By Emily Putnam-Hornstein, Mark Ghaly, and Michael Wilkening

# Integrating Data To Advance Research, Operations, And Client-Centered Services In California

DOI: 10.1377/hlthaff.2019.01752
HEALTH AFFAIRS 39,
NO. 4 (2020): 655–661


**ABSTRACT The value of using administrative records for operational and evaluation purposes has been well established in health and human services. However, these records typically reflect the reach of a single government agency or program and fail to capture the experiences of individuals as they engage with different agencies or programs over time. Thus, the potential for these data to improve everyday operations, coordinate services, develop targeted interventions, and advance the science behind broader social policies has yet to be fully realized. A first step toward realizing that potential is to transition from an agency-centered to a client- or person-centered organization of data. We systematically linked tens of millions of records across California's largest health and human services programs. Our results underscore how the integration of records can help shift discussions from the programs that administer services to the people who are served.**

**Emily Putnam-Hornstein** (ehornste@usc.edu) is director of the Children's Data Network, University of Southern California, in Los Angeles.

**Mark Ghaly** is secretary of the California Health and Human Services Agency, in Sacramento.

**Michael Wilkening** is special adviser on innovation and digital services, Office of Gov. Gavin Newsom, in Sacramento.

Social and economic factors distinct from medical care are powerful predictors of health outcomes and disease burden throughout a person's life.[1,2] From a population health perspective, this means that evidence-based policies that affect the broader conditions in which people are born, grow, and live can exert a powerful influence on health and well-being.[2,3] From an operational perspective, data-driven efforts to better coordinate human and social supports with the medical and health care sectors provide opportunities to deliver services that are more client centered, efficient, effective, and tailored.[4]

For these reasons, there has been broad interest in ways to enhance government agencies' ability to systematically assemble, securely share, and responsibly use administrative client data.[4–6] Numerous initiatives have emerged from philanthropic organizations, with funding for activities that range from advancing the rigorous testing of interventions and policies via randomized trials[7] to improving the use of evidence through research-practice partnerships between universities and the public sector.[8] Other efforts have received federal funding to develop statewide longitudinal data systems focused on education and the workforce.[9] Still other efforts have received private and public funding to promote the use of social impact bonds or "pay-for-success" models.[10]

Unfortunately, the potential of integrated data remains largely unrealized, to the detriment of both clients and communities. While there are numerous reasons why this work has not progressed more quickly,[11] none are insurmountable. In this article we describe the process by which the California Health and Human Services Agency (CHHS) partnered with university-based researchers to carry out its first agencywide data integration effort, which resulted in the linkage of more than thirty million client records. Based on these linkages, we present examples of the cross-program, client-centered insights that can be produced, and we describe the next steps for sustaining and expanding this effort.

## Background

CHHS is a public agency that consists of twelve statewide departments, five offices, and various boards and commissions.[12] Collectively, CHHS invests significant resources in the delivery of programs designed to address negative social determinants for the state's most vulnerable and at-risk residents through the delivery of both short- and long-term public benefits—including food assistance, child care, health care coverage, housing support, employment support, child support services, child welfare, and many more. California's 2019 budget act allocated $163 billion ($41.9 billion from the general fund and $121.1 billion from other funds) for all health and human services programs.[13] The agencies and departments that administer these programs collect rich data about the characteristics of their clients. Statistical information derived from these client records can be an important way to inform program planning and accountability, while also driving improvement initiatives.[14–16]

Nevertheless, isolated program data are a blunt instrument for policy development and service coordination.[16] While each program captures data concerning discrete client encounters, typically absent is information concerning concurrent services and benefits that the same individual or family may have received through other CHHS programs. Also missing are data organized to document the timing, sequencing, and outcomes of service and program encounters both within and across departments. The absence of records integrated at the client level across CHHS programs limits the understanding of the collective size and impact of investments in public benefits, and it prevents a full assessment of population needs so that available resources can be strategically coordinated to reduce inequality. Because records are not integrated across programs, insights about client outcomes can be understood only through the lens of a single program, even though the client might well have received services from multiple programs.

