USING INTEGRATED DATA TO INFORM REGIONAL POLICY AND ACTION

The Center on Urban Poverty and Community Development (the Center) was created in 1988 with founding grants from the Cleveland Foundation and the Rockefeller Foundation. Over three decades, it has become one of the nation’s preeminent university-based poverty research programs, known especially for its community partnerships and innovative applications of integrated data to inform the search for social solutions. The location of the Center within the Jack, Joseph, and Morton Mandel School of Applied Social Sciences at Case Western Reserve University provides access to tremendous resources including outstanding students and faculty, modern facilities and technology, and strong managerial and ethical oversight.

The Center’s tagline is “Research and data innovations to strengthen families and communities.” The intention is to be a source of rigorous information that shapes strategies to address the upstream causes and downstream effects of poverty and inequality. The Center’s data systems include some of the most comprehensive longitudinal and continually updated regional integrated datasets in the country. The Center places primary value on using these assets to inform civic priorities with cogent data that can be the key to successful program and policy design.

The Center carries out a program of applied research, evaluation, policy analysis, and information dissemination. It follows a model of working closely with stakeholders to design questions and formulate analyses that will contribute to the knowledge base and action in various program or policy arenas. It also works to communicate the data and findings from these studies to inform policy conversations locally, regionally, and nationally. Center faculty and staff are frequent presenters in local and national forums, publish in scientific journals, and serve as trusted sources for policymakers and the media.

The Center’s expertise and research experience is relatively broad due to the systemic nature of poverty and inequality. Its staff includes social and data scientists, programmers, evaluation specialists and community engagement professionals. They have experience in conducting research and evaluation on a range of topics such as maternal and child health, early childhood programs, public assistance policy, neighborhood decline and revitalization, child welfare, youth employment, housing quality and affordability, homelessness, evictions, community services and assets, social determinants of health, and more. The structure of the Center is flexible to facilitate teamwork and the deployment of staff expertise across projects in an efficient manner.

The projects undertaken draw upon several core data capacities maintained by the Center. This means that most projects are not starting from scratch, but are benefiting from data and tools built up over previous projects. The Center’s notable core capacities are summarized below:

- **ChildHood Integrated Longitudinal Data (CHILD) System** is composed of linked administrative records of individuals in Cuyahoga County, Ohio, born since 1989 and continuing to the present. Based on data use agreements with 37 health, educational, and human services agencies, CHILD currently holds nearly 200 million records in a highly secure research environment. This represents an unduplicated count of over 700,000 individuals. The records cover health, education, human services, justice system, housing services, employment services, and residential locations. CHILD data have been used for numerous studies and have been drawn upon to inform a number of strategic initiatives in the region. (For details on the CHILD system, see the attached documentation beginning on page 4.)

- **Neighborhood Strategy Technology Web Application (NST)** enables a diverse group of government and community agencies to collaborate on improving the communities of Cleveland and Northeast Ohio. They can quickly access information on property ownership, occupancy and vacancy, housing conditions, foreclosure, sheriff’s sale, REO history, detailed tax information, code enforcement actions, demolition information, property surveys conducted by area agencies, and more. In addition to linking and making data available, NST also serves as a communication and analytic platform for agencies to
house their own data. Mobile Apps can access NST, so that residents can easily search out information about properties they are considering renting or are concerned about in their neighborhoods. These detailed property-level data can also be linked to residential addresses to evaluate the influence of housing conditions on individual outcomes.

- **NEOCANDO Neighborhood Data Warehouse** provides statistical indicators covering a wide array of topic areas such as demographic and socioeconomic information, crime, safety and health conditions, public service use, educational status and performance, built environment, and housing market trends. The warehouse delivers information for various external applications such as Cleveland Progress Index and Health Data Matters. NEOCANDO data are often linked to geocoded records from the CHILD system for studies of how neighborhood conditions affect outcomes for individuals.

