too low. We also re-clean and try again to geocode addresses we thought could be coded. Once records successfully geocode or locate with a specific latitude and longitude, ArcGIS produces a set of variables with its version of the standardized address. At the end of geocoding only 0.02% of addresses failed to meet the minimum ArcGIS standard. To set up data linkage, we made a file with the best address we could identify after several attempts, with the following results after merging back the addresses we did not submit:

- 95.2% of BC street addresses had the same state, county, city, and ZIP
- 97.7% of BC mailing addresses had the same state, county, city, and ZIP
- 94.3% of DT street addresses had same state, county, city, and ZIP

After identifying the address latitude and longitude, the next step was to submit the geocoded file to locate where the address is within census area variables such as block, block group, and census tract. We use these areas to identify families to compare with HVP families. Rooting case and comparison families in the same neighborhood or community provides a way to control for local demographic characteristics.

We merged this with the original record from its source (BC, DT), retaining cleaned address and census area variables. This added records we did not use for geocoding. At the end of our work, the following shows the median score statistic and percent assigned to blocks, the lowest census geography level:

- BC street address Median score 99.5 93.5% assigned blocks
- BC mailing address Median score 99.5 89.7% assigned blocks
- DT street address Median score 99.5 92.1% assigned blocks

We made formats to flag residential treatment addresses and other high-risk addresses identified during this journey. A document with a fuller description of this work is available on the FHOP website [6].

CLASSIFYING OCCUPATION

Background

Federal agencies use the North American Industry Classification System (NAICS) to classify business establishments and to collect, analyze, and publish statistical data related to the U.S. business economy [29]. Developed under auspices of the Office of Management and Budget (OMB), NAICS was adopted in 1997 to replace the Standard Industrial Classification system. To allow for a high level of comparability in business statistics among North American countries, NAICS was developed jointly by the U.S. Economic Classification Policy Committee, Statistics Canada, and Mexico's Instituto Nacional de Estadistica y Geografia. NIOCCS is a free web-based tool used to translate industry and occupation text into standardized codes [30]. Coding is based on the U.S. Census Industry and Occupation Classification system with options for coding to the Census 2000, 2002 and 2010 schemes. Output files include NAICS and U.S. Standard Occupation Classification (SOC) codes associated with the Census [31].

Cleaning Occupation in Birth and Death Certificates

BC files include text fields describing usual occupation and business for mother and father (1989-forward), and DT files have similar fields (2014-forward). To date, FD files do not have work-related variables.

After standardizing race/ethnicity and education as described earlier, we output two temporary BC files, one each for mother and father. We concatenated these files, assigning sex by source (father = male = 1; mother = female = 2). The DT file required only one output file which we concatenated with the birth file.

The last step in preparing files for NIOCCS involved cleaning work variables (occupation, business/industry) using a macro developed specifically for this task. The macro first removes special characters, blanks character strings indicating absence of data (e.g., refused, withheld) and fixes high frequency misspelled words or abbreviations (‘U S A F’ becomes ‘USAF’). Then we flag data quality issues that might impact reporting work-related information. We output two files, one with no records indicating either occupation or employer and the other with some information about occupation or employer.

We repeated essentially the same task for occupation variables in the death files. Then we concatenated the cleaned birth and death files with some indication of occupation or employer. The next step involved summarizing cleaned business and occupation fields. Going in with 8,058,569 records, we emerged from the summary with 1,469,866 records. Preliminary record cleaning reduced our volume more than 5-fold. Per NIOCCS rules, we output the summary to comma-delimited text files.

NIOCCS Classification

We began online processing using NIOCCS3, the version then available to the public. The first file of 150,000 records took 6 days to process. This would translate into about two months to convert the summarized list we had compiled.

After discussing our situation (including the sheer volume) with officials at the National Institute for Occupational Safety and Health (NIOSH), they offered to let us process these data using NIOCCS4, which was not yet available publicly. Given the size of our task, they gave permission to upload our files to the NIOSH server. We gladly became their guinea pigs!

According to NIOSH, NIOCCS4 uses a machine learning algorithm that is both faster and more efficient. Files that took days in NIOCCS3 take a few hours in NIOCCS4. NIOSH used this version for its own internal projects and made it public in 2021.

After receiving the coded SAS dataset back from NIOSH, we made formats for each industry and occupation classification from the Census, NAICS and SOC codes. We now have a simple 2-step process. Clean the incoming text string using the relevant macro, then assign codes for the string. We updated the BT and DT files accordingly.

