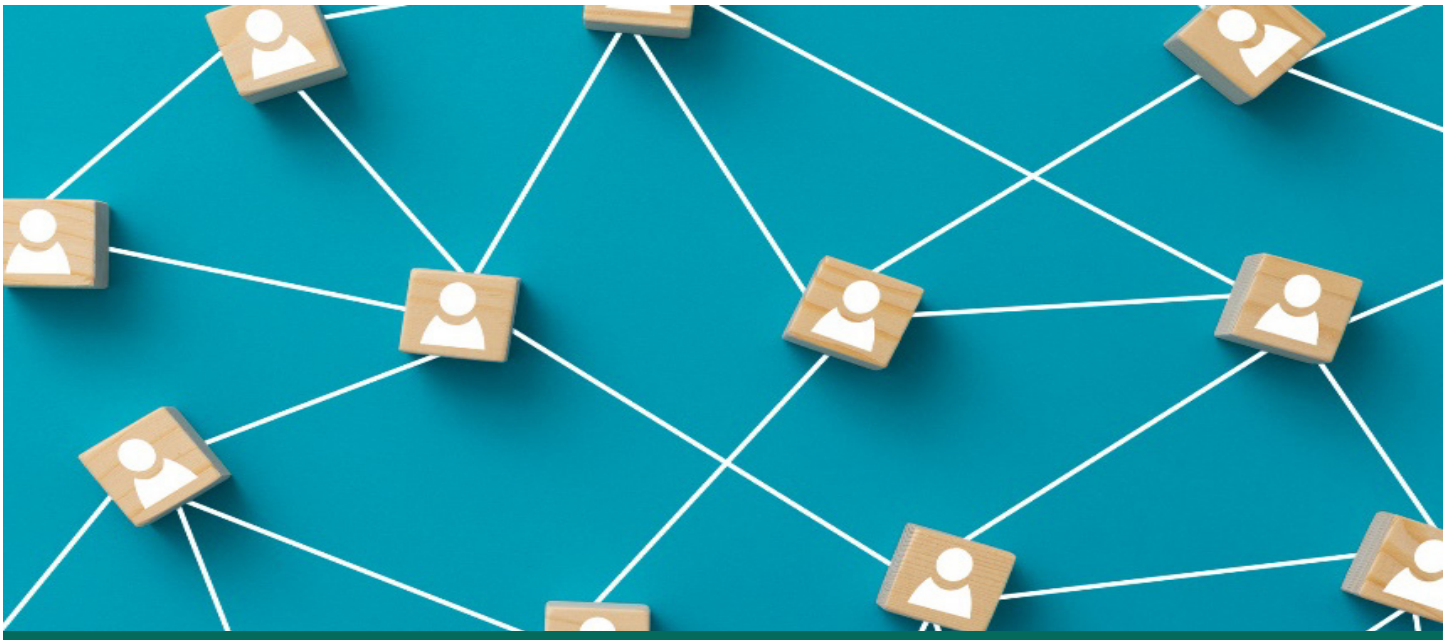
Mathematica
Progress Together



Using Hospital Data to Predict Child Maltreatment Risk

November 2021

Emily Putnam-Hornstein, John Prindle, Rebecca Rebbe, Jeannie Huang, Cynthia Kuelbs, Sarah Vega, and Regan Foust

Submitted to:

Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
200 Independence Avenue, S.W.
Washington, D.C. 20201
Attention: Jenessa Malin and Christine Fortunato
Project Officers
Contract number: HHSP233201500035I
OPRE report number: 2021-254

Submitted by:

Mathematica
111 East Wacker Drive, Suite 3000
Chicago, IL 60601-4303
Phone: (312) 994-1002
Fax: (312) 994-1003
Matt Stagner, Project Director

This page has been left blank for double-sided copying.

Introduction

This appendix is an overview of the CDN/Rady project, conducted as part of the CMI Data Linkages work. It was written by the site team, with the Mathematica team working with the site to ensure consistency in information, level of detail, and presentation across sites.