Given the complex nature of CHHS's operational, fiscal, and regulatory commitments, this "program-centric" design of data collection increasingly impedes administration and planning. Further complicating efforts, records are currently maintained across distinct data systems using unique client identification keys that are assigned program by program: There is no universal or common client identifier captured across CHHS programs. Fortunately, advances in machine learning and probabilistic matching techniques have facilitated increasingly rigorous, accurate, and efficient ways for government agencies to connect client records to support the design and administration of large-scale programs.[17–19]

In 2017 CHHS partnered with researchers at the Children's Data Network at the University of Southern California to pilot an agencywide effort to systematically integrate, organize, and analyze administrative client records. The effort was conceptualized as a "record reconciliation"; the goal was both to demonstrate the feasibility of linking tens of millions of records quickly, accurately, securely, and cost-effectively and to facilitate the cross-program and cross-departmental exchange of statistical information about common clients.[20] This initial pilot was based on records from 2016 and led to the creation of encrypted linkage keys that connected client-level records across eight of CHHS's largest health and human services programs, from food stamps and public reproductive health programs to child welfare services. In 2019 this pilot was extended to incorporate records from additional years (2015–18) and to include vital birth and death records. Agreements were also signed with the CHHS Office of Statewide Health Planning and Development to additionally integrate emergency department, ambulatory surgery, and hospitalization records with the other data.

## Study Data And Methods

**DATA AGREEMENTS** Two key CHHS data sharing agreements govern data integration activities for research and operational purposes. First, an intra-agency data exchange agreement covers the exchange of data among departments within CHHS in compliance with all applicable federal, state, and local laws, regulations, and policies.[21] As the sole agreement for data exchange among CHHS departments, it eliminates the need for the departments to enter into "point-to-point" agreements except where an alternative agreement is required by the federal government or federal law. Second, to carry out record linkages and produce curated data sets, an interagency data sharing agreement was signed by CHHS, participating departments, and the Children's Data Network.

**DATA** For the pilot, we integrated the records of individuals eligible for services from a CHHS program in the period January 1, 2015–December 31, 2018. We additionally linked information concerning birth (children born and the legal parents associated with the birth) and death events, as recorded in California's vital records. Analysts from each CHHS program extracted a defined set of personally identifiable information concerning individuals who were eligible for services for at least one month in the period

2015–18. Records for developmental services were not available for 2015, and vital birth and death records were available only for 2015–17. *Personally identifiable information* was defined as any information maintained by CHHS or its departments that could be used on its own or with other information to identify an individual. Data elements used for linkage purposes included both unique (such as Social Security number) and nonunique (for example, first and last names) fields. Personally identifiable information was used solely for deduplicating client records within a given program data file and linking client records across program data files. Exhibit 1 provides a list of participating departments, along with program descriptions.

**RECORD TRANSFER** Records were extracted, encrypted, and then transmitted by individual programs within CHHS departments to the Children's Data Network. Some programs transferred a file already assembled to reflect a calendar year cohort (for example, all unique individuals eligible in a given year), while other programs transferred files that reflected monthly logs of eligible clients. In accordance with data security protocols, all program data sets were processed on a dedicated, non-networked server. Once the information was decrypted, a series of procedures were used to clean, standardize, and organize records into a Structured Query Language (SQL) database. Record-level identifiers were assigned as a way to inventory transferred information. A within-program client identifier—typically the program's internal alphanumeric client key—was documented and retained. Given that birth records could contain

