California Health and Human Services Agency (CalHHS)
DRAFT Record Reconciliation Methodology

Prepared by the Children's Data Network

DRAFT REVISED July 19, 2023

California Health and Human Services Agency (CalHHS)
Record Reconciliation Methodology
(DRAFT REVISED July 19, 2023)

Table of Contents

Project History 3
Partners 4
Procedures 5
Project Agreements 5
Human Subjects, IRB, and VSAC Approvals 6
Time Period 6
Data Elements 6
Data Transfer 7
Data Cleaning and Hygiene Checks 7
Data Linkage 9
Linkage Algorithm 9
Within-Program Match / De-Duplication 10
Between-Program Match / Cross-Program Linkage 11
Pairwise Program Linkage Key 11
Matching Results 11
Return Data File 11
Structure / Record Layout 11
Data Dictionary 12
File Return 16
Confidentiality Measures 17
Appendices 18

California Health and Human Services Agency (CALHHS)
Record Reconciliation Methodology

Project History

Each year, California's Health and Human Services Agency (CalHHS) invests significant resources in programs designed to help California's most vulnerable and at-risk residents. The administration of these public programs is accompanied by the collection of rich data about the characteristics of clients served. Statistical information derived from these client records serves as an important vehicle for informing program planning and accountability, while also driving improvement initiatives.

Yet, given the complex nature of CalHHS's operational, fiscal, and regulatory commitments, the use of program-specific administrative data increasingly proves to be inadequate. While each program captures data concerning an individual client's encounters, typically absent is information concerning concurrent services and benefits that same individual may have received through other CalHHS programs. Also missing is critical data needed to understand the timing, sequencing, and outcomes of service and program encounters both within and across the Departments.

The current "program-centric" design of statewide data collection efforts is a barrier to policy and program administration and planning. It limits understanding of the collective size and impact of investments, and prevents the full assessment of population needs so that available resources are strategically coordinated and equitably allocated. Arguably, it also restricts innovation by reinforcing insights about clients and their service encounters through the lens of isolated programs. "Person-centered" program planning requires statistical information organized from the perspective of our clients.

To that end, in 2018, CalHHS partnered with the USC Children's Data Network (CDN) to develop a 1st ever "record reconciliation" that linked, organized, and analyzed administrative, client-level records across major CALHHS programs. Resulting in the creation of CalHHS Common Client Identifiers (CCIs) assigned to clients served by the largest programs administered by the Departments, this data integration effort facilitates the exchange of statistical information concerning common clients as separately governed by the CalHHS Intra-Agency Global Data Sharing Agreement. It helps CalHHS 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 CalHHS programs. Most importantly, it supports CalHHS 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.

Furthermore, these efforts fomented the development of a secure, cloud-based research enclave for hosting record-level research data sets and accompanying linkage keys. Once fully operational, this environment will provide carefully controlled, role-based access to analysts within CalHHS. In the longer term, the goal is to develop protocols that, with necessary approvals, will give external university-based and other research partners access to curated data sets and statistical resources within this analytic environment. This secure platform will advance rigorous evaluation, improve the reproducibility of research, create efficiencies in data management, and further the engagement of university-based researchers with government. Additionally, this Agency Data Hub will enhance record security and client confidentiality through data access and security protocols that can be more carefully audited.

Partners

The 1st record reconciliation (2018) involved 2016 data from 8 programs representing 4 CalHHS Departments including the California Department(s) of Health Care Services (DHCS), Developmental Services (DDS), Public Health (DPH), and Social Services (DSS). The 2nd record reconciliation (2019) expanded this effort to 2015-2018 data from the same agencies and added the California Department of Public Health's (DPH) Vital Statistics birth and death records. The 3rd reconciliation (2020) expanded the year range from 2015 through 2019 and added emergency department, ambulatory surgery, and hospital discharge records from the Office of Statewide Health Planning and Development (OSHPD), and 4th (2021) extended the year range from 2015 through 2020; the 5th (2022) extended it through 2021.

