# Preparing and Linking Administrative Data: Promising Practices and Lessons Learned from the Child Maltreatment Incidence Data Linkages Project

**Tara Strelevitz and Claire Smither Wulsin**

## Introduction

Accurate and ongoing surveillance of the incidence of child maltreatment and related risk and protective factors can help inform policy and programs, as well as shape prevention and intervention efforts. One promising approach to capturing this information is by linking local, state, or federal administrative records.

The Child Maltreatment Data Linkages (CMI Data Linkages) project identified five research groups (sites) with experience using linked administrative data to examine child maltreatment incidence and related risk and protective factors. The project supported these sites to enhance their approaches to administrative data linkage through acquiring new data sources, using new methods, or replicating existing methods. This brief highlights promising practices for preparing and linking data. We discuss lessons related to (1) processing and cleaning data, (2) completing linkages, and (3) collaborating with partners to execute linkages. Additional detail can be found in the full report, *Linking Administrative Data to Improve Understanding of Child Maltreatment Incidence and Related Risk and Protective Factors: A Feasibility Study.*

### Promising practices: preparing and linking data

- Use existing data cleaning and diagnostic protocols.
- Consult with staff who have specialized expertise on the content of administrative data sets.
- Have or develop technical familiarity with the data sources and specific data elements.
- Tailor linkage approaches to the content of the specific data source.
- Clearly communicate with the research team and third parties (if applicable) regarding the linkage approach to establish realistic expectations regarding how the linkage algorithms will operate.
- Use machine-learning techniques and tools to efficiently link larger databases.

**Table 1. CMI Data Linkages Projects**

**Replicating the Alaska Longitudinal Child Abuse and Neglect Linkage (ALCANLink) methodology**
**Alaska Department of Health and Social Services and Oregon Health Sciences University (ADHSS/OHSU)**

The ALCANLink approach used a population-based, mixed-design strategy to integrate two sets of data: (1) those births that were sampled and mothers who subsequently responded to the Pregnancy Risk Assessment Monitoring System survey and (2) child welfare and other administrative data. Alaska partnered with Oregon to replicate this methodology and to estimate and compare the cumulative incidence to first report, screen-in, substantiation, and removals by age 9.

**Methods to estimate the community incidence of child maltreatment**
**Children's Data Network and the California Child Welfare Indicators Project (CDN/CCWIP)**

This site focused on developing a methodology that used administrative data to estimate the number of children who were victims of abuse or neglect. The site produced upper and lower bounds of estimates that reflected the number of children who the child welfare system identified as victims of abuse or neglect, as well as those who were victims but not identified as such by the system. The site tested the methodology using data from California and explored the potential for using it in other states.

**Using hospital data to predict child maltreatment risk**
**Children's Data Network and Rady Children's Hospital-San Diego (CDN/Rady)**

This site tested the predictive value of integrating hospital data with vital birth records, statewide child protection records, and vital death records to identify children who might be at an elevated risk of maltreatment. The site focused on validating a statewide predictive risk model by determining the extent to which children identified to be at high risk of maltreatment are also at elevated risk of injury, poor health outcomes, and mortality in childhood. The site used machine-learning methods to train probabilistic algorithms for linking hospital-system data to other administrative data sources. These data linkages aimed to better characterize the demographics and public service trajectories of Rady Children's Hospital patients.

**Understanding the effect of the opioid epidemic on child maltreatment**
**Center for Social Sector Analytics and Technology (CSSAT)**

This site contributed to the knowledge about the opioid epidemic's potential effects on child maltreatment. Drawing from several data sources across Washington State, this project examined the associations among multiple indicators of child maltreatment, child welfare system involvement, and individual- and community-level risk factors.

**Examining child maltreatment reports using linked county-level data**
**University of Alabama School of Social Work (UA-SSW)**

This site examined how risk and protective factors relate to child maltreatment reports at the county level across the nation. The site linked county and state data from the National Child Abuse and Neglect Data System to county and state data from the U.S. Census, Bureau of Labor Statistics, Center for Disease Control and Prevention, National Center for Health Statistics, and other sources. The site aimed to explain widely varying state- and county-level maltreatment rates and to develop valid ways to use county-level child maltreatment risk.

## Processing and cleaning data

**Sites often used protocols they established in earlier projects to access, process, and clean newly received data.** They also primarily used data that had been through these procedures before. Several sites used data-cleaning and diagnostics protocols they had developed and applied before undertaking their CMI Data Linkages projects. These protocols involved standardizing some fields necessary for data linkage, such as addresses and dates of birth. They also involved checking the means and ranges of key variables to find outliers or unexpected values. One site (CDN/Rady) noted that this process helped them identify variables with values they did not understand that would require clarification. Specialized expertise with some elements of newly acquired data sets, such as diagnostic codes in hospital data, supported sites' data processing and cleaning. In at least two sites (ADHSS/OHSU and UA-SSW), the research team relied partly or fully on data sources that had undergone quality control and cleaning during data collection or preparation for public use. These sources included survey data, vital statistics data, and data from the National Child Abuse and Neglect Data System.

In at least one site (CSSAT), initial assessments of data from one provider revealed issues related to data quality. A variable related to hospitals was determined to be unusable, and the initial extract had missing and corrupted data. As a result, the research team needed to request the re-extraction and re-transfer of the files. This process took several months, resulting in delays in the project timeline. However, the site was able to use older data to begin analyses that could be refreshed once it received the corrected data.

