APPENDIX C. METHODS

Methods

Project Agreements

In order to accomplish this legislatively-mandated analysis, the AB 2083 Children and Youth System of Care State Joint Interagency Resolution Team (JRT) (i.e., the AB 2083 Team) leveraged data governance established within four (i.e., two existing and two new) agreements:

1. No changes were needed to an ***existing CDSS <> CDN Research Agreement*** that permits CDN to receive CWS records for the purposes of cross-program/department/agency linkage and research/evaluation.
2. An amendment was needed to the ***existing CalHHS <> CDN Record Reconciliation Agreement (RRA), as amended*** (*Record Reconciliation Project Agreement*, *Amendment I*, *Amendment II*, *and Amendment III*). This agreement permits CDN to receive (any) data from each department for the purposes of linkage with other CalHHS departments and cross-program analysis. It also permits the creation and secure transfer of records from the departments to the CDN; de-duplication and probabilistic linkage of client and beneficiary records by the CDN; creation and secure delivery of a common, encrypted client identifier from the CDN to the Departments; and the generation of aggregated demographic profiles of clients served across multiple programs disaggregated by client demographics and characteristics including age, race/ethnicity, sex at birth, and county. The ***new amendment (Amendment IV)*** permits cross-agency data linkage and analysis, includes the Center for Data Insights and Innovation (CDII) as a party to the agreement, and allows the CDN to access into the Agency Data Hub for validation purposes. Please see *California Health and Human Services Agency (CalHHS) DRAFT Record Reconciliation Methodology*, which includes each agreement and amendment.
3. A ***new CDE <> CDN Data Use Agreement (DUA)*** was needed. This new DUA permits CDN to receive statewide Public School (PS) records from the California Department of Education (CDE); link those records to child protection (Child Welfare Services/Case Management System (CWS/CMS)) records under the jurisdiction of the California Department of Social Services

(CDSS), Medi-Cal records under the jurisdiction of the California Department of Health Care Services (DHCS), and Developmental Service records under the jurisdiction of the California Department of Developmental Services (DDS); and meet reporting requirements included in AB 2083 and AB 153. This agreement authorized CDE to disclose to CDN a onetime data pull of information for the observation period included in this report.

4. A ***new CalHHS Business Use Case Proposal (BUCP)*** was needed. This new BUCP permits CDN to analyze CalHHS records for the purposes of AB 2083.

Defining the Target Population

For the purposes of this analysis, the target population (i.e., Children and youth in foster care who have experienced severe trauma) was operationalized as: ***Children and youth age 0 to 21 years with an open foster care placement at some point during the observation window (2020/21 academic year) as documented in statewide child protection records (i.e., CWS/CMS)*** and will hereafter be defined as the Foster Care (FC) population. The FC population was considered the 'analytic spine,' the focal point to which Developmental Services, Medi-Cal, and California Public School records were each probabilistically matched and analyzed.

The CDN, in close consultation with the AB 2083 Team and state departmental staff, identified key variables and outcomes of interest pertaining to the FC population, including demographic characteristics, cross-program service interactions, and distribution across California counties. Once finalized, these key variables were incorporated into a set of table shells that, once populated, would answer the research questions.

Data Sources

Department of Social Services (CDSS) child welfare data are sourced from the CWS/CMS. Categories include: Child demographics, current and prior out of home placement and case characteristics, as well as child abuse and neglect allegation history.

Department of Health Care Services (DHCS) Medi-Cal eligibility and claims data are sourced from the Management Information System/Decision Support System (MIS/DSS). Categories include: Mental and behavioral health service utilization, inpatient and outpatient services, Adverse Childhood Experience Scores (ACES), and Healthcare Effectiveness Data and Information Set (HEDIS) quality measures.

Department of Developmental Services (DDS) data are sourced from the Client Master File, Client Development Evaluation Report (CDER), and Purchase of Service data. Categories include: Qualifying conditions; medical, physical, and behavioral challenges; Regional Center status; residence information; and services and supports.

California Department of Education (CDE) data are sourced from the California Longitudinal Pupil Achievement Data System (CALPADS). Categories include: Student demographics, enrollment, attendance, and discipline information for both special education and traditional public-school students who were enrolled in California public schools between July 1, 2020 and June 30, 2020 and who youth who were youth age 5 by June 29, 2021. More information can be found on the CDE Data & Statistics (CA Dept of Education) webpage.