**EXHIBIT 1**

**Departments of the California Health and Human Services Agency and sources of data that were included in the record reconciliation pilot**

| Department | Program or other source | Description |
|---|---|---|
| California Department of Social Services | CalFresh | Known federally as the Supplemental Nutrition Assistance Program (SNAP), CalFresh provides monthly food benefits to low-income individuals and families and economic benefits to communities. |
| | CalWORKs | Known federally as Temporary Aid to Needy Families (TANF), CalWORKs is a welfare program that gives cash aid and services to eligible California families. |
| | Child Welfare Services | Child Welfare Services is California's program for child protection and associated foster care services and preventive interventions. |
| | IHSS | IHSS provides in-home assistance to eligible aged, blind, or disabled people as an alternative to out-of-home care and enables recipients to remain safely in their own homes. |
| California Department of Developmental Services | Developmental services | This department is the agency through which California provides services and supports to people with developmental disabilities, including intellectual disabilities, cerebral palsy, epilepsy, autism, and related conditions. |
| California Department of Health Care Services | Medi-Cal | Medi-Cal is California's Medicaid program. This public health insurance program provides needed health care services for low-income people, including families with children, seniors, people with disabilities, pregnant women, and low-income people with specific diseases. |
| | Family PACT | Family PACT provides comprehensive family planning education, assistance, and services to low-income Californians of childbearing age. |
| California Department of Public Health | WIC | WIC provides nutrition education and counseling; breast-feeding support; referrals to health care and other community resources; and vouchers for families to purchase specific foods that provide key nutrients needed by pregnant and breast-feeding women, infants, and young children. |
| California Department of Public Health, Center for Health Statistics and Informatics and State Registrar | Vital birth and death records | Vital birth and death events are recorded via the state's registration process. The center is responsible for compiling registered information. The Vital Statistics Advisory Committee ensures that all research using vital statistics is consistent with the guidelines provided by the center and satisfies state statutes governing the use of these data. |
| Office of Statewide Health Planning and Development | ED, patient discharge, ambulatory surgery records | The office manages the collection and provision of out- and inpatient encounters in California-licensed hospitals and clinics for approved research and program operations. |

**SOURCE** Authors' analysis of documentation from the record reconciliation pilot. **NOTES** Data from the Office of Statewide Health Planning and Development were not reported in this study because those records had not yet been linked to other CHHS program data. CalWORKs is California Work Opportunity and Responsibility to Kids. IHHS is In-Home Supportive Services. Family PACT is Family Planning, Access, Care, and Treatment. WIC is Special Supplemental Nutrition Program for Women, Infants, and Children.

information for three people, every record was split into person-specific records: one each for the child who was born; the mother who gave birth; and the father or second legal parent, if named.

**RECORD LINKAGE MODEL** We used an open-source, machine-learning record linkage software program, ChoiceMaker (version 2.7.1), for both within-program matching (or deduplication) and between-program linkages (such as linking records from the Special Supplemental Nutrition Program for Women, Infants, and Children [WIC] to child welfare records). ChoiceMaker employs probabilistic matching and modeling techniques for record linkage.[22] To develop the record linkage model, data scientists at the Children's Data Network 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 that was 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,[23,24] an overall probability is generated to describe the likelihood that two records describe the same person (that is, a match likelihood).

To support an iterative model development process, samples of record pairs were systematically extracted for clerical review. For each record pair, a reviewer determined whether the records should be categorized as referring to the same person (they matched), two different people (they differed), or a hold (not enough information). Manually marked record pairs were 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 was subsequently applied to new record pairs, we found that ChoiceMaker probabilities closely predicted how a human expert would mark those records.

**LINKAGES** The algorithm was first configured to identify within-program matches, or records from a single program file that were probabilistically determined to represent the same person—even though they were recorded under different source client keys. Records with at least an 80 percent probability of being a match were coded as duplicates. These within-program matches typically reflected records in which there was missingness on a key personal identifier used to search for and assign a client key in a source data system. Following efforts to identify duplicate records, the software then deployed the linkage model to document between-program matches in a pairwise fashion. Once again, a threshold of 80 percent probability was used to classify two records from different programs as containing information about the same person. Additional methodological details and linkage information are available from the authors upon request.

**ANALYSES** After record linkages were completed, the files were stripped of personally identifiable information, and analytic files were created. Alphanumeric linkage keys generated through the linkage process allowed an examination of a client's cross-program interactions within and across years. All individuals were classified based on demographic information (sex, race/ethnicity, and age) and geography (such as county of residence or legislative district), as recorded in the administrative records for a given program. Descriptive statistics were calculated based on the full, unduplicated census of individuals eligible for services in each CHHS program. All analyses were coded in Stata, version 16.