A document with a fuller description of this work is available on the FHOP website [6].

PREPARING FOR LINKAGE

To prepare the BC and FD files for linkage, we cleaned the mother's name using the macros we had developed, selected variables from the mother and infant files, and for the BC files pulled residence variables from the address files and calculated work variables from formats based on occupation and employer. Then we merged these files. For the period 1990-2019, this produced 15,920,663 BC records and 89,926 FD records.

SUMMARY

In preparing VS data for the HVP evaluation, we used validated software to classify pregnancy and birth indicators in BC and FD files. For many years we wished to clean names and addresses to improve linkage, which the HVP evaluation provided an opportunity to do. We also were able at last to classify work-related variables.

California Department of Social Services Files

CAL-OAR FILES

Because no centralized repository exists to identify and collect information on all families participating in CalWORKs HVP, we used a variety of data sources to complete this task. As part of ensuring ongoing evaluation of county CalWORKs programs, CDSS implemented the California CalWORKs Outcomes and Accountability Review (Cal-OAR) process in 2019, part of which involves CDSS receiving monthly downloads of select data fields from the State Automated Welfare System (SAWS) Consortia databases used by county social services departments.

At CalWORKs HVP inception, Consortia databases were modified to add a field allowing counties implementing CalWORKs HVP to indicate a client's HVP participation status. As of April 2021, 35 of the 44 CalWORKs HVP counties reported some data about HVP participation into the Consortia databases. However, completing this field was not mandatory and not every county used it consistently over time. As a result, we could not use this to identify all CalWORKs HVP clients.

We prioritized case identification using HVP client datasets requested directly from 19 counties, not only because the client lists were likely more complete, but also because county data contained other valuable information that allowed for better client characterization, such as HV model and program enrollment and exit dates. Of the remaining 25 counties from which we did not directly request data, we used the Cal-OAR HVP indicator field from 10 counties to identify HVP clients.

HVP families from the remaining 14 counties were identified via client lists that the prospective evaluation arm solicited from counties at three different timepoints to serve as the sampling frame for client surveys. In 6 of these 14 counties, we used the client survey list rather than Cal-OAR because of inconsistent HVP indicator use.

MEDI-CAL MONTHLY EXTRACT FILES

CDSS provided the Medi-Cal Monthly Extract Files (MMEF) which we standardized following our usual processes to facilitate linkage with other datasets. After standardizing, we stacked the files and cleaned names and addresses (N = 149,838,919). Using SSNC from the AFDC file (N = 4,927,283), we merged with the cleaned file to select AFDC families (N = 70,530,854). Then we output one record for each episode of AFDC eligibility for the period 2015 through June 2021. The end product has linkage, demographic, name, and address variables for all AFDC families at each eligibility episode (N = 15,064,244).

The MMEF files come with a combined city/state variable, length 26. In addition to shortening the city name (Huntington Pk) which makes linkage less successful, we found 11,521 different ways to spell the names of 1,277 California cities and pleasantly learned that our city cleaning macro corrected all.

After cleaning city names, we followed our standard procedure when city, ZIP, or county is missing. If city is missing but ZIP is present, we assign city based on ZIP. If county is missing we assign county based on city.

Note that reported characteristics such as race/ethnicity and language change over time. For example, during different episodes a person will identify themself as Hispanic, other, or multi-race.

CHILD PROTECTIVE SERVICES FILES

We received the Child Protective Services files on 08-Dec-2021, which was too late to process. We will begin work with these files after we select comparison cases.

Employment Development Department Files

CDSS provided the following Employment Development Department (EDD) files that we standardized following our usual processes to facilitate linkage with other datasets:

- The crosswalk file with the employee's birth date and SSNC, the Employer Account Number (EAN), and the Employer Identification Number (EIN) (N = 626,741). This file has 1 to n records per SSNC.
- The employee file for the period 2015 Quarter 1 to 2021 Quarter 1 (N = 95,869,290). This file has the Employer Account Number (EAN), employee name, SSNC, and quarterly earnings but no demographics.