CALHHS Department	Program
Department of Developmental Services	Developmental Services Program
DHCS Department of Health Care Services	Family Planning, Access, Care, and Treatment (Family PACT) Program Medi-Cal Program
California Department of Public Health	Women, Infants, & Children (WIC) Program Vital Statistics Birth Records Vital Statistics Death Records
CDSS California Department of Social Services	Cal Fresh Program CalWORKs Program Child Welfare Services / Case Management System (CWS/CMS) In-Home Supportive Services (IHSS)
HCAI	Emergency Department Ambulatory Surgery Discharge

Procedures

Project Agreements

A Record Reconciliation Project Agreement (APPENDIX A.1) was established between CalHHS and the CDN for the 1st record reconciliation. The scope of work included the following activities:

(1) the extraction and secure transfer of records from CalHHS Departments to the CDN;
(2) the probabilistic de-duplication and linkage of client records by the CDN;
(3) the creation and secure delivery of an encrypted, client-level, between-program linkage key from the CDN to CalHHS Departments;
(4) the generation of an aggregated, de-identified demographic profile of clients served across multiple programs for return to CalHHS and subsequent dissemination.

This agreement was updated for the 2nd record reconciliation (APPENDIX A.2). Changes included:

(1) the inclusion of the Office of Statewide Health Planning and Development (OSHPD) as a signatory;
(2) the addition of Vital Statistics Birth and Death records from the California Department of Public Health (CDPH);
(3) modified amendment procedures so that existing CalHHS Departments do not need to re-sign when additional CalHHS Departments are included as signatories; and
(4) modified to reflect that once a signatory, each CalHHS Department has the authority to incorporate additional program data or record fields as part of the reconciliation effort without an amendment.

The agreement was amended for a second time on February 20th, 2020 (APPENDIX A.3). This amendment expressly allows the Children's Data Network to choose a vendor to build the RDH pilot and push the encrypted linkage keys generated from the RRs, as well as designated analytic data from CalHHS departments to the secure RDH environment. In addition, it extends the security requirements that existed between CalHHS and CDN to cover any data transfers between the CDN and the RDH. This amendment also extended the end date of the pilot to February 15, 2022. Amendment III (APPENDIX A.4) updates the agreement to reflect new agency and departmental names (i.e., changes the California Health and Human Services Agency acronym from CHHS to CalHHS, and changes the Office for Statewide Health Planning and Development (OSHPD) to the Department of Health Care Access and Information (HCAI)), and extends the end date of the pilot to February 15, 2023. Amendment IV (APPENDIX A.5) permits cross-agency data linkage and analysis for the purpose of AB2083, includes CDII as a party to the agreement, and allows CDN to access into the Agency Data Hub for the purpose of research and validation of record linkage performed by CDII in the Data Hub environment. Amendment V (APPENDIX A.6) updates the agreement to change the acronym of the USC-CDN to the CDN to align with approved human subjects protocols, removes the (non CalHHS) California Department of Education (CDE) as a party to the agreement at the conclusion of the AB2083 analysis, and changes extends the end date of the pilot to February 15, 2024.

Human Subjects, IRB, and VSAC Approvals

Original project protocols were approved by both the University of Southern California Institutional Review Board (IRB) and state Committee for the Protection of Human Subjects (CPHS). These protocols were amended for the 2nd record reconciliation. See CPHS Record Reconciliation Approval Letter Protocol ID 2018-080 (APPENDIX B.1). Vital Statistics Advisory Council approved the inclusion of vital birth and death records. Please refer to VSAC Data MOU_2019 (APPENDIX B.2) and VSAC Approval P2 - Putnam-Hornstein_E_19-03-0043 Approval Letter (APPENDIX B.3).

Client records concerning the administration of CalHHS programs are currently maintained across distinct administrative data systems. The unique identifier assigned to individual clients is specific to a given data system; there is no single client identifier common across CalHHS programs. As such, Personally Identifiable Information (PII) and personally identifiable Protected Health Information (PHI) was required to carry out records reconciliation and generate a common CalHHS CCI.

PII is defined as any information maintained by CalHHS and the Departments that can be used on its own or with other information to identify an individual. This includes, but is not limited to: (a) information that can be used to directly distinguish an individual's identity such as his or her name, social security number, date and place of birth, mother's maiden name, or home address; and (b) any other information that is linked or linkable to an individual. PHI is defined as personally identifiable information related to the past, present, or future physical or mental health condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual. PII / PHI was used solely for de-duplicating client records within a given program data file and linking client records across program data files.