**CONSTITUENCIES SERVED**

The Center regards its primary constituents as public officials and other policymakers, philanthropic and nonprofit leaders, human service agencies and their administrators, community organizations, and the public in Greater Cleveland. Drawing on its data resources, substantive expertise, analytic tools, and collaborative relationships, the Center has had the privilege of being directly involved in or a key partner and resource for a number of important civic priorities. Recent examples include:

- **Data for a Lead Safe Cleveland**: The Center is serving as a data partner that will inform agenda setting and action driven by an emerging coalition of philanthropic and government leaders. The goal is to address lead poisoning at its source—that is, housing that exposes children to lead dust and paint. The research products include a profile of Cleveland’s affordable rental housing stock and landlords, an assessment of strategies for lead hazard control, a study of the downstream costs of lead exposure, and the development of metrics to monitor progress and outcomes.

- **Eviction—Circumstances and Consequences for families**: Partnering with Cleveland Legal Aid and housing court, the Center is conducting a study of evictions. The study includes interviews with families facing eviction to evaluate their needs and circumstances and record linkage of evictions with other social service records to highlight some of the costs of eviction to society.

- **Addressing costs related to vacant housing**: The Center has been involved in a series of analyses pinpointing the spillover effects of vacant housing on property values and neighborhood problems such as crime. Thriving Communities, Cleveland Neighborhood Progress, Dynamo Metrics, and the National Institute of Justice are among the funding and/or policy partners. The studies have documented increases in crime near vacant housing, and have estimated counterfactual models showing that demolition or rehabilitation of vacant housing can halt the adverse effects on property values.

- **Child Welfare—Juvenile Justice crossover youth**: There is both local and national concern about these high-risk populations, but little is known about how the pathways through the child welfare and juvenile justice systems intersect. The Center has received funding from national foundations and the federal government to examine these pathways in Cuyahoga County, make comparisons with other counties, and recommend prevention strategies.

- **BUILD health housing**: Environmental Health Watch is the lead agency for this partnership to improve housing health and advocate for vulnerable families. A key component is to use data to identify at-risk properties, and to monitor code enforcement and other actions to improve these properties. The Center is working with all partners to link their relevant data and build a phone app that will allow families to get up-to-date information on the conditions of units that they are considering renting.

- **Pay for Success**: The Center has been engaged in two social impact financing projects where IDS was essential to the planning and evaluation. The Center conducted the analyses to inform the development of the Partnering for Family Strength project, the first county-sponsored PFS project in the U.S., and now serves as the external evaluator. More recently, Invest in Children received a grant from the Department of Education to examine the feasibility of using a Pay For Success social finance model to expand access to high quality preschool to all 3- and 4-year olds in Cuyahoga County. The Center used data from the CHILD system to estimate the potential impact and cost savings necessary for determining the feasibility of the social impact financing mechanism.

- **Early Childhood Education**: In collaboration with Cuyahoga County’s Office of Early Childhood/Invest in Children, PRE4CLE, and Starting Point, the Center has been involved in a number of needs assessment and monitoring efforts to understand the preschool landscape
in Cuyahoga County and the City of Cleveland. This work includes comparisons of underlying demand, capacity and enrollment to identify areas in need of high-quality preschool as well as evaluation studies exploring the relationships between high-quality preschool enrollment, attendance, and kindergarten readiness.

- **Early Childhood Services:** For nearly 20 years, the Center has worked in partnership with Invest in Children and nonprofit service providers to evaluate county-funded services available to young children and their families living in Cuyahoga County. Recent evaluations include assessments of home visiting programs like SPARK and early childhood mental health, special needs child care for early learning sites serving children with medical, social, and environmental needs, and county-wide monitoring of lead testing and exposure rates.

- **Infant mortality:** The City of Cleveland is home to one of the highest infant mortality rates in the county. More troubling is the significant and persistent racial disparity in birth outcomes, with African American babies dying at approximately 2.5 times the rate of white babies. The Center serves as the research and evaluation partner for the City of Cleveland’s MomsFirst home visiting program that works with pregnant women who are at higher risk, due to biological, social, and environmental determinants of health, of experiencing a poor birth outcome.

