CALIFORNIA HEALTH AND HUMAN SERVICES (CHHS) PROGRAM DATA DASHBOARD METHODOLOGY

2.05.2020

BACKGROUND

CHHS has partnered with the USC Children's Data Network (CDN) to develop an annual "record reconciliation" that leads to the linkage and organization of administrative, client-level records across major CHHS programs each year. This data integration effort facilitates the exchange of statistical information concerning common clients as separately governed by the CHHS Intra-Agency Global Data Sharing Agreement. It helps CHHS and the Departments avoid inefficiencies that inevitably arise from ad hoc record linkage efforts specific to individual use cases, leading instead to a well-documented and routinized process for inventorying, cleansing, standardizing, and linking client-level records across programs. It also ensures that the same rigorous record linkage methodologies are used across CHHS programs. Most importantly, it supports CHHS and the Departments efforts to achieve better outcomes for all Californians through a richer evaluation of policy options, the improved stewardship of taxpayer dollars, and a more coordinated design and delivery of public services.

RECORD LINKAGE

An open source, machine-learning record linkage software program, ChoiceMaker (Version 2.7.1), was used for both within-program matching (or de-duplication) and between-program linkages (e.g., WIC records to child welfare records). ChoiceMaker uses probabilistic matching and modeling techniques for record linkage. To develop the record linkage model, data scientists developed a set of logical instructions, or model features, to examine commonalities between fields originating in different records. Individual features were then combined into a single linkage model used to determine the degree to which two records contained similar or dissimilar information. Each coded feature emerges with a weight, which indicates its relative predictive significance in determining a match. Based on a machine learning mathematical model called Maximum Entropy, an overall probability is generated to describe the likelihood that two records describe the same person (i.e., match).

As part of the model development process, analysts examine samples of record pairs, and for each pair, they indicate whether the records should be categorized as the same person (match), two different people (differ), or a hold (not enough information). The manually marked sample is then returned to ChoiceMaker Analyzer, a module of the software. The linkage model incorporates, or "learns", from the human decisions that were made and subsequently updates feature weights to best reproduce those decisions. This process is called training a model. When a trained model is subsequently applied to new record pairs, one finds that ChoiceMaker probabilities closely predict how a human expert would mark the new pairs.

Once the linkage process was completed, a unique pairwise Linkage Key was assigned to each record pair. This identifier is an 8-digit, alpha-numeric field that can be utilized across programs within CHHS as a master Common Client Identifier to facilitate the exchange of statistical program information, both within and between individual CHHS departments. Records with a .80 or greater match probability assigned by the model were retained as matched pairs.

DATA SOURCES

CalFresh

- Individual clients and cases
- Medi-Cal Eligibility Data System (2015-2018)

California Work Opportunity and Responsibility to Kids (CalWORKs)

- Individual clients and cases

- Medi-Cal Eligibility Data System (2015-2018)

Child Welfare
- Individual clients associated with Open Cases and Out of Home Placements (child, mother, father)
- Child Welfare Services Case Management System (CWS/CMS) (2015-2018)

Department of Developmental Services (DDS)
- Regional Center individual clients
- DDS administrative data (2016-2018)

Family Planning, Access, Care, and Treatment (Family PACT)
- Individual clients
- Medi-Cal Eligibility Data System (2015-2018)

Foster Care
- Individual child clients
- Child Welfare Services Case Management System (CWS/CMS) (2015-2018)

In-Home Supportive Services (IHSS)
- Individual clients
- Medi-Cal Eligibility Data System (2015-2018)

Medi-Cal
- Certified eligible individual clients
- Medi-Cal Eligibility Data System (2015-2018)

Women, Infants & Children (WIC)
- Individual clients
- California Department of Public Health (CDPH) administrative data (2015-2018)

TIME PERIOD

Program "Person" Count: Person level counts represent a census of clients in the specific CHHS program for the month of July of the designated year (2015-2018) by geography.
Multi-Program Participation: For each program specific July client population, participation in other CHHS programs during the designated year (Jan 1, XXXX-Dec 31, XXXX) is presented.

GEOGRAPHY

Linked data from the CHHS Record Reconciliation to code 4 geographic units for each record. All geographic levels are derived from the client's home address fields.

1. County
2. US Congressional District
3. State Assembly District
4. State Senate District

The availability of information required for geographic assignment (i.e. address fields), however, varied across participating programs. For example, Medi-Cal records include geocoded address data, while

other programs had only the minimum zip code field. As such, we adopted the agency's methodology for assigning geographic units to records used in the prior Data Dashboard:[1][2]

1. Utilize geocoded address data from the Medi-Cal Program file to code geographies. Geocoding is the process of extracting latitude and longitude coordinates of a location from its street address. These coordinates allow placement of households into geographic units of interest.