Each of the above agencies serves youth of different ages and each of their respective data systems are subject to their own timing and rules for data reporting.

File Contents

Each department was responsible for extracting and securely transmitting three types of information:

1. *Information About Program Participants*: In order to permit cross-program linkage, departments were asked to submit a linkage file that included ONLY direct identifiers for participants with a service interaction between July 1, 2020 and June 30, 2021 and who were age eligible for at least one day during that time period (i.e., 0 through their 21st birthday in 2020/2021[1]). Direct identifiers for program participants are the minimum level of data necessary in order to answer this legislatively mandated analysis. Requested data elements included: Program-specific Client IDs, Names, Referral/Claim IDs, Address, Race/Ethnicity, Sex at Birth, DOB, and Service Start and End Date. Not all data elements were available for each program. Many of the requested data elements represent Personally Identifiable Information (PII)/Protected Health Information (PHI). Per California Committee for the Protection of Human Subjects (CPHS) Protocol, PII/PHI was used solely for de-duplicating client records within a given program data file and linking client records across program data files.

[1] The CDE provided data included students who were eligible and age 5 by 6/29/2021 as CDE Does not serve the full age range of the population included in CWS/CMS.

CalHHS departments: Due to their ongoing participation in annual CalHHS Record Reconciliations,[2] encrypted cross-program linkage keys had already been generated among 2020 and 2021 DHCS, DDS, and CDSS records (Please see *the CDN website* for more details about the RR linkage methodology). As no additional PII was needed to create the intra-agency linkage keys, CalHHS departments were not required to securely transmit direct identifiers for the purposes of the analysis. CDN analysts subsequently subset each department's existing linkage keys to the appropriate age group and service delivery window using analytic data provided by CalHHS departments.

CDE: CDE analysts securely transmitted PII for all students who were enrolled in California public schools between July 1, 2020 and June 30, 2021 and who were age eligible for at least one day during that time period (i.e., students who were age 5 by 6/29/2021).

Identifiable data were transferred to the CDN via a Secure File Transfer Protocol (SFTP). In accordance with CalHHS, CDE, and CDN data security protocols, all datasets were then transferred to the CDN Data Lab, the highly physically and technologically secure environment in which probabilistic linkage occurs.

PLEASE NOTE: After linkage at the CDN Data Lab, unique, encrypted pairwise linkage keys are assigned to each record pair and all direct identifiers are stripped from the file. The resulting confidential limited research datasets (i.e., containing only the linkage keys and source record identifiers) are securely transferred to the CDN's secure analytic server. There, they can be merged with cross-department client-level analytic information and made available to the project's authorized researchers for analysis.

2. *Information About Program Participation*: In order to adhere to the separation principle, departments were also asked to provide Analytic Files that did NOT include direct identifiers, only mutually agreed-upon dimensions of interest and an individual-level source record identifier so that analytic information could be merged with encrypted linkage keys for cross-program analyses. Variables were coded per departmental preferences and described in associated data dictionaries.

[2] See CDN/CalHHS Record Reconciliations and Research Data Hub

Analytic files were provided for participants meeting age and service interaction criteria – clients age 0-21 years with a service interaction between July 1, 2020 and June 30, 2021 and who were age eligible for at least one day during that time period (i.e., 0 through their 21st birthday in 2020/2021)[3]. Files included one record per client.

In accordance with CalHHS, CDE, and CDN data security protocols, analytic data were transferred to the CDN's secure analytic server via a Secure File Transfer Protocol (SFTP). There, analytic information and encrypted linkage keys could be combined to create cross-department client-level datasets and made available to authorized researchers for analysis.

3. *Information About the Data*: Each CalHHS department, and the CDE, shared a detailed data dictionary for variables transmitted. Please refer to *APPENDIX B. Data Dictionaries*.

Data Linkage

This project involved the linkage of Developmental Services, Medi-Cal, and Public School records to the 'analytic spine' (i.e., Children and youth age 0 through their 21st birthday with an open foster care placement at some point during the observation window (2020/21 academic year) as documented in statewide child protection (i.e., CWS/CMS) records.