- **Promise Neighborhoods:** Residents of Cleveland’s Central neighborhood, where 82% of children live below the federal poverty line, struggle with issues associated with multigenerational poverty. The Sisters of Charity Foundation of Cleveland has sought to make a population-level impact on the health and well-being of children, youth, and families living in one of the most economically distressed neighborhoods in the county. The Center has used its linked records on children and housing to inform the planning and implementation of this program and to track benchmarks for success.

Beyond its regional role, the Center also reaches a national audience and has an impact on discussions at the federal level. It accomplishes this through its active participation in important networks such as the Urban Institute’s National Neighborhood Indicators Partnership, the Ohio Education Research Center, and the Actionable Intelligence for Social Policy network based at the University of Pennsylvania. For example, the Center’s work on lead poisoning was recently nationally recognized as an outstanding model for how local data intermediaries can contribute to informed action to address community challenges. The Center’s housing research has also been featured in recently policy forums and congressional testimony. The Center’s NST platform was featured in a profile of proven practices for addressing the foreclosure crisis by the Federal Reserve.
1. INTRODUCTION

Governmental and nonprofit agencies and institutions routinely generate electronic administrative records related to the population they serve. These records capture information about individual and family characteristics, eligibility, risk assessment, services received, and information about outcomes or results. But individuals who use programs and services often participate in other programs at the same time, traversing multiple systems as they move along in their development. The ongoing integration of administrative records across agencies and time has the potential to provide new types of information that can be used to evaluate outcomes, drive decision making, target resources, and gain an understanding of how the collective work of agencies and systems are addressing the needs and concerns in their communities.

The ChildHood Integrated Longitudinal Data (CHILD) System is composed of linked administrative records of individuals in Cuyahoga County, Ohio, born since 1989 and continuing up to the present. The linkage of records across time and systems is performed via a combined approach involving probabilistic matching and machine learning. The records contain geographic information that enables aggregation to the neighborhood, city, county, or other jurisdiction level and also allows linking with other data systems at various levels of geographies such as parcel, address, or census tracts. The end result is a longitudinal data system in which individuals are observed if and when they are served by one of the numerous agencies and systems that contribute electronic records to CHILD.

Figure 1: Overview of the types of records contained in the CHILD System

CHILD integrates data from 35 different administrative systems represented in nine broad categories above.
The CHILD System began in the late ’90s as part of the Cuyahoga County Invest in Children (IIC) initiative. \(^1\) IIC is a community-wide, public-private partnership of government leaders and agencies, nonprofit organizations, and local foundations. The purpose of IIC is to assure that all the county’s young children and their families receive the supports they need so that they and their families, communities, and schools are ready for them to enter kindergarten and succeed. Since no one agency or system within the initiative can achieve this outcome its own, the CHILD System was required to support joint planning and the evaluation of collective impact. Since its inception, the CHILD System has gone from comprising information across seven to 35 administrative systems and has been extended to cover the population as it transitions to adulthood (see Figure 1 on page 4).

From the 37 administrative data systems, CHILD holds nearly 200 million records (see Table 1). The CHILD System currently contains information related to the following observable events: birth, death, prenatal and ongoing home visiting, early childhood mental health services, subsidized child care, public assistance benefits (e.g., Medicaid, Cash Assistance, SNAP), housing subsidies and public housing, lead test results, special needs child care, public preschool and UPK attendance, kindergarten readiness assessments, public school attendance and test scores, child welfare involvement, juvenile court