We unduplicated the employer file to one record per SSNC and birth date combination (N = 418,262). Then we cleaned names in the employee file, unduplicated to one record (N = 9,404,654), and merged with the birth date file (N = 9,822,916). No record with SSNC and birth date merged with any record in the employee file. The last step was to merge with the MMEF AFDC file, to identify household members (N = 3,303,861). Birth date present
418,262

- First name present 2,885,304
- Last name present 2,885,599

We merged this file with the MMEF file to obtain demographic information about those who worked. Further work to understand employment waited until after making the case and comparison list file.

CalWORKs HVP Client Case Files

DATA PREPARATION

Facilitated by CDSS HVP staff, we reached out to 19 counties to directly request client-level data needed for this evaluation. As a condition for funding, counties agreed to provide data necessary to CDSS for program evaluation purposes. Along with this initial outreach, CDSS included a memo requesting contact information for relevant county staff familiar with case management data collection, and requested the following type of types of information from each county:

- Identifying information for participating clients and those in their households, including names, birthdates and addresses, so that records may be linked to other administrative databases such as CA Vital Statistics and hospitalization data to assess whether HVP participants have different health and social outcomes compared to non-HVP participants.

- Data and notes related to each home visiting encounter, such as frequency, length and mode of visits, assessments results, needs identified, service referrals and usage, and other benefits accessed as a result of home visiting.

- Data addressing social and economic condition of the family, such as food, housing, employment and financial stability, participation in educational programs or workforce training by client or others in the household, key events affecting household stability such as marriage, separations, and other major life and health events.

At first contact, we clarified for each county the type of data that would allow us to address legislatively mandated indicators. We followed this initial request with multiple conversations, sometimes with a single county representative, and at other times with large teams involving program staff, analysts, and database contractors. The purpose of these meetings was to learn more about each data system and refine the appropriate data fields to transmit, given each system's scope and limitations.

This negotiation often lasted months with each county because of vastly different data system set-ups, agency-client consent processes, and interagency data transmission permissions. Our challenges, lessons learned and recommendations for improving the data request process are documented in the main body of this report. While the first sample dataset was received January 2021, we received final datasets from 21 databases containing data for 19 HVP counties for the evaluation from September to November 2021.

In addition to working with county staff, we worked closely with a case management database vendor used by the seven of the 19 counties. This helped to decrease the variability in data format received from these counties. In these instances, we asked the county to help the vendor identify the relevant data fields to include in the query, and give the vendor permission to perform the data query and transmit the data. The vendor then performed the data extraction directly from the backend of their databases and transmitted the datasets directly to CDSS for forwarding to this research team.

Once we received the data, we documented the contents of and relationships between each county's data tables. Relevant tables were imported into SAS using a standardized procedure to ensure that as much data as possible could be analyzed across counties in a consistent format. The number of tables received from each county ranged from 1 to 41, with most sending about 10 to 15 tables.

As part of the documentation, we cross-walked each table with information relevant to each legislatively mandated indicator to ensure that similar information was extracted from each table, and to determine the level of granularity with which we could confidently report. For example, consider developmental screening. While some counties reported dates, results, and disposition for every attempted and completed screening, others only documented whether screening was completed. Thus, for this measure, we were limited to reporting only the number of screened clients, rather than screening details, which would require limiting the number of clients included in the analysis and run the risk of identifying the program model.

For each source, we processed tables that contained any demographic information for any HVP family members, such as names, birthdates, sex, addresses, family role, primary language, race/ethnicity, and highest education level. Values for sex, language and race/ethnicities were standardized across counties and state data sources. These processed tables were collated into one large case list table, which was then deduplicated within and across sources, and matched with other data sources to gather more information about each

case family, described in next section. We also added some data gathered by Resource Development Associates (RDA) to interview clients. In the end the CalWORKs file contained 11,379 records, with data from the following:

- CalOAR 519 records
- COUNTIES 10,454 records
- RDA 406 records

Next we moved on to process tables similarly for variables needed to calculate the legislatively-mandated indicators. We determined the final variables needed for the calculations, and worked within each county's dataset to recode, transform and standardize the relevant tables, subsetting the appropriate population to include for the measure. We then combined each source into one dataset to perform final data recodes and summaries. We were not able to calculate some legislatively mandated indicators because data was not available to calculate them.

CALWORKS HVP CASE DATA

After CalWORKs HVP case data was available, the next task was to clean the file and calculate other needed variables. We identified 249 duplicate records so set them aside. We also identified another 807 records where one or more people in one family also appeared in records for another family, often with the same start and end dates. Pending further processing, we kept these records in the file.