Again, due to their ongoing participation in annual CalHHS Record Reconciliations,[5] encrypted cross-program linkage keys had already been generated among DHCS, DDS, and CDSS records (Please see *the CDN website* for more details about the RR linkage methodology). Therefore, linkage for this analysis involved the additional integration of Public School (PS) records for students enrolled during academic year 2020/21 with CWS/ CMS. The linkage process is described below.

Data Preparation

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

[3] The CDE provided data for students who were enrolled in California public schools, grades K – 12 (inclusive of transitional kindergarten) between July 1, 2020 and June 30, 2021, who were age 5 by 6/29/202.

For each data file, a unique client identifier, typically the program's internal client ID, was chosen as the key identifier. Following the initial data transfer and file reading, analysts performed a series of data hygiene checks for records in each program dataset.[4] These hygiene checks not only help analysts confirm that all data have been received, but also that the data have been read into and are being displayed appropriately by analytic software packages.

Linkage Process

Linkage largely involved separate data linkage processes. **Within-Program Matching/De-Duplication** involved developing a routinized methodology for the large-scale de-duplication of records originating from a single data source 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 CDSS and DDS, DHCS, or CDE.

An open-source, machine-learning record linkage software, was used to link CalHHS program records. The algorithm developed for the CalHHS Record Reconciliation process utilizes both probabilistic matching and modeling techniques for record linkage. A summary of the original model development and improvement process is provided below. For more detailed information please see Please see *the CDN website*.

> *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 *signals*, to check whether the selected field (such as First Name) values point toward a decision. The algorithm uses Match Signals, Differ Signals and Hold Signals. A collection of such signals is applied together as part of a single *model* for whether records match. After all the signals are evaluated, the algorithm assigns each signal a positive numerical value, which indicates its relative predictive significance. Based on a machine learning mathematical model called Maximum Entropy, the program

[4] Hygiene checks involve documentation of key information, including: transmitted file name, transmission date, transmission format, total file size, total file # of fields, total number of records, total number of records identifying a unique individual, number and percentage of records with complete first name and last name fields, number and percentage of records with complete DOBs, number and percentage of records w/SSN field, number and percentage of records with each (individual) address field completed, SSN distribution, age at first and day of the observation period, sex at birth distribution, ethnic distribution, and summary of fields that vary when unique client identifiers appear in duplicate.

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 (i.e., not enough information). The manually marked sample is then returned to a separate module of the software, where the model then incorporates, or "learns," the human decisions and subsequently updates the signal's original weights. This human training process may be repeated several times until researchers are satisfied with the algorithm's predictive output.

Using a machine-learning algorithm, the software then determines the signal 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 algorithm probabilities closely predict how a data expert would mark the new pairs.

Technical Procedure. After data hygiene checks and pre-processing, data from each department were imported into its respective table in the SQL database. This process also assigned a CDN_ID, an internal unique identifier to unique individuals in each dataset. We then compared record pairs designated their CDN_IDs using the final mature model.

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

Within-Program Match/De-Duplication

Using the linkage algorithm, we first identify within-program matches, or identifiers from within a single program file that are probabilistically determined to represent the same individual, even though they are recorded as unique individuals. Such within-program matches typically reflect cases of duplicate records due to a missing value on a key identifier, or twin siblings. If the algorithm determines that an individual (i.e., unique client ID) to be a match with another individual (i.e., unique

client ID) in the program file, these records are flagged. Records with a .80 or greater match probability assigned by the algorithm are coded as duplicates.

Between-Program Match/Cross-Program Linkage

The algorithm then identifies between-program matches. Specifically, for each pair of CalHHS program datasets, probabilistic algorithms are used to assess the likelihood that an individual with a record in one program dataset is the same individual in a second program dataset. If the algorithm determines that an individual (i.e., unique client ID) is a match with another individual (i.e., unique client ID) in the program file, a linkage keys is created and the match probability is recorded.

It is important to note that matches were not necessarily 1-to-1. For example, when linking clients from CDSS to CDE, a client from CDSS, represented by a unique client ID/CWS/CMS identifier, might be probabilistically linked by the algorithm to two or more unique student records IDs/CDE program identifiers. This could be due to: (a) a duplicate client/student record; or (b) two records in a given program that are probabilistically similar across a number of fields. In these cases, we create and record two separate Linkage Keys and corresponding match probability records.[5]

Pairwise Program Linkage Key

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

After linkage at the CDN Data Lab, unique, encrypted pairwise linkage keys were assigned to each record pair and all direct identifiers were stripped from the file. The resulting confidential limited research datasets (i.e., containing only the linkage keys and source record identifiers) were securely transferred to the CDN's secure analytic server. There, they were merged with cross-department client-level analytic information and made available to authorized project researchers for analysis.