Time Period

To generate a CalHHS CCI, each of the Departments extracted a defined set of PII /PHI concerning all clients / beneficiaries served during a designated time period. The 2nd record reconciliation concerned all clients / beneficiaries served during of the calendar years 2015, 2016, 2017, and 2018 (i.e., between January 1, 2015 and December 31, 2015, between January 1, 2016 and December 31, 2016, etc.). The 3rd, 4th, and 5th record reconciliation concerned all clients / beneficiaries served during of the calendar years 2019, 2020, and 2021, respectively (i.e., between January 1 and December 31 each year). Separate files were requested for each calendar year.

Data Elements

The following data elements were requested from each program / department:

Client ID (program specific)	Race/Ethnicity	Referral / Claim /
First Name	Sex	Other ID
Middle Name / Initial	Birthdate	Address Street
Last Name	Social Security Number	Number

Address Street Name	Address County	Service Start Date
Address Unit Number	Address State	Service End Date
Address City	Address Zip	Medi-Cal Number

Not all data elements were available for each program. Please see Record Reconciliation Data Elements (APPENDIX C) for a detailed list of data elements available in each program dataset. Many of the requested data elements represent Protected Health Information (PHI) – per the CPHS Protocol, PHI was used solely for de-duplicating client records within a given program data file and linking client records across program data files.

Data Transfer

Data for all clients served by each program during each separate calendar year (2015-2018) were transferred to the Children's Data Network via Secure File Transfer Protocol (SFTP). In accordance with CalHHS and CDN data security protocols, all program datasets were then transferred to two non-networked machines. One machine was used for cleanup and the other for database storage. These computers do not have wireless or wired connections to the Internet (i.e., they are "Air Gapped"). Each Air Gap machine is password-protected and can only be accessed in a secure location by authorized CDN personnel.

Data Cleaning and Hygiene Checks

Once the files were received and stored, the data underwent a series of procedures to clean, standardize, and organize client records into a SQL database. A SQL Database is a relational data base structure where files can be merged using Structured Query Language and common client identifiers.

For each data file, a unique client identifier, typically the program's internal client ID, was chosen as the key identifier. Unique client identifiers (ID) for each program are listed, as follows:

CALHHS Program	**Unique Client Identifier**
Birth Records	Birth state file number
Death Records	Death state file number
Developmental Services Program	Client_ID_Number
Family Planning, Access, Care, and Treatment (Family PACT) Program	HAP_ID
Medi-Cal Program	AKA_CIN
Women, Infants, & Children (WIC) Program	Individual_ID
Cal Fresh Program	CDSS_UID (Encrypted SSN)
CalWORKs Program	CDSS_UID (Encrypted SSN)
Child Welfare Services / Case Management System (CWS/CMS)	FKCLIENT_T
In-Home Supportive Services (IHSS)	CDSS_UID (Encrypted SSN)
OSHPD Emergency Department	PAT_ID / DATA_ID
OSHPD Ambulatory Surgery	PAT_ID / DATA_ID
OSHPD Discharge	PAT_ID / DATA_ID

Following the initial data transfer and file reading, analysts performed a series of data hygiene checks for records in each program dataset. As part of these checks, analysts documented the following information:

1. Transmitted file name
2. Transmission date
3. Transmission format
4. Total file size
5. Total file # of fields
6. Total number of records
7. Total number of records identifying a unique individual
8. Number and percentage of records with complete first name and last name fields
9. Number and percentage of records with complete DOBs
10. Number and percentage of records w/SSN field
11. Number and percentage of records with each (individual) address field completed
12. SSN distribution
13. Age at first and day of the observation period
14. Gender distribution
15. Ethnic distribution
16. Summary of fields that vary when unique client identifiers appear in duplicate

For each new dataset received, a data hygiene check including these 16 fields was returned to the respective CALHHS programs. As an external validity check, analysts also compared field #7 (the total number of unique individuals in each program) to published agency data. This information was communicated in an email that accompanied the hygiene check reports.