---


### Table 1: CHILD Data Sources, Years, Number of Records

<table>
<thead>
<tr>
<th>DATA SOURCE</th>
<th>DATA YEARS</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VITAL RECORDS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ohio Births</td>
<td>1989-present</td>
<td>Vital Statistics</td>
</tr>
<tr>
<td>Ohio Deaths</td>
<td>1989-present</td>
<td>Vital Statistics</td>
</tr>
<tr>
<td><strong>HUMAN SERVICES</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCFS Child Abuse and Neglect Reports</td>
<td>1989-2017</td>
<td>Demographics, abuse/neglect, placements, service use</td>
</tr>
<tr>
<td>Public Assistance</td>
<td>1992-present</td>
<td>Demographics, service use (Medicaid, SNAP, TANF)</td>
</tr>
<tr>
<td>MomsFirst Prenatal Home Visiting</td>
<td>2007-present</td>
<td>Demographics, service use, assessments</td>
</tr>
<tr>
<td>Help Me Grow Home Visiting</td>
<td>1999-2014</td>
<td>Demographics, service use, assessments</td>
</tr>
<tr>
<td>Early Childhood Mental Health</td>
<td>2008-2012</td>
<td>Demographics, service use</td>
</tr>
<tr>
<td>County Newborn Home Visiting</td>
<td>2010-present</td>
<td>Demographics, service use, assessments</td>
</tr>
<tr>
<td>Homeless Services</td>
<td>2009-present</td>
<td>Homeless clients and entry/exit information</td>
</tr>
<tr>
<td>Cuyahoga Metropolitan Housing Authority</td>
<td>1989-present</td>
<td>Public Housing and Housing Choice Voucher (Section 8) application data and dates of use</td>
</tr>
<tr>
<td><strong>EDUCATION/CHILD CARE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal Pre-Kindergarten (UPK)</td>
<td>2007-present</td>
<td>Demographics, attendance, assessments, providers</td>
</tr>
<tr>
<td>Special Needs Child Care</td>
<td>2001-present</td>
<td>Demographics, assessments, service use, providers</td>
</tr>
<tr>
<td>Cleveland Metropolitan School District</td>
<td>2005-present</td>
<td>Demographics, enrollment, attendance, and assessment data</td>
</tr>
<tr>
<td>Suburban School Districts</td>
<td>2005-present</td>
<td>Demographics, enrollment, attendance, and assessment data for 17 inner ring suburbs</td>
</tr>
<tr>
<td>Charter Schools (Intergenerational)</td>
<td>2005-present</td>
<td>Demographics, enrollment, attendance, and assessment data</td>
</tr>
<tr>
<td>Subsidized Child Care</td>
<td>1997-present</td>
<td>Demographics, service use, providers</td>
</tr>
</tbody>
</table>
involvement, homelessness services, and county jail spells. In addition, the following data sources are pending: Ohio Means Jobs, unemployment insurance wage records, child support enforcement, and Breakthrough charter schools. There is also a pilot underway to link CHILD records with electronic health records of children covered by one of the area’s major Medicaid accountable care organizations.

2. STRUCTURE, MATCHING AND UPDATING PROCEDURES

On a regular schedule, new data become available from the various administrative entities that provide data to the CHILD System. SAS programs are used to process and standardize the data to prepare for matching [linking] the new records over time and across systems. Variables that are used for matching vary from one dataset to another, but can include names, birth dates, social security numbers, home address, race, gender, and relationships. We geocode all incoming addresses using MapMarker, MapInfo, and ArcGIS. Once the data are prepared for matching, we transform the data into SQL tables that provide the input to the linking software, ChoiceMaker.

ChoiceMaker is an open source product that is unique among data linking software in that it is based on machine learning. The matching methods are completely transparent to the user and can deliver extremely high accuracy. ChoiceMaker incorporates many techniques including Soundex phoneticization, Edit-Distance, and Jaro-Winkler. Each algorithm compensates for a different type of error, including typographical errors, transcription errors, or misspellings. Record matching is a two-stage process. Stage one focuses on finding possibly matching records (blocking); this stage is extremely fast with ChoiceMaker. In the second stage, ChoiceMaker employs an Artificial Intelligence approach, called maximum entropy modeling, to build models trained on the specific data of the user’s database. The user develops clues (complex, multi-field matching criteria) that are tailored to the idiosyncrasies of each users’ data. The machine learning technology then learns the user’s decision-making process during the model building phase, by having the user mark a set of training data as match or differ. Once the model has been developed, production runs create an output file of possible matched pairs which are each assigned a probability. The user can assign the probability cut points to determine which pairs are rejected (differ), which pairs match, and which pairs require human review, depending on the desired level of accuracy (see Figure 2).