[5] The CDE provided data for students who were age 5 by 6/29/2021 rather than inclusive of ages 0 – 3

Analytic File Production

Variable Coding

The AB 2083 Team consisting of state departmental staff and subject matter experts, identified key variables and outcomes of interest pertaining to the FC population, including demographic characteristics, cross-program service interactions, and distribution across California counties. Each of these key variables were incorporated into a set of table shells that, once populated, would answer the research questions. Working backward from the table shells and in close partnership with program analysts, CDN analysts processed and cleaned each program's analytic file. Formats were created for dimensions consistent with definitions shared in program specific data dictionaries (*APPENDIX B*). Then, they developed code to populate each cell of the table shells.

Defining the Analytic Spine

The Quarter 1, 2022 California CWS/CMS Data Extract was used to select all children ages 0-21 as of July 1, 2020 who were in an out of home placement at some point during the 2020/2021 Academic Year (i.e., between July 1, 2020 and June 30, 2021). This includes all clients in care at the start of the study period, plus any who entered care before the end of the period.[6]

Out of home placement episodes (i.e., episodes in care) where the agency legally responsible for the placement, care and supervision of the child included county child welfare or county probation agencies. If a child was no longer under the jurisdiction of the court, including Private Adoptions Agencies, or Kin-Gap they were excluded. Clients with missing birth dates were excluded. Additionally, those whose placement episode started on or after their 21st birthday were also excluded.

Individuals meeting the criteria for inclusion in the FC population became the "spine." CWS (i.e., analytic) information for individuals appearing in the spine population was merged into the research dataset using the CDSS source record identifier.

Demographic and placement characteristics as well as child abuse and neglect allegation history were coded for all clients. For youth who remained in care at the

[6] Because the denominator includes all children served during the academic year, the proportions reported may not be consistent with more standard point-in-time estimates.

end of the study period, only placements up until June 30, 2021 were included. If a client had more than one placement episode during the study period, the last was selected. Similarly, within a placement episode, if the client had more than one out of home placement, the most recent placement relative to the end of the study period was chosen. Only child abuse and neglect referrals received through the end of the study period were included.

Creating Cross-Program Dataset

To create the cross-program dataset, pairwise linkages between CWS/CMS and the AB 2083 involved departments (DDS, DHCS, CDE) data were used to determine which children from the "analytic spine" (i.e., age 0-21 in an out of home placement at some point during the study period) were also eligible for or enrolled in from DDS, DHCS, CDE during the academic year. If matched, the program-specific source record identifier for that client was merged onto to the "analytic spine" file. This sequence resulted in an individual-level dataset (i.e., the final research dataset) about the FC population without individual identifiers, but with information about participation in DDS, Medi-Cal, and PS. If a CWS/CMS record from the analytic spine matched with more than one client record in the program data set, the matched record with the highest match probability was selected.

Information on the program specific report dimensions can be found in the Data Dictionary (*APPENDIX B*). This dictionary details each dimension, including definitions and descriptions, source data, formats, and special considerations.

Data Analysis

Analysts conformed to CalHHS de-identification guidelines,[6] which requires both primary masking of any cells less than or equal to 10 and secondary masking of complementary cells to prevent re-calculation. Additionally, they periodically checked the initial results against data that is publicly available,[7–11] data published in academic research papers, and with program analysts and leadership.

Results

The analysis had three main goals. Goal 1 was to develop descriptive statistics for the FC population and its subgroups, disaggregated by important demographic and CWS characteristics. Goal 2 was to characterize the FC population's DDS, Medi-Cal,

and Public School[7] involvement. Goal 3 was to document the distribution of the FC population across California counties.

Record Reconciliations Methodology

More information and products related to the work and partnership between CalHHS, CDE and the USC Children's Data Network to utilize "record reconciliations" that link and organize administrative, client-level records across major programs and data sources can be found on CDN's webpage: CDN/CalHHS Record Reconciliations and Research Data Hub.

[7] Includes non-public school settings (NPS)