Overview

The project incorporates hospital records into existing linked data to determine whether adding hospital records increased the predictive power of an existing predictive risk model (PRM) in identifying children and youth who experienced child maltreatment. The project team linked pediatric hospital records to child protection service (CPS) records and vital birth and death records to characterize children with medical encounters for accidental and non-accidental trauma relative to children with other medical encounters. The data were used to validate statewide predictive risk models built to predict future child protection involvement and explore the use of hospital records as predictors in predictive risk models (PRMs) focused on child maltreatment.

Partnership history

The Children's Data Network (CDN) is a data and research collaborative focused on linking and analyzing administrative records. In partnership with public agencies, philanthropic funders such as First 5 LA, the Conrad N. Hilton Foundation, the Heising-Simons Foundation, affiliated researchers, and community stakeholders, CDN works to generate knowledge and advance evidence-rich policies that will improve the health, safety, and well-being of the children of California.

The CDN maintains data use agreements with numerous agencies that give us permission to link cross-sector data together and configure them longitudinally. The population-based cross-sector data can be leveraged to develop applied and actionable research, support cost-effective program evaluations, and address policy-relevant questions. Emily Putnam-Hornstein, Ph.D., who served as principal investigator, oversaw all aspects of the project and was the main point of contact with state partners.

RCHSD is a nonprofit, 551-bed pediatric-care facility dedicated to excellence in care, research, and teaching. RCHSD's mission is to restore, sustain, and enhance the health and developmental potential of children through excellence in care, education, research, and advocacy. Subaward Principal Investigator Jeannie Huang, M.D., and Co-Investigator Cynthia Kuelbs, M.D., were responsible for extracting of RCHSD data.

The Centre for Social Data Analytics is in the School of Social Sciences and Public Policy at Auckland University of Technology. The Centre applies strong data science to linked data, undertaking research that tells us new things about the human experience. These valuable insights can help us measure the impact of social interventions, identify the need for new policy, and predict the likely outcomes for individuals and groups in society. PRM Validation Subaward Principal Investigator Rhema Vaithianathan, Ph.D., and team were responsible for providing technical assistance (TA) and guidance on PRM validation and coding.

Background

First, this project will allow the CDN to use machine learning methods to train probabilistic algorithms for linking hospital-system collected health data to other administrative data sources. This will ensure a

high quality record match and lay a foundation for the ongoing linkage with subsequent years of hospital-based health data at both the local and state level.

This unprecedented analysis filled in the gaps for RCHSD patients by identifying key public service interactions before, during, and after hospital encounters. Ultimately, this analysis helped to form a more complete picture of the characteristics and public service trajectories of RCHSD patients compared to the universe of all children born in San Diego County. In addition, it enhanced our understanding of the social determinants of the physical and behavioral health of children served at RCHSD, including child maltreatment).

This project facilitated work to validate (and potentially improve) a model that predicts a child's risk of future involvement with the child protection system once a referral for child abuse or neglect has been made.[17] This preliminary work suggested a relationship between identification as "high risk" and near-fatal injury and death during childhood. Applying this model to all children born in San Diego County (Birth Model) and to all children referred for maltreatment in San Diego County (Child Protective Services [CPS] Hotline Model) helped us assess the extent to which children who are identified by this model to be "at high risk" of maltreatment also are at elevated risk of injury and mortality in childhood (that is, to validate the model). It also increased understanding of the RCHSD patient population and the pediatric population of San Diego at large. In addition, it provided a mechanism for exploring any potential value that integrating hospital records into the PRM could have on increasing the model's predictive value.

Research Questions

1. To what extent are children identified by a statewide predictive risk model (PRM) "at high risk" of maltreatment also at elevated risk of injury, poor health outcomes, and mortality in childhood? Two different models will be used to validate the PRM: (1) "birth model" predicting a child's risk of a CPS referral at birth and (2) "CPS hotline model" exploring injury and mortality risk.
2. What is the predictive value of integrating hospital data as predictors in the PRM?

Data

Sources

The data used in this work are presented in Table C.1.