The 2016 data extracts were also compared for CALHHS programs that participated in the 1st record reconciliation. Because administrative data systems are not static, some change was expected between the 1st and 2nd reconciliation 2016 extracts. The results showed only minor differences.

Finding no major discrepancies, programs were then asked to confirm the accuracy of the information recorded for the transferred data ahead of linkage. All programs confirmed the accuracy of the counts. The total number of records and the total number of records identifying a unique individual in for each year in each of the participating CALHHS programs was, as follows:

CHHS Program	Record Type	2015	2016	2017	2018
Developmental Services Program	All		365,114	383,137	403,995
	Unique				
Medi-Cal Program	All	21,414,490	22,128,622	21,123,464	19,627,202
	Unique				
Family Planning, Access, Care, and Treatment (FPACT) Program	All	2,291,767	2,023,671	1,801,256	1,591,740
	Unique				
Women, Infants, & Children (WIC) Program	All	17,492,380		15,269,506	14,161,642
	Unique				
Vital Statistics Birth	All				
Vital Statistics Death	All				
Cal Fresh Program	All	5,955,517	5,808,335	5,600,217	5,256,461
	Unique				
CalWORKs Program	All	1,751,942	1,652,768	1,508,745	1,347,329
	Unique				
Child Welfare Services / Case Management System (CWS/CMS)	All	761,895	728,570	703,822	671,285
	Unique				
In-Home Supportive Services (IHSS)	All	574,929	603,626	629,103	649,595
	Unique				

Note: Information for the year 2019 and OSHPD Record information was not available at the time of drafting, but will be added in future versions.

Data Linkage

The project involved two separate data linkage processes. **Within-Program Matching / De Duplication** involved developing a routinized methodology for the large-scale de-duplication of records originating in different data sources using machine learning and probabilistic linkage algorithms. **Between-Program Match / Cross-Program Linkage** involved determining the lower and upper bound estimates of clients who are jointly or concurrently served by programs administered by CALHHS Departments.

Linkage Algorithm

ChoiceMaker, an open-source, machine-learning record linkage software, was used to link CALHHS program records. ChoiceMaker utilizes both probabilistic matching and modeling techniques for record linkage.

Model Development – The software compares selected fields of two records at a time. For each field in the pair of records, the software applies a set of logical instructions, called *clues*, to check whether the selected field (such as First Name) values point toward a decision. ChoiceMaker uses Match Clues, Differ Clues and Hold Clues. A collection of such clues is applied together as part of a single *model* for whether records match. After all the clues are evaluated,

ChoiceMaker assigns each clue a positive numerical value, which indicates its relative predictive significance. Based on a machine learning mathematical model called Maximum Entropy, the program produces a probability to describe the likelihood that the two records describe the same person (i.e., match).

Model Improvement – In order to ensure the quality of linkages, human reviewers then review a random sample of record pairs, and for each pair, they indicate whether the records should be categorized as a match, differ, or hold (not enough information). The manually marked sample is then returned to ChoiceMaker Analyzer, a module of the software. Using a machine-learning algorithm, ChoiceMaker then incorporates, or "learns", the human decisions and subsequently update the clue's original weights. This human training process may be repeated several times until researchers are satisfied with the ChoiceMaker's predictive output.

Using a machine-learning algorithm, the ChoiceMaker software then determines the clue weights that best reproduce these expert decisions. This process is called *training* a model. When a trained model is subsequently applied to completely different pairs, one finds that ChoiceMaker probabilities closely predict how a data expert would mark the new pairs.

Technical Procedure – ChoiceMaker is a standalone software which receives input from the designated SQL databases. After data hygiene checks and pre-processing, data from each agency was imported into its respective table in the database. This process also assigned a CDN_ID, an internal unique ChoiceMaker identifier to unique individuals in each dataset. ChoiceMaker then compared record pairs designated their CDN_IDs using the final mature model.

After the linkage process was successfully completed, an analyst combined the produced decisions for each pairs and other relevant variables into an extract, uniquely designated by an Extract Number. Analysts then utilized this extract to produce relevant statistics and ad-hoc reports.