We cleaned names using the macros we developed from names in birth and death files, did a preliminary cleaning of street addresses and city names. After cleaning city names, we followed our standard procedure when city, ZIP, or county is missing. If city is missing but ZIP is present, we assign city based on ZIP. If county is missing we assign county based on city. We then geocoded the street addresses using ArcGIS.

In the MMEF file, the variable SERIAL (the MEDS family number) had a length of seven. In the case data, 5,650 records had no SERIAL, 2 had length 5, and 48 had length up to 11. Thinking that SERIALs of length 6 or 7 were missing leading zeros, we imputed those and used the first 7 digits of the 48 longer. The SSNC was available in only 100 case records.

We identified 510 records with no start date and 7,128 with no end date. We calculated a new start date as the minimum of available start dates and a new end date as the maximum of the available end dates. If the end date was still missing, we set it at 30-Jun-2021. For cases with an end date but no start date, we calculated a start date six months before the end date.

After processing, we identified that the RDA person identifier was constructed when the data was standardized and there was no family identifier. We set those aside, since very little was available in them (N = 406). A table summarizing characteristics of the valid records as we understood them at the time (N = 10,948) is in the main report.

LINK CALWORKS HVP CASE DATA TO MMEF DATA

As discussed, the CalWORKs HVP case data has missing or inaccurate information that makes it difficult to navigate linkage across the various data sources this evaluation requires and to better describe the characteristics of family members. Much of the missing information is available in the MMEF EOS file.

In particular, we were concerned that the family have one start and end date to cover the period during which any family member participated in CalWORKs HVP. Toward this end, we summarized the CalWORKs case file by the CDSS family identifier to get the earliest start date and latest end date for the family. We also summarized the number of people in the family by roles (mother, father, other adult, child) and identified at the family level the HVP program model that helped them. This summarized the individual-level file to a family-level file of 5,251 families. At the present time, we show the following program level information:

- 2,001 families enrolled in Parents as Teachers
- 1,428 families enrolled in Healthy Family America
- 838 families enrolled in Early Head Start
- 520 families enrolled in Nurse Family Partnership
- 366 families enrolled in a local program
- 98 families were in a program we could not identify
- No family enrolled in more than one program model

At the family level:

- 15 families had two mothers and 3 had two fathers.
- 3,407 families were headed by a single mother
- 806 families had 1 to 4 adults living in the home, with 651 apparently acting as the primary caregiver.
- 432 families were two-parent households.
- At this point, 610 families had no identified adult caregiver
- 2,542 families had 1 child, 1,255 had 2 children, 222 had 3 children, and 161 had 4 to 11 children

The HVP Case and Comparison File

We obtained data from CDPH and Los Angeles County identifying their cases that had received home visiting under their auspices. We standardized them (CDPH N = 9,178, Los Angeles County N = 4200).

Current Status

After many twists and turns in data acquisition, completely unaware of data quality issues we faced, we received the last case data in November 2021. As this methods report goes forward, we are in the midst of identifying CDPH and Los Angeles County cases that overlap with

CDSS HVP cases so a determination can be made as to whether to keep or drop them. Then the remaining CDSS and Los Angeles County cases will be dropped from the CDPH list and we will select comparison families for the HVP. Once comparison families are selected, we will prepare the analysis files including quantitative outcomes already prepared and calculate the statistics needed to report outcomes.

Endnotes

1 Romano PS, Zache A, Luft HS, Rainwater, J, Remy LL, Campa D. (Dec. 1995) The California Hospital Outcomes Project: Using administrative data to compare hospital performance. Joint Commission Journal on Quality Improvement, 21(12), 668-682

2 Romano P, Luft HS, Remy L: *Annual Report of the California Hospital Outcomes Project. Volume Two: Technical Appendix. 2003*. California Health and Welfare Agency, Office of Statewide Health Planning and Development. Sacramento, CA.

3 Remy L, Clay T, Oliva G: *The California Child and Youth Injury Hot Spot Project Report for the Period 1995 to 1997, Volume Three, Technical Guide*. Sacramento, CA: California Department of Health Services. Aug 2000. Last access 31-Aug-2021 at: https://fhop.ucsf.edu/sites/fhop.ucsf.edu/files/wysiwyg/Vol_3_Tech_Guide.pdf

4 Kuzniewicz MW, Vasilevskis EE, Lane R, Dean ML, Trivedi NG, Rennie DJ, Clay T, Kotler PL, Dudley RA. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 2008; 133:1319–1327

5 Vasilevskis EE, Kuzniewicz MW, Cason BA, Lane RK, Dean ML, Clay T, Rennie DJ, Vittinghoff E. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. Chest. 2009 Jul;136(1):89-101. doi: 10.1378/chest.08-2591. Epub 2009 Apr 10.