Table C.1. Data sources

Data source	Description of records and sample
RCHSD patient records (RCHSD)	Records for all in- and out-patient encounters between 2010 and 2016
Vital Birth/Death records (CDPH)	Records for all births and deaths in California between 2010 and 2016
Child protection records (CDSS)	Statewide CWS/CMS records for the years 2010–2016

Source: Project documents.

[17] Please see http://www.datanetwork.org/research/assessing-childrens-risk-using-administrative-records-a-proof-of-concept-predictive-risk-modeling-prm-project for more information about the model that has been developed.

Linking process

RCHSD securely transmitted records for all clients seen between 2010 and 2016 to the CDN for linkage. The CDN linked RCHSD client records to birth, child protection, and death records using previously published machine learning methods,[18] and created a unique identifier (Table C.2). Briefly, linkages are developed using probabilistic matching methodologies that incorporated identifying information including names and dates of birth of both children and parents. The names and dates of birth of both children and parents and other unique-and non-unique identifiers were used to develop the proposed linkages. ChoiceMaker, the probabilistic record linkage software that CDN uses, is based on a machine learning technique called Maximum Entropy. ChoiceMaker is based on Clues, which are Boolean tests of similarity between fields in a record pair. Each clue is assigned a weight, which is learned through machine learning on manually reviewed training pair records. ChoiceMaker outputs a probability of a match, all pairs above a certain threshold are matches, pairs less than a certain threshold are differs and those pairs in between are held out for review. The dataset was then stripped of all direct identifiers, leaving only the encrypted unique identifier and an identifier that links back to RCHSD records and then placed on a secure server. Once analytic data were appended to the deidentified file, analysts were able to carry out the analyses. In this way, the team was able to maintain the separation between PII and analytic information (i.e., the separation principle, which is considered best practice for ensuring confidentiality and security in administrative record linkage.

Table C.2. Methods for linking data

Data source	Variables used to link data	Linkage approach
RCHSD patient records (RCHSD)	Patient First Name Patient Last Name DOB [date of birth] Address Zip Code SSN [Social Security number] Sex	Probabilistic
Vital Birth/Death records (CDPH)	Birth surname of decedent's mother Date of decedent's birth Decedent's place of birth First name of decedent Initials of decedent's spouse, if married Last name of decedent Middle name of decedent SSN of decedent Surname of decedent's father Date of birth/delivery Date of child's death Date of decedent's death Father's date of birth Father's SSN First name of child	Probabilistic

[18] (Putnam-Hornstein et al., 2020)

Data source	Variables used to link data	Linkage approach
	Last name of child	
	Last name of father	
	Mother's date of birth	
	Mother's first name	
	Mother's maiden name (birth surname)	
	Mother's SSN	
Child Protection records (CDSS)	Agency name	Probabilistic
	Badge number	
	Birth date	
	Birth place	
	Birth year – Caretaker 1	
	Birth year – Caretaker 2	
	City name	
	Common first name	
	Common last name	
	Common middle name	
	Court case number	
	Death date	
	Death place	
	Department division name	
	Driver license number	
	Driver license state code type	
	Gender code	
	Mailing city name	
	Mailing zip number	
	Name	
	Partner birth date	
	Receiver safe surrender child name	
	SSN	

Source: Project documents.

Analytic methods

To form a complete picture of the characteristics and public service trajectories of RCHSD patients, we matched administrative CPS records from San Diego County with RCHSD health encounter records for the cohort of all children born between 2010 and 2014. CPS and health care encounter data spanned the period from 2010–2016 and included information about the timing and type of referrals alleging maltreatment. After CDN linked records, staff added ICD-9 and 10 codes to document health care encounters. We used Schnitzer's (2011) condensed list of ICD codes that are suggestive or confirmatory of a medical encounter for child abuse or neglect. Using these linked records, we produced descriptive statistics about our population and examined the nature of medical encounters that led to official reports of maltreatment, identifying characteristics that affected reporting likelihood; and used medical data to validate assessed differences in risk among children reported to CPS for abuse or neglect. Four analyses were conducted:

1. **Identified children with the highest risk score and an injury encounter**. We looked at all unique children in our data, classified their risk based on the highest risk score assigned for any referral

during the analytic window, and coded all associated injury encounters, regardless of when the injury occurred relative to the child abuse and neglect referral.