Within-Program Match / De-Duplication

Once the linkage algorithm was constructed, ChoiceMaker was first configured to identify within-program matches, or identifiers from within a single program file that were probabilistically determined to represent the same individual, even though they were recorded as unique individuals. This information regarding data quality and client duplication was designed for internal program use. Such within-program matches typically reflect cases of duplicate records due to a missingness on a key identifier, or twin siblings.

If ChoiceMaker determined that an individual (unique client ID) to be a match with another individual (unique client ID) in the program file these records were flagged. Records with a .80 or greater match probability assigned by ChoiceMaker were coded as duplicates.

Between-Program Match / Cross-Program Linkage

ChoiceMaker was then configured to identify between-program matches. Specifically, for each pair of CHSS program datasets, probabilistic algorithms were used to assess the likelihood that an individual with a record in one program dataset was the same individual in a second program dataset.

If ChoiceMaker determined that an individual (unique client ID) to be a match with another individual (unique client ID) in the program file the record a Linkage Key was created and the match probability was recorded.

It is important to note that matches were not necessarily 1-to-1. For example, when linking clients from WIC and CWS/CMS, a client from WIC, represented by a unique client ID / WIC identifier, might be probabilistically linked by ChoiceMaker to two or more unique client IDs / CWS/CMS program identifiers. This could be due to: (a) a duplicate client record in CWS/CMS; or (b) two records in a given program that are probabilistically similar across a number of fields. In these cases, ChoiceMaker created and recorded two separate Linkage Keys and corresponding match probability records.

Pairwise Program Linkage Key

Once the inter program linkage process was completed, a unique pairwise Linkage Key was assigned to each record pair by ChoiceMaker. This identifier is an 8-digit, alpha-numeric field that can be utilized within agencies as a master Common Client Identifier (i.e., Linkage Key) to facilitate the exchange of statistical program information, both within and between individual CALHHS departments. Records with a .80 or greater match probability assigned by ChoiceMaker were retained as linkages.

Matching Results

Pairwise Program Match Statistics (APPENDIX D) details the distribution of match probabilities for each pairwise program match from ChoiceMaker.

Return Data File

Structure / Record Layout

For each program return data file, a “spine” consisting of all the unique client identifiers from the original program data file was created. For each unique client ID, information regarding both within- and between-program matches was provided.

Once this base file was established for each department / program, the files were customized in order to meet strict program confidentiality requirements. A unique file structure was developed that enabled programs with information regarding within program matching and de-duplication, within-program and -department matches, and linkage keys to all other CALHHS programs.

Return files included a main program level file (PROGRAM NAME_final_main) and three look-up files to identify duplicate clients (PROGRAM NAME_final_dup), as well as linked birth records (PROGRAM NAME_final_bsmf) and death records (PROGRAM NAME_final_dsmf). For each program the final_main file included information regarding several content areas including:

Data Dictionary

File Identifiers - Information regarding the data source and extract for data files returned to each agency was recorded. Specific variables included:

- Record_Source: Abbreviated name of data source.
- Extract_No: Number generated by CDN for linkage administrative purposes.

Client Identifiers - Matched data files returned to each agency were unduplicated at the client level. Specific variables included:

- Unique Client ID: *Unique Client ID* for PROGRAM NAME
- CDN_ID: Unique ID generated by CDN based on *Unique Client ID* for processing in ChoiceMaker
- PROGRAM NAME_UID: Unique Identifier for unduplicated records, populated when dup_yn =1

Annual Program Participation Indicators - Matched data files returned to each agency were unduplicated at the client level, binary indicators of annual program participation for each of the four analysis years were retained on each record. This allows agency analysts to examine annual cohorts. Specific variables included:

- in_2015: Binary variable. (1) if *Unique Client ID* has program participation in 2015, (0) otherwise
- In_2016: Binary variable. (1) if *Unique Client ID* has program participation in 2016, (0) otherwise
- in_2017: Binary variable. (1) if *Unique Client ID* has program participation in 2017, (0) otherwise
- in_2018: Binary variable. (1) if *Unique Client ID* has program participation in 2018, (0) otherwise