ChoiceMaker has a component called Analyzer that permits us to review the system’s accuracy (including false positives and false negatives) based on a set of data that we have reviewed manually. Another advantage of using ChoiceMaker is that there is a growing national and international community of ChoiceMaker users. Through this community, it is possible to share information about models and clues, to save time in developing a new system.

---

Once the data have been linked with ChoiceMaker, we save the linked data in a series of SAS datasets, including analysis files that can be de-identified before sharing with researchers. To date, the CHILD System has grown through the careful matching and updating of Event Tables out of which the analysis files have been created and analyses have been performed. As the demands for data from the CHILD System grow, it will be necessary to consider the creation of a Master Event Table including direct or derived variables from all systems at once. This master table would not include any Personally Identifiable Information (PII) so that researchers could access this table to create their own analysis files. The master table would be linked to a metadata repository that would facilitate the process of planning new studies, selecting study populations, and conducting analyses.

3. POPULATION REPRESENTED IN THE CHILD SYSTEM

Representation in the CHILD System depends on the variables of interest. Any cross section of the data includes all children born in 1989 and later who experienced the event of interest in the specific time period queried. The administrative data can track onetime events such as births, or recurrent events such as home visits; however, the majority of events represented in the CHILD System occur more than once with specific dates attached. As the schedule of appearance in administrative records is highly variable across children, researchers routinely create count variables of service receipt over a specified period of time or identify spells of service engagement and other metrics.

For some children, birth certificate data will be the only time they appear in the CHILD System, because they have not experienced any of the events captured in the administrative records. It is currently estimated that approximately 90% of Cleveland children and 60% of children residing in the county appear in multiple records in CHILD. This proportion will continue to increase as additional suburban school districts and charter schools enter into data sharing agreements with CHILD.

Recently, we have begun to link children to their parents or guardians and are following them into adulthood as well. Birth certificate, public assistance, child welfare, and homelessness services records are currently being used to identify the parent and other household members related to children born since 1989. Continuing this process will allow us to not only begin to identify key events that are associated with generational poverty but, more importantly, opportunities for interrupting this cycle over the life course. Two-generation data could also be used to better understand the ripple effect of a major life event throughout a family system. We are also beginning to integrate workforce data into our system allowing further tracking of individuals into adulthood.

4. LINKING THE CHILD SYSTEM TO NEIGHBORHOOD OR ADDRESS LEVEL DATA

The residential addresses and geocodes in the CHILD System provide a means to link data from CHILD with data from other systems that contain information at the address or neighborhood level (using geocodes such as block group, census tract, or other geographical unit). For example, the Center on Urban Poverty and Community Development (Poverty Center) at Case Western Reserve University has a free and publicly accessible longitudinal database of social, economic, and property data called NEO CANDO. Neighborhood level variables can be extracted from the NEO CANDO system and appended to individual records based on geocodes. Examples include rates of poverty, unemployment, foreclosures, violent and property crimes, and vacant housing. Individual addresses can also be linked to property records compiled in another Poverty Center database, the Neighborhood Stabilization Technology integrated parcel information system (NST). Historical data from NST include information at the address level on housing type, conditions and values, land use, mortgage originations, deed transfers, foreclosure filings and completions, vacancies, code violations, demolitions, tax delinquencies, building permits, and community development investments.

5. CONFIDENTIALITY AND SHARING IDENTIFIABLE INFORMATION

Creating an integrated data system (IDS) remains a difficult endeavor despite their growing presence in the field of human services. In particular, the process of accessing data from

---

1A small number of birth records are missing each year due to Ohio residents giving birth outside the state or their birth certificate information being suppressed due to adoption.