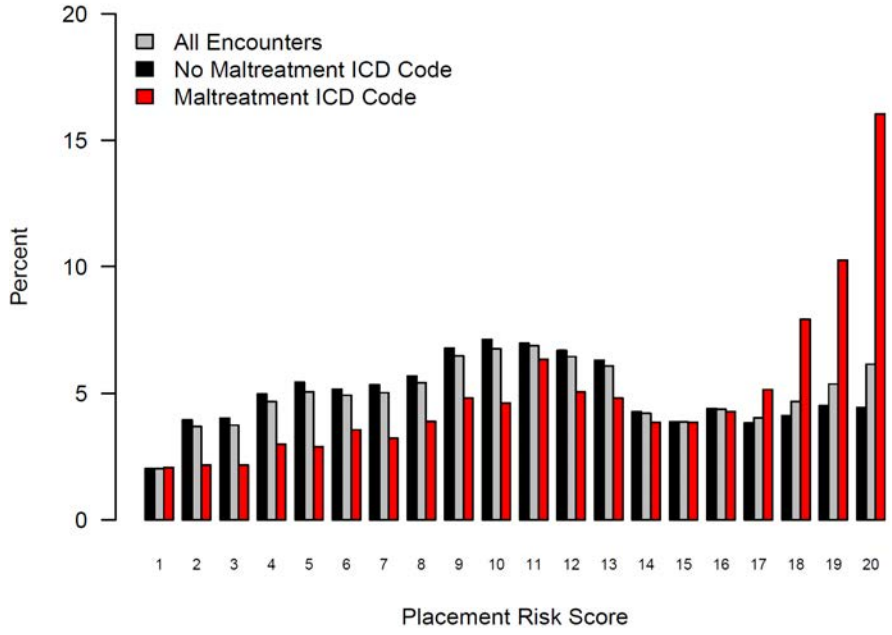
2. **Randomly selected risk score and an injury encounter.** We looked at all unique children in the data, randomly selected a CPS referral and associated risk score for each child during the analytic window, and coded their associated injury encounters, regardless of when the injury occurred relative to the selected child abuse and neglect referral.
3. **Identified the highest risk score before an injury encounter.** We looked at all unique children in the data and coded the child's risk level based on the highest risk score assigned during the analytic window, but *before* a specific injury encounter.
4. **Randomly selected risk score before an injury encounter.** We looked at all unique children in the data, randomly selected a referral and associated risk score for each child during the analytic window, and coded a medical encounter as having occurred only if the selected referral date was *before* the injury encounter.

Findings

Key findings include the following:

- Nearly one in five patients (19.7 percent) born between 2010 and 2014 who were followed by RCHSD until about age 5 had one or more report of alleged maltreatment made to CPS. This aligns with other research the team completed using statewide CPS, birth, and death records.
- Among children with a documented encounter coded as maltreatment (suggested/confirmed) in RCHSD data, most (84.9 percent) were also reported to CPS. The CPS reporting rate is lower for injuries and events with codes "suggestive," as opposed to "indicative," of maltreatment. These results reinforce the value of the list of those medical encounters for identifying the larger universe of children who may have been victims even if it was not officially documented in the medical records.
- From the perspective of the local child protection agency, 61 percent of the children who were born in San Diego between 2010 and 2014 and reported for maltreatment had a medical encounter with RCHSD.
- We found a strong association between the risk level (generated by the CPS risk model) and maltreatment-related injuries in hospital records. In addition, we found no relationship between the risk level (generated by the CPS risk model) and medical encounters that were not maltreatment-related (Figure C.1). As such, our findings not only validate a model trained to predict future CPS involvement, but also clearly indicate that maltreatment injuries are a credible indicator of child harm when viewed from the lens of the child protection system. Figure C.1. Relationship between risk for placement and hospital encounters.