Client Level Information used in Record Linkage - Summary information regarding client level variables used in the linkage process was also returned. Specific variables included:

- SOCIAL_SEC_NBR_complete: Binary variable. (1) if *Unique Client ID* has at least SOCIAL_SEC_NBR (SSN) observation populated, (0) otherwise
- SOCIAL_SEC_NBR_valid*: Binary variable. (1) if *Unique Client ID* has at least 1 SSN observation valid, (0) otherwise.
- Last_Name_complete: Binary variable. (1) if *Unique Client ID* has at least 1 Last_Name observation populated, (0) otherwise
- Last_Name_valid: Binary variable. (1) if *Unique Client ID* has at least 1 valid Last_Name observation (not a placeholder name, such as 'UNKNOWN'), (0) otherwise

- First_Name_complete: Binary variable. (1) if *Unique Client ID* has at least 1 First_Name observation populated, (0) otherwise
- First_Name_valid: Binary variable. (1) if *Unique Client ID* has at least 1 valid First_Name observation (not a placeholder name, such as 'UNKNOWN'), (0) otherwise
- GENDER_complete: Binary variable. (1) if *Unique Client ID* has at least 1 GENDER observation populated, (0) otherwise
- RACE_ETH_complete: Binary variable. (1) if *Unique Client ID* has at least 1 RACE_ETH (race/ethnicity) observation populated, (0) otherwise
- DATE_OF_BIRTH_complete: Binary variable. (1) if *Unique Client ID* has at least 1 DATE_OF_BIRTH observation populated, (0) otherwise

** A valid SSN must contain all 9 numeric digits, with no group contains all 0 characters (e.g: 000-12-3456). The valid SSN must not be among those known for non-administrative purposes (078-05-1120, 111-11-1111, 123-45-6789, 219-09-9999, 999-99-9999) and cannot be between 987654320-987654329).*

Address Data **–** The availability of address fields differed by program. For a complete list by program please see Record Reconciliation Data Elements (APPENDIX C). The 2nd record reconciliation effort also included geocoded address data for MediCal data that was used for linkage where possible. Address data was used for preparation of ancillary data products including the CALHHS Data Dashboard. Specific variables included:

- ADDR_complete: Binary variable. (1) if *Unique Client ID* has at least 1 ADDR observation populated, (0) otherwise
- ADDR_CITY_complete: Binary variable. (1) if *Unique Client ID* has at least 1 ADDR_CITY observation populated, (0) otherwise
- ADDR_ZIP_complete: Binary variable. (1) if *Unique Client ID* has at least 1 ADDR_ZIP observation populated, (0) otherwise
- ADDR_ZIP_valid: Binary variable. (1) if *Unique Client ID* has at least 1 valid ADDR_ZIP observation (not a placeholder name, such as 'XXXXX'), (0) otherwise
- RESIDENCE_ZIP_valid: Binary variable. (1) if *Unique Client ID* has at least 1 valid RESIDENCE_ZIP observation (not a placeholder name, such as 'UNKNOWN'), (0) otherwise
- RESIDENCE_COUNTY_valid: Binary variable. (1) if *Unique Client ID* has at least 1 valid RESIDENCE_COUNTY observation (not a placeholder name, such as 'UNKNOWN'), (0) otherwise

Within-Program Match / De-Duplication - Within each program and for each unique client ID that was identified as a duplicate, a flag duplicate flag was provided for linkage to the PROGRAM NAME_final_dup file. The flag (dup_flag) (1) indicated if a record was probabilistically linked to another record in the PROGRAM_NAME (2 or more records are predicted to be the same individual based on available information), (0) otherwise. Records were identified as duplicates if the match probability was .80 or greater.