4NEO CANDO system, Center on Urban Poverty and Community Development, MSASS, Case Western Reserve University [http://neocando.case.edu](http://neocando.case.edu)

agencies is governed by federal regulations such as the Privacy Act, HIPAA, and FERPA, as the release of PII is a risk to personal privacy and confidentiality. Federal, state, and local regulations dictate strict limitations as to what identifiable information can be released. Due to university policies, university legal staff typically work with lawyers from data contributing agencies to enter into a process of data sharing bound by a legal document, often called a data use agreement (DUA) or memorandum of understanding (MOU), governing acceptable uses of the data. Most IDS, including the CHILD System, also submit protocols for specific projects and receive approval from an Institutional Review Board (IRB) charged with the protection of human subjects.

The CHILD system is protected following a strict set of procedures. Only a small number of staff who are certified in human subjects protection and have signed oaths of confidentiality work with identifiable records. All work with these records is done on a highly secure server and personal identifiers are stored separately and linked through a random ID. Researchers and analysts, who are also certified by the IRB, work with de-identified data sets to produce summary statistics.

6. GOVERNANCE

The CHILD System has a multidimensional governance model designed to ensure excellent data stewardship. Strict DUAs are executed between the Poverty Center and all data providers that explicitly state the terms of use and expectations of confidentiality and security. Through these agreements, the data providers govern the use of their data. The CHILD System is also governed by the IRB at Case Western Reserve University. The IRB assures that all research conducted using data from the CHILD System are in compliance with all federal protections of human subjects, including privacy and risk.

In 2016, the CHILD Advisory Group was first assembled to provide guidance for the CHILD System. The purpose of the Advisory Group is to facilitate and maximize the application of CHILD in community planning, policy, and evaluation, as well as to provide advice on governance and sustainability initiatives. The Advisory Group can be particularly helpful in developing a framework for the effective use and development of the CHILD system as a community and governmental resource.

This advisory body comprises representatives from several key constituencies including:

- Data providers: public agencies and nonprofit organizations that supply data to the Poverty Center for inclusion in CHILD.
- System partners: organizations that serve many of the same children and families, but do not currently provide data to CHILD.
- Funders: governmental, federated, and philanthropic funding entities that invest resources in serving the populations represented in CHILD.

The Poverty Center also routinely collaborates with other advisory groups for specific projects or grant applications to increase capacity and generate discussion among the agencies that supply data, the foundations that fund grants, and the officials who develop policies at local, state, and federal levels.

7. LIMITATIONS

As with all integrated administrative data systems, the richness and quality of the data are dependent upon access to and quality of administrative records, incidence of service use, and mobility patterns of the population. The following limitations of CHILD should be noted. First, all data in the CHILD System were originally collected for a wide variety of program purposes rather than rigorous research analyses. As a result, the accuracy and reliability are not as high as would be expected for data collected in controlled settings. Second, the CHILD System does not currently include data from all providers of a particular service, such as nonprofits not under government contract. Third, we cannot be sure whether the lack of an event means that it did not happen or that it was not recorded. For example, not all child maltreatment situations are reported to the authorities and some individuals experiencing homelessness may not seek help from agencies. Fourth, given the lag time associated with data extract receipt, data in the CHILD System cannot be used for real-time decision making. Rather, analyses highlight trends and patterns in service receipt and outcomes over time for planning and decision making going forward.

---

8. SUSTAINABILITY

Since its founding, the Poverty Center has been committed to finding better solutions to problems associated with urban poverty by using data and research to inform decision making. In the last 20 years, the Poverty Center has invested over $2 million in its IDS efforts, with the majority of funds committed on a project-by-project basis to support targeted research to inform vital policy or programmatic concerns. This piecemeal approach has garnered impressive results, though there is growing recognition among university and community leaders that a structured and strategic investment