6 Data Management Methods. See https://fhop.ucsf.edu/data-management-methods

7 See http://www.us-zip.org/armed_forces_pacific/apo/

8 Maternal and Child Health Bureau. Federally Available Data (FAD) Resource Document. April 13, 2021; Rockville, MD: Health Resources and Services Administration. Available at: https://mchb.tvisdata.hrsa.gov/PrioritiesAndMeasures/NationalPerformanceMeasures

9 See: https://www.hcup-us.ahrq.gov/tools_software.jsp

10 Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software (CCS), 2006. U.S. Agency for Healthcare Research and Quality. Available: http://www.ahrq.gov/data/hcup/ccs.htm#download.

11 Remy LL, Oliva G (2008) Acute episodes of mental illness among the population of reproductive age 1991-2005. For the Maternal, Adolescent, and Child Health Branch, California Department of Health Services. Available at: https://fhop.ucsf.edu/fhop-publications-hospitalizations-trends-and-outcomes

12 Remy LL, Oliva G, Clay T (2008) Maternal morbidity and outcomes including mortality, California 2001-2006. Family Health Outcomes Project, University of California San Francisco. Available at: https://fhop.ucsf.edu/fhop-publications-hospitalizations-trends-and-outcomes

13 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

14 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

15 Remy LL, Clay T, Byers V, Rosenfeld P (2019) Hospital, health, and community burden after oil refinery fires, Richmond, California 2007 and 2012. Environmental Health 18:48. https://doi.org/10.1186/s12940-019-0484-4

16 Remy LL, Clay T. (2020) Longitudinal trends for diabetes during pregnancy: California 1983-2015. Arch Clin Obst Gyn Res. Oct-2020, 1(1).

17 HCUPnet, Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality, Rockville, MD. https://hcupnet.ahrq.gov/.

18 Remy L, Clay T. (2020) Managing Longitudinal Research Studies: The crosswalk between ICD-9 AND ICD-10. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

19 Utter GH, Cox GL, Atolagbe OO, Owens PL, Romano PS: Conversion of the Agency for Healthcare Research and Quality's Quality Indicators from ICD-9-CM to ICD-10-CM/PCS: The Process, Results, and Implications for Users. Health Services Research. DOI: 10.1111/1475-6773.12981.

20 Maternal and Child Health Bureau. Federally Available Data (FAD) Resource Document. April 13, 2021; Rockville, MD: Health Resources and Services Administration. Available at: https://mchb.tvisdata.hrsa.gov/PrioritiesAndMeasures/NationalPerformanceMeasures

21 Clinical Classifications Software for Services and Procedures. Last accessed: 22Jun2007 at: http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp#table1

22 Romano PS, Remy LL, Luft HS. (May 1996) Second Report of the California Hospital Outcomes Project: Acute Myocardial Infarction. Volume Two: Technical Appendix, Acute Myocardial Infarction. California Health and Welfare Agency, Office of Statewide Health Planning and Development.

23 Hunt KA, Weber EJ, Showstack JA, Colby DC, Callaham ML. Characteristics of frequent users of emergency departments. Ann Emerg Med. 2006 Jul;48(1):1-8. doi: 10.1016/j.annemergmed.2005.12.030. Epub 2006 Mar 30.

24 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

25 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

26 Acknowledgement of Paternity/Parentage. See: https://www.cdph.ca.gov/Programs/CHSI/Pages/Acknowledgement-of-Facts-of-Paternity-Parentage.aspx

27 Remy L, Clay T. (2011) Managing Longitudinal Research Studies: Decisions to make in defining and bridging race/ethnicity. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

28 https://data.chhs.ca.gov/dataset/licensed-healthcare-facility-listing

29 North American Industry Classification System (NAICS). Last access 28-Jun-2020 at: https://www.census.gov/eos/www/naics/.

30 See: https://wwwn.cdc.gov/nioccs3/. Last access 28-Jun-2020.

31 See: https://www.census.gov/topics/employment/industry-occupation/guidance/code-lists.html. Last access 28-Jun-2020.