The duplicate look-up file look-up files ***PROGRAM NAME_final_dup*** contained duplicate IDs and match probabilities were provided. Within a specific program, the flags and associated information can be used for data quality checking. Specific variables included:

- Extract_No: Number generated by CDN for linkage administration purposes
- Record_source: Abbreviated name of data source. 'RR"year"PROGRAM_NAME'
- Unique Client ID: *Unique Client ID* for PROGRAM_NAME
- PROGRAM_NAME_UID: Unique Identifier for unduplicated records
- dup_*Unique Client ID_1*: *Unique Client ID* for other PROGRAM_NAME record that is probabilistically linked and identified as a duplicate
- dup_*Unique Client ID_2*: *Unique Client ID* for other PROGRAM_NAME record that is probabilistically linked and identified as a duplicate
- dup_*Unique Client ID_3*: *Unique Client ID* for other PROGRAM_NAME record that is probabilistically linked and identified as a duplicate
- dup_*Unique Client ID_n*: *Unique Client ID* for other PROGRAM_NAME record that is probabilistically linked and identified as a duplicate (up to n^{th} duplicate ID)

Within-Department and Cross-Department Linkage

Within-Department Matches - Two departments (i.e., CDSS and DHCS) submitted data for more than one program under their jurisdiction (i.e., CWS/CMS, CalWORKs, CalFresh & IHSS; Medi-Cal & FPACT, respectively). For these agencies, binary flags indicating within-department program matches were provided. For example, for a unique client ID in the CalWorks file, flags indicating matches with the CWS/CMS, CalFresh and IHSS programs were provided. For matches, the pairwise Linkage Key and match probability were provided. Data could then be directly exchanged between these programs using the pairwise Linkage Key. So, for instance, within CDSS, analysts from CalWorks could select all IHSS matches, or a subset of matches, using the IHSS_YN match flag, and send the pairwise Linkage Keys for those matches to IHSS in order to exchange data.

Cross-Department Matches - For pairwise matches across departments, binary program flags are not provided so that matches cannot be directly observed. For all matches, pairwise Linkage Keys for each record and associated match probabilities were provided. For non-matches, an algorithm was used to create synthetic (orphan) pairwise Linkage Keys and match probabilities. In this way, programs cannot identify matches without exchanging Linkage Keys and receiving return matches from a specific program. So, for instance, if DDS was interested in exchanging data with CalFresh they would provide the DDS/CalFresh Linkage Keys for all clients of interest. CalFresh would then match these IDs to their data using the pairwise Linkage Key and then return the data for client's who matched.

The CalHHS Intra-Agency Global Data Sharing Agreement allows for this program-to-program exchange of data. The Linkage Key facilitates this linkage. Specific variables include:

- CalFresh_Linkage_Key: Pairwise Linkage Key which identifies Unique *Client ID* that is probabilistically linked to another *Unique Client ID* in CalFresh

- CalFresh_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in CalFresh
- *CalWorks_flag: Binary variable. (1) if *Unique Client ID* is linked to another *Unique Client ID* in the CalWorks dataset, (0) otherwise
- CalWorks_Linkage_Key: Pairwise Linkage Key which identifies *Unique Client ID* that is probabilistically linked to another *Unique Client ID* in CalWorks
- CalWorks_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in CalWorks
- *CWS_CMS_flag: Binary variable. (1) if *Unique Client ID* is linked to another *Unique Client ID* in the CWS/CMS dataset, (0) otherwise
- CWS_CMS_Linkage_Key: Pairwise Linkage Key which identifies *Unique Client ID* that is probabilistically linked to another *Unique Client ID* in CWS/CMS
- CWS_CMS_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in CWS/CMS
- DDS_Linkage_Key: Pairwise Linkage Key which identifies *Unique Client ID* that is probabilistically linked to another *Unique Client ID* in DDS
- DDS_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in DDS
- *IHSS_flag: Binary variable. (1) if *Unique Client ID* is linked to another *Unique Client ID* in the IHSS dataset, (0) otherwise
- IHSS_Linkage_Key: Pairwise Linkage Key which identifies *Unique Client ID* that is probabilistically linked to another *Unique Client ID* in IHSS
- IHSS_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in IHSS
- *MediCal_flag: Binary variable. (1) if *Unique Client ID* is linked to another *Unique Client ID* in the MediCal dataset, (0) otherwise
- MediCal_Linkage_Key: Pairwise Linkage Key which identifies *Unique Client ID* that is probabilistically linked to another *Unique Client ID* in MediCal
- MediCal_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in MediCal
- WIC_Linkage_Key: Pairwise Linkage Key which identifies Unique *Client ID* that is probabilistically linked to another *Unique Client ID* in WIC
- WIC_match_prob: ChoiceMaker probability that the *Unique Client ID* is linked to another *Unique Client ID* in WIC

*Pairwise Match Fields – Inclusion of variables marked *program-dependent*

Birth and Death Records **-** Although vital statistics birth and death records were matched using the same protocols as other agency data, because they involve one-time events the cross-program indicators constructed were different. Specifically, birth and death matches were given "within-department match status" for all programs, therefore a binary flag indicating a match with birth or death was provided within every agency/ program file.