Figure C.1. Relationship between highest risk score for placement and medical encounters



Next steps

We are in the process of developing multiple papers, presentations, and applications for future funding based on this work. Specifically:

1. Initial analyses investigating CPS system responses formed the basis for a COVID-focused NIH submission, "The Impact of COVID-19 on Child Maltreatment-Related Medical Encounters and System Responses Using Linked Administrative Data," submitted on October 5, 2020.
2. The team was invited to present "Use of RCHSD Healthcare Encounter Data to Validate a Child Protection Risk Model" as a poster session at the Rady Children's Hospital Interprofessional Innovations Transforming Healthcare Symposium on October 20, 2020,
3. The team has been invited to present to the Helfer Society's 2021 Pre-Conference Institute: Abusive Head Trauma & Medical Evaluation of Child Physical Abuse.
4. The team submitted an abstract, "A Longitudinal Study of Healthcare Utilization Among Infants Reported for Maltreatment", to the Pediatric Academic Societies Meeting 2020.
5. Additional papers that explore the relationship between Adverse Childhood Experiences health outcomes and the variation in identifying race/ethnicity in administrative data sources are in the works as well.

Lessons Learned About Administrative Data Linkage Practices Related to Examining the Incidence and Risk of Child Maltreatment

We learned a number of lessons based on our data linkage project—all of which would inform the considerations and recommendations we could provide to other states or localities working on data linkage initiatives related to child maltreatment incidence. Including medical records in data linkage

projects is a strategic move because it makes more funding sources and grant mechanisms available for health care projects.

SSNs are infrequently available in ER /hospital records for children. We discovered this through our linkage work with Rady Children's Hospital, but have confirmed this reality with the Office Statewide Health Planning which manages state hospitalization data. Although the inclusion of SSN numbers in medical records is higher for adults, it is clear that many pediatric medical settings do not use SSNs for medical billing. Specifically, we found that less than one third of records had a SSN. This has implications for the accuracy of the linkages that can be completed and underscores the necessity of having other non-unique personal identifiers available for linkage.

Although most (~80 percent) of maltreatment documented through official child abuse and neglect ICD codes is associated with a report to child protective services by medical reporters, the CPS reporting rate is lower for injuries and events with codes "suggestive" of maltreatment. That said, data from this study reinforce the value of the Schnitzer classification list of ICD-9/10 codes for identifying the larger universe of children who may have been victims even if it was not officially documented in the medical records. (Translations between ICD-9 and ICD-10 codes has been well documented and does not pose hurdles to coding medical records.)

It is possible to set up data sharing agreements to support the use of medical and CPS records for research, but it is a different process than simply signing an agreement with a public agency. Hospitals / ERs operate within a complex health care system. Clarity is needed about who has the authority to grant access to sensitive data and which parties are needed as signatories. These roles, responsibilities, and governance structures should be documented at the outset of a project.

Maltreatment injury base rates are, as expected, low. Thinking strategically about the numbers of years that need to be accessed and the structure of longitudinal data is critical to ensuring a sample that is of sufficient size.

Consistent with what was observed in Allegheny County, Pennsylvania, findings from San Diego, California, suggest that medical records collected in other jurisdictions are also of high enough quality to serve as useful sources for linkages and validating predictive risk models.

REFERENCES

Putnam-Hornstein, E., M. Ghaly, and M. Wilkening. (2020). Integrating Data to Advance Research, Operations, and Client-Centered Services in California. *Health Affairs*, vol. 39, 2020, pp. 655–661. Available at https://doi.org/10.1377/hlthaff.2019.01752.

Schnitzer, P.G., P.L. Slusher, R.L. Kruse, and M.M. Tarleton, Identification of ICD Codes Suggestive of Child Maltreatment. *Child Abuse & Neglect*, vol. 35(1), 2011, pp. 3–17. Available at https://doi.org/10.1016/j.chiabu.2010.06.008