   Geographic shape files corresponding to the geographies can be found in the US Census Bureau's Legislative Areas National Geodatabase (tlgdb_2019_a_us_legislative.gdb). Documentation.

   County - *10.5 County*
   US Congressional District - *7.1 116th Congressional Districts*
   State Senate District - *7.2 State Legislative Districts*
   State Assembly District - *7.3 State Legislative Districts – Lower*

2. Where geocoded data is not available, geographic units are coded as follows:
   a. County
      i. Utilize available **county** variable in the individual program data set for coding.
      ii. If county is not available, utilize **zip code** to aggregate into county methods from the US Dept of Housing and Urban Development. *Please see - HUD / USPS Zip Code / County Cross Walk File - ZIP_COUNTY_062019.xls*[3]
   b. Congressional District / CA Assembly District / CA Senate District
      i. Use the following guidelines from the CA Senate Office of Demographics to aggregate **zip code** into select geographies:
         US Congressional District - *Final CD 2013 Zip Codes.xls*
         State Senate District - *SD_2014_2015 Zip Percentages.xls*
         State Assembly District - *Final AD 2013 Zip Codes.xls*
   c. Missing Values
      If a zip code did not have a corresponding geography, it is assigned to missing.

DATA SUBCATEGORIES

Sex: Categories include: "Male" and "Female." Several programs also include the category "Unknown/Other Gender."

Race/Ethnicity: Categories include: "Black", "White", "Hispanic", "Asian/Pacific Islander" and "Native American/Other/Unknown."

Age: Age was defined as of July 1 of the designated year. Categories include: "17 and Under" and "18 and Over" as well the subcategories: "18-64" and "65 and Over." *Information on additional program specific age categories can be found in the Program Specific Notes Section.*

---

[1] California Department of Social Services Research Services Branch. (n.d.). *Health & Human Services Program Dashboard Sources & Methodology*. Retrieved January 21st, 2020 from: https://www.cdss.ca.gov/Portals/9/DSSDB/Dashboards/CHHS/DPPSources.pdf

[2] Please note: There can be a 1 to 1 relationship between zip code and geographic units, as well as a 1 to many. In the event of a 1 to many relationships, records with that the zip code must be assigned to the specific geographic unit category based on the proportional formula set out in the state's guidance.

[3] When a duplicated individual can be coded in more than one geography, the geography is randomly assigned.

Program Counts: Categories include: “1 program”, “2 programs”, “3 programs”, “4 programs” and “5 or more programs.” Foster Care is a subset of Child Welfare and is only counted as one program. ACA is a subset of Medi-Cal and is only counted as one program.

Department Counts: Categories include: “1 department”, “2 departments”, “3 departments” and “4 departments.” Programs by department are as follows: CDPH (WIC), CDSS (CalFresh, CalWORKs, Child Welfare, Foster Care, IHSS), DDS (DDS), and DHCS (FPACT, Medi-Cal).

CLIENT CONFIDENTIALITY AND DATA DE-IDENTIFICATION

In accordance with the CHHS Data De-Identification Guidelines implemented to prevent the release of personally identifying information, counts of less than eleven and complementary cells have been suppressed. Small cells are annotated as follows:

- “0” no data in cell
- “*1” cell suppressed for small numbers (n=1-11)
- “*2” cell suppressed for complementary cell
- “*3” blank / not applicable for program

PROGRAM SPECIFIC NOTES

CalFresh

- Includes counts of persons and cases by geographies of cases (FFSSR)
- Includes program specific age categories: “18-59” and “60 and Over”

California Work Opportunity and Responsibility to Kids (CalWORKs)

- Includes counts of persons and cases by geographies of cases (SERIAL)

Child Welfare

- Includes program specific age categories: “age 0”, “1 to 2”, “3 to 5”, “6 to 10”, “11 to 15”, “16 to 17” and “18 to 20” for child clients and categories “18-64” and “65 and Over” for mothers and fathers.
- Includes Sex = “Unidentified/unknown”

DDS

- Program data not available for 2015.

Foster Care

- Foster Care child clients are a subset of the Child Welfare (CWS) program population and are counted only once in the program and department summary counts
- Includes program specific age categories: “age 0”, “1 to 2”, “3 to 5”, “6 to 10”, “11 to 15”, “16 to 17” and “18 to 20”
- Includes Sex = “Unidentified/unknown”

Medi-Cal

- Includes counts for Medi-Cal ACA expansion. ACA clients are a subset of the Medi-Cal program population and are counted only once in the program and department summary counts

Women, Infants & Children (WIC)

- Includes two program specific age categories options: “less than 19”, “20 to 24”, “25 to 29”, “30 to 34”, and “35 and over” which provides detailed age breakouts for adult clients and “age 0”, “age 1”, “age 2”, “age 3”, “age 4”, “5 to 19” and “20 and over” which provides detailed age breakouts for child clients.

- Includes Sex = “Unidentified/unknown”