**HUMAN SUBJECTS AND INSTITUTIONAL REVIEW BOARD APPROVALS** Data security protocols, record linkages, and analytic plans were reviewed and approved by the University of Southern California's Institutional Review Board, the CHHS Committee for the Protection of Human Subjects, and California's Vital Statistics Advisory Committee.

## Study Results

**POPULATION** On the health side of CHHS, there were 19.8 million unique individuals with certified Medi-Cal eligibility in 2015–18 and 3.7 million Californians who received reproductive health services through the Family Planning, Access, Care, and Treatment (Family PACT) program. On the human services side of CHHS, there were approximately 8.7 million people who received monthly food benefits from CalFresh; 2.7 million clients in the California Work Opportunity and Responsibility to Kids (CalWORKs) program, a welfare program that gives cash aid and services to eligible people in California; 3.9 million WIC enrollees; roughly 800,000 people who received benefits from the In-Home Supportive Services (IHSS) program; 627,500 children and parents associated with an open child welfare case; and half a million people with needs assessed for, or who were receiving, developmental services. During the study period, individuals who interacted with one or more CHHS programs were also associat-

ed with roughly 2.1 million registered birth events (either as a child born or as a parent) and approximately 300,000 deaths. The percentage of records determined to be duplicates within each program was relatively low, ranging from 0.02 percent for IHSS to 6.8 percent for Family PACT. For larger programs such as Medi-Cal, CalFresh, CalWORKs, and WIC, the share of duplicates was always less than 2 percent.

**AGGREGATED MULTIPROGRAM DATA** Examples of descriptive information that can be generated from integrated health and human services records are in the online appendix.[25] In appendix exhibit 1 we present information about the distribution of children versus adults who interacted with each CHHS program in 2017.[25] Notable variations emerged by program. In Medi-Cal, 36.0 percent of beneficiaries were younger than age eighteen. Meanwhile, the share of children among people who received CalWORKs benefits was 74.6 percent. In appendix exhibit 2 we show the numbers and percentages of children who were involved in multiple CHHS programs during the study period (2015–18).[25] We found that among the roughly 1.4 million young children enrolled in WIC in 2017, two-thirds were concurrently or sequentially enrolled in CalFresh during the study period, and 8.3 percent received developmental service supports. Appendix exhibit 3 illustrates the numbers of programs with which children interacted during the study period.[25] We found that among children with an open child welfare case in 2017, 30.8 percent interacted with five or more CHHS programs during the study period, but this was true of only 6.2 percent of children in CalWORKs. In appendix exhibit 4 we illustrate additional metrics that can be produced by presenting cross-program statistics stratified by demographic variables for a specific program (CalWORKs) in a given calendar year (2017).[25] We found that a larger percentage of white children in CalWORKs had open child welfare cases (5.6 percent), compared to black and Hispanic children (4.9 percent and 4.0 percent, respectively).

To promote transparency and encourage interest in integrated data, aggregated (deidentified) cross-program data from these linkages are available on the CHHS Open Data Portal.[26]

**CLIENT-LEVEL LINKED RECORDS** To facilitate the use of integrated data for operational and evaluation activities, individual program files were transferred back to the CHHS department with authority for the source records. Each file was returned with cross-program linkage keys and associated match probabilities. To ensure client confidentiality and careful governance during the pilot, files were returned with pairwise (program to program) encrypted linkage keys, rather than a single "master" client identification number. Additionally, returned files included linkage keys that reflected actual matches to records in other programs, as well as randomly generated linkage keys that, when used, would not return data. We adopted these approaches to ensure that a client's cross-program participation (and accompanying service information) could be determined only through a separately governed data exchange approval process within CHHS.[27]