For records with the flag indicating a record match, the PROGRAM_NAME_Unique_ID could be use to link to the birth (**PROGRAM_NAME_final_bmsf**) and death (**PROGRAM_NAME_final_dmsf**) look-up files for more information about these events. In addition to linkage match probabilities, each file also includes information regarding the year of the event, as well as the state file and local registration numbers. The birth file also includes variable indicating the client's role (1=Self/Child, 2=Parent 1 (Mother), 3=Parent 2 (Father)) on the birth record. Specific variables include:

PROGRAM NAME_final_bsmf

- Extract_No: Number generated by CDN for linkage administration purposes
- Record_source: Abbreviated name of data source. 'RR"year"PROGRAM_NAME'
- Unique Client ID: *Unique Client ID* for PROGRAM NAME
- BMSF_Match_Probability: ChoiceMaker Probability for linkage to a record in Birth Statistical Master File (BSMF)
- Bthrole: BIRTH Role (1=Self/Child, 2=Parent 1 (Mother), 3=Parent 2 (Father))
- Bthyear: BSMF File Year
- STATE FILE NUMBER: BSMF State File Number
- LOCAL REGISTRAR NUMBER: BSMF Local Registrar Number

PROGRAM NAME_final_dsmf

- Extract_No: Number generated by CDN for linkage administration purposes
- Record_source: Abbreviated name of data source. 'RR"year"PROGRAM_NAME'
- Unique Client ID: *Unique Client ID* for PROGRAM NAME
- DMSF_Match_Probability: ChoiceMaker Probability for linkage to a record in Death Statistical Master File (DSMF)
- Dthyear: DSMF File Year
- STATE FILE NUMBER: DSMF State File Number
- LOCAL REGISTRAR NUMBER: DSMF Local Registration Number

File Return

Prior to returning linked files back to participating programs with the pairwise program-specific Linkage Keys attached, a detailed program-specific data dictionary and sample data file was shared with each program. Feedback regarding the file format and additional data needs was solicited. DHCS requested that both forms of the Common Client Identifier be included in the returned files. No other significant changes were requested by participating agencies.

The data files were processed and returned to agencies via SFTP. A detailed Linkage Guide was sent, which included step-by-step instructions for programs to link data with one another using the Linkage Key. This guide is attached as Step-By-Step Linkage Guide (APPENDIX E).

For the 3rd and subsequent Record Reconciliations, the CDN has permission to securely return linkage keys + source record IDs to respective departmental folders on the RDH per the RR Project Agreements and associated amendments, BAA, and DSA (APPENDICES A.1-5).

Confidentiality Measures

Once the Linkage Key was created, in accordance with confidentiality protocols, the research team removed all confidential identifiers transferred by the agencies in the original files and replaced them with a CDN_ID – the unique ID assigned by ChoiceMaker for processing.

Appendices

APPENDIX A.1 - Record Reconciliation Project Agreement
APPENDIX A.2 - RR Project Amendment I
APPENDIX A.3 - RR Project Amendment II
APPENDIX A.4 - RR Project Amendment III
APPENDIX A.5 - RR Project Amendment IV
APPENDIX A.6 - RR Project Amendment V
APPENDIX B.1 - CPHS Record Reconciliation Continuing Approval Letter Protocol ID 2018-080
APPENDIX B.2 - VSAC Data MOU
APPENDIX B.3 - VSAC Approval P2 - Putnam-Hornstein_E_19-03-0043ApprovalLetter
APPENDIX C - Record Reconciliation Data Elements
APPENDIX D - Pairwise Program Match Statistics
APPENDIX E - Step-By-Step Linkage Guide