## Completing linkages

To link individual-level records, sites used deterministic, probabilistic, and combined approaches. Sites selected linkage methods based on the type of data they used, their previous approaches to linkages, and the composition of their project teams. In the ADHSS/OHSU site, linkages involved a combination of deterministic and probabilistic methods, scoring, and manual matches. A state agency, Integrated Client Services (ICS), completed data linkages on behalf of the research team. After several rounds of matching, records were linked based on the highest scoring match. To integrate the Pregnancy Risk Assessment Monitoring System (PRAMS) and vital records data, ICS used slightly different methods for each data source. A deterministic match based on the birth certificate number was used to link PRAMS and vital records data. A probabilistic match based on names and date of birth was used to link vital records to Child Protective Services data.

One site (UA-SSW) used a direct method to link data at the county level. Data sets were merged based on a geographic identifier, the Federal Information Processing Standard code. The site matched all counties with other data sources, with the exception of about 200 that were missing data from the National Child Abuse and Neglect Data System.

**Three sites (CSSAT, CDN/CCWIP, and CDN/Rady) used machine-learning techniques to complete data linkages.** The CSSAT site relied on a cloud-based software product for data integration, known as AWS Glue. The site adopted this method after its originally planned approach (which involved deterministic and probabilistic methods) became infeasible because of an institutional review board (IRB) requirement that a third party complete the linkages. The software uses a machine-learning algorithm to identify and link records across databases. The research team was able to adjust software settings to avoid false-positive matches. The team also used the blocking statistical method to block on gender to reduce the unexplained variability from the number of record-pair comparisons.

The CDN/CCWIP and CDN/Rady sites used a custom model that they had developed for previous work. The model generated match probabilities based on similarities in linkage fields. Analysts manually reviewed uncertain matches and used the results to train the model and improve its performance as new data were integrated. In the CDN/Rady site, the newly added data source of

hospital records did not include Social Security numbers, a variable these sites typically use to link data. When the team first ran its linkages program, the match rate was much lower than expected. After consulting with data partners to understand the missing data pattern, the team was able to revise the linkage program to reflect the high level of missing of Social Security numbers in the algorithm. The match rate was higher and in the expected range after implementing the revised linkages program.

Two sites linking individual-level data (CDN/Rady and CDN/CCWIP) reported correct match rates of 85 to 92.5 percent, respectively. The research team indicated that these rates were within the expected and acceptable range for the field (Rebbe 2019).

## Collaborating with partners to execute linkages

**Two sites (CSSAT and ADHSS/OHSU) worked with outside partners to complete the data linkages separate from the research team. Research teams and their partners needed to develop technical and communicative approaches for working effectively.** In the CSSAT site, the research team's agreement with the state IRB stipulated that a named individual outside the principal investigator's organization have direct access to personally identifying information to conduct the linkages. Because of a change in personnel, this task was assigned to a staff member in a partner organization, and data linkages were not this person's primary field of expertise. To get the linkages done, the team opted to use a cloud-based software product (AWS Glue) that offered visual interfaces to control the linkage process instead of programming code. A drawback of this approach was that the linkage algorithm used in the software was not transparent to the research team, making it difficult to monitor the quality of linkages.

In the ADHSS/OHSU site, data partners required that a state agency, ICS, complete data linkages on behalf of the research team. This agency receives and links data from multiple state programs and agencies every month. Because of the partners' requirement, to ensure that the original Alaska Longitudinal Child Abuse and Neglect Linkage (ALCANLink) process could be replicated and to limit unnecessary sharing of data, the site team needed to take steps to understand the linkage process and algorithm ICS would use to link new data, such as data from the PRAMS survey. Involving a separate agency in data linkage also meant that the research team was not able to monitor the quality and completeness of linkages during that process. It was therefore important to establish a high level of confidence and trust in the linkage approach from the outset. The site team held an in-person meeting with representatives from ICS to discuss the basic approach and linkage flow for each data source. The team then documented this flow in project materials and its IRB application. Ultimately, the team determined that ICS's linkage approach was close enough to the ALCANLink method.

## Conclusion

The experiences and findings of the CMI Data Linkages sites offer important lessons about the process of preparing and completing administrative data linkages to study the incidence of child maltreatment and related risk and protective factors. The lessons underscore the potential for these approaches to inform understanding of child maltreatment.

**The sites' projects illustrate how linkages of varying levels of complexity—regarding the level of linkages and number of data sources—can yield new information for the field.** Linkages need not involve individual-level data from numerous sources to yield useful insights. For example, the CDN/CCWIP project relied on linkages of just two types of data: vital records and child welfare data. Although the cleaning, processing, and linking of these data involved complex methods, the project relied on a small number of data sources. Similarly, the UA-SSW project used relatively straightforward geographic-level linkages, rather than individual-level linkages, and publicly available data.

**The CMI Data Linkages sites implemented promising practices for preparing and linking data that enabled them to address high-priority questions about child maltreatment incidence and related risk and protective factors.** Ultimately, the information produced through these approaches might support stakeholders in estimating the extent of child maltreatment and inform efforts to prevent maltreatment through appropriately targeted supports for communities, families, and children. The promising practices highlighted in this brief represent important guidance for researchers who might be interested in replicating the approaches taken by these five sites.

## References

Rebbe, R., J.A. Mienko, E. Brown, and A. Rowhani-Rahbar. "Hospital Variation in Child Protection Reports of Substance Exposed Infants." The Journal of Pediatrics, vol. 208, 2019, pp.141–147.e2. https://doi.org/10.1016/j.jpeds.2018.12.065.

**April 2022**
**OPRE Brief:** 2022-108
**Project officers:** Jenessa Malin and Christine Fortunato
**Project Director:** Matt Stagner


## Connect with OPRE