## Discussion

Health and human services agencies are charged with delivering defined services and managing discrete programs.[4,28] Using integrated data to conceptualize client-centered, cross-program outcomes or to align programmatic activities is a secondary operational objective, at best. Similarly, developing and sustaining an infrastructure that possesses both the necessary agency authority and the resources to link records and host integrated data sets is clearly a challenge—as evidenced by the lack of government agencies that have successfully done so.[29]

Yet findings from California's record linkage efforts document several important dynamics and reinforce the value of cross-program data. First, from the perspective of minors served by CHHS, more children than not had concurrent or sequential involvement with other programs within the agency (appendix exhibits 2–4).[25] Even in the largest program, Medi-Cal, three-quarters of the children interacted with at least one additional health or human services program. Linkages underscore the opportunities to develop targeted strategies that might be delivered through more coordinated services in California, with a focus on improving outcomes, preventing adversities, and advancing equity throughout the life course.

Second, data integration efforts need not take years or cost millions of dollars. To be clear, what has been created is not a system designed to produce "real time" cross-program data. Nonetheless, CHHS now has a well-documented and routinized process for inventorying, cleansing, standardizing, and linking client-level records across its health and human services programs. The frequency with which these linkages are conducted can be modified to meet evolving operational needs. The systematic and periodic creation of cross-program linkage keys enables CHHS and its departments to avoid inefficiencies that otherwise arise from ad hoc data integration efforts specific to individual use cases. It also ensures that the same rigorous record link-

age methodologies are used across programs. While client records concerning the administration of CHHS programs continue to originate in distinct administrative data systems, CHHS now has linkage keys that can be used to connect those records while still ensuring the proper governance.

Finally, and most importantly, this data integration effort supports CHHS's efforts to achieve better outcomes for all Californians through a richer evaluation of policy options, improved stewardship of taxpayer dollars, and more coordinated design and delivery of public services. Using individual-level program linkage keys, researchers and policy makers can begin to conduct person-centered research that examines the timing, sequencing, and outcomes of service and program encounters both within and across departments. Because California's population is so diverse, and because the state has a decentralized, county-level approach to delivering services, in principle there are many opportunities to evaluate and compare the effectiveness of different programs for individuals and their families. These opportunities are only rarely exploited. Variations across demographic groups and geographic regions can be used to help reveal important questions about service access, population need, and equity.[30]

To further an agency shift toward client-centered services and cross-departmental collaboration, CHHS is working with the Children's Data Network and other partners to develop a secure, cloud-based research enclave for hosting record-level research data sets and accompanying linkage keys. Once operational, this environment will provide carefully controlled, role-based access to analysts within CHHS. 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. It is anticipated that 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, we believe that a research data hub will enhance record security and client confidentiality through data access and security protocols that can be more carefully audited.

## Conclusion

The ambitious data linkage effort undertaken by CHHS provides a remarkable new source of integrated administrative data. The resulting population-based, cross-program data can be leveraged to better characterize the public service trajectories, experiences, and outcomes of Californians over time. With exceptionally broad coverage of the population, these data provide a unique opportunity to improve coordination among programs for the people CHHS serves and to document the impact of the programs implemented. ■

The authors received grant support for this research from the Heising Simons Foundation, First 5 LA, and the Conrad N. Hilton Foundation. The authors acknowledge the departments of the California Health and Human Services Agency (CHHS) for their steadfast commitment to the use of data to improve the lives of all Californians. The leadership and vision of Linette Scott, Scott Christman, Jim Greene, Adam Dondro, Chris Krawczyk, Akhtar Khan, Jim Switzgable, David Sanabria, Alicia Sandoval, Jeannie Lin-Walsh, Mike Valle, Michelle Baass, and Marko Mijic have been crucial at all stages of data integration, from inception to implementation. The importance of the dedicated efforts of the members of the CHHS Data Subcommittee, data coordinators, and departmental staff cannot be overstated. The authors also acknowledge the incredible team of data scientists, researchers, and students at the Children's Data Network (CDN). Agency-university partnerships are too often in name and on paper only. The ongoing collaboration between CHHS and CDN is a model for what is possible. Its success can be credited to Regan Foust, Stephanie Cuccaro-Alamin, Huy Nghiem, John Prindle, Himal Suthar, Andi Lane Eastman, Siddharth Raj, Jonathan Hoonhout, and Jacquelyn McCroskey.

## NOTES

1 Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep. 2014; 129(Suppl 2):19–31.
2 Palmer RC, Ismond D, Rodriquez EJ, Kaufman JS. Social determinants of health: future directions for health disparities research. Am J Public Health. 2019;109(S1):S70–1.
3 Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. Health Aff (Millwood). 2018;37(4): 585–90.
4 Department of Health and Human Services, Office of the Chief Technology Officer. The state of data sharing at the U.S. Department of Health and Human Services [Internet]. Washington (DC): HHS; 2018 Sep [cited 2020 Feb 27]. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf
5 Culhane DP, Fantuzzo JW, Rouse HL, Tam V, Lukens J. Connecting the dots: the promise of integrated data systems for policy analysis and systems reform [Internet]. Philadelphia (PA): University of Pennsylvania, Actionable Intelligence for Social Policy; 2010 Mar 22 [cited 2020 Feb 27]. Available from: https://1ff99847xy5y3clikf2gqqb8-wpengine.netdna-ssl.com/wp-content/uploads/2019/09/Connecting-the-Dots-AISP-Version.pdf
6 Coulton CJ, Goerge R, Putnam-Hornstein E, de Haan B. Harnessing big data for social good: a grand

challenge for social work [Internet]. Cleveland (OH): American Academy of Social Work and Social Welfare; 2015 Jul [cited 2020 Feb 27]. (Grand Challenge for Social Work Initiative Working Paper No. 11). Available from: https://grandchallengesforsocialwork.org/wp-content/uploads/2015/12/WP11-with-cover.pdf
7 Baron J. The Coalition for Evidence-Based Policy will close, as an exciting new chapter begins [Internet]. Washington (DC): Coalition for Evidence-Based Policy; 2015 Apr 24 [cited 2020 Feb 27]. Available from: http://coalition4evidence.org/wp-content/uploads/2015/04/Coalition-Board-of-Advisors-Update-04-24-15.pdf
8 William T. Grant Foundation. Institutional challenge grant: overview [Internet]. New York (NY): The Foundation; [cited 2020 Feb 27]. Available from: http://wtgrantfoundation.org/grants/institutional-challenge-grant
9 Bloom-Weltman J, King K. Statewide longitudinal data systems (SLDS) survey analysis: descriptive statistics [Internet]. Washington (DC): Department of Education, National Center for Education Statistics; 2019 Oct [cited 2020 Feb 27]. Available from: https://nces.ed.gov/pubs2020/2020157.pdf
10 Kelly J. Feds put $76 million in play to support "pay-for-success" projects. Chronicle of Social Change [serial on the Internet]. 2019 Mar 5 [cited 2020 Feb 27]. Available from: https://chronicleofsocialchange.org/child-welfare-2/feds-put-millions-in-play-to-support-pay-for-success-projects/34085
11 Currie J. "Big data" versus "big brother": on the appropriate use of large-scale data collections in pediatrics. Pediatrics. 2013;131(Suppl 2):S127–32.
12 California Health and Human Services Agency. About us [Internet]. Sacramento (CA): CHHS; c 2017 [cited 2020 Feb 27]. Available from: https://www.chhs.ca.gov/home/about-us-chhs/
13 California Department of Finance. California state budget: 2019–20. Health and human services [Internet]. Sacramento (CA): The Department; [cited 2020 Feb 27]. Available from: http://www.ebudget.ca.gov/2019-20/pdf/Enacted/BudgetSummary/FullBudgetSummary.pdf
14 Card DE, Chetty R, Feldstein MS, Saez E (University of California Berkeley). Expanding access to administrative data for research in the United States [Internet]. Rochester (NY): SSRN; 2010 [cited 2020 Feb 27]. Available for download (free account required) from: http://www.ssrn.com/abstract=1888586
15 Virnig BA, McBean M. Administrative data for public health surveillance and planning. Annu Rev Public Health. 2001;22:213–30.
16 Hotz VJ, Goerge R, Balzekas J, Margolin F, editors. Administrative data for policy-relevant research: assessment of current utility and recommendations for development [Internet]. Chicago (IL): Northwestern University/University of Chicago Joint Center for Poverty Research; 1998 Jan [cited 2020 Feb 27]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.1385&rep=rep1&type=pdf
17 Gu L, Baxter R, Vickers D, Rainsford C. Record linkage: current practice and future directions [Internet]. Canberra: CSIRO Mathematical and Information Sciences; [cited 2020 Feb 27]. (CMIS Technical Report No. 03/83). Available from: http://dc-pubs.dbs.uni-leipzig.de/files/Gu2003RecordlinkageCurrentpracticeandfuturedirections.pdf
18 Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient machine learning for big data: a review. Big Data Research. 2015;2(3):87–93.
19 Winkler WE. Record linkage software and methods for merging administrative lists [Internet]. Washington (DC): Census Bureau, Statistical Research Division; 2001 Jul 23 [cited 2020 Feb 27]. (Statistical Research Report No. 2001/03). Available from: https://www.census.gov/srd/papers/pdf/rr2001-03.pdf
20 Children's Data Network. CHHS annual record reconciliation [Internet]. Los Angeles (CA): The Network; [cited 2020 Feb 27]. Available from: https://www.datanetwork.org/research/chhs-annual-record-reconciliation/
21 California Health and Human Services Agency. CHHS memorandum of understanding and intra-agency data exchange agreement [Internet]. Sacramento (CA): CHHS; 2016 May [cited 2020 Feb 27]. Available from: https://chhsdata.github.io/dataplaybook/documents/datasharing/CHHS%20Data%20Sharing%20-%20Legal%20Agreement.pdf
22 ChoiceMaker. ChoiceMaker record matching: project summary [Internet]. Princeton (NJ): ChoiceMaker; c 2013 [cited 2020 Feb 27]. Available from: https://oscmt.sourceforge.io/dev-doc/project-summary.html
23 Berger AL, Della Pietra SA, Della Pietra VJ. A maximum entropy approach to natural language processing. Comput Linguist. 1996;22(1):39–71.
24 Borthwick A. A maximum entropy approach to named entity recognition [dissertation] [Internet]. New York (NY): New York University; 1999 Sep [cited 2020 Feb 27]. Available from: https://cs.nyu.edu/media/publications/borthwick_andrew.pdf
25 To access the appendix, click on the Details tab of the article online.
26 California Health and Human Services Agency. CHHS Open Data: Health and Human Services Program Dashboard [Internet]. Sacramento (CA): CHHS; [cited 2020 Feb 27]. Available from: https://data.chhs.ca.gov/dataset/health-human-services-program-dashboard
27 California Health and Human Services Agency. CHHS data exchange agreement business use case proposal instructions [Internet]. Sacramento (CA): CHHS; [updated 2018 Nov 25; cited 2020 Feb 27]. Available from: https://chhsdata.github.io/dataplaybook/documents/datasharing/Business%20Use%20Case%20Proposal%20-%20Instructions.pdf
28 Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation. Status of state efforts to integrate health and human services systems and data: 2016 [Internet]. Washington (DC): ASPE; 2016 Dec [cited 2020 Feb 27]. Available from: https://aspe.hhs.gov/system/files/pdf/255411/StateHHSSystems.pdf
29 Doar R, Gibbs L. Unleashing the power of administrative data: a guide for federal, state, and local policymakers [Internet]. Washington (DC): Results for America; 2017 Oct [cited 2020 Feb 27]. Available from: https://results4america.org/wp-content/uploads/2017/10/Unleashing-the-Power-of-Administrative-Data.pdf
30 Lorch SA, Enlow E. The role of social determinants in explaining racial/ethnic disparities in perinatal outcomes. Pediatr Res. 2016;79(1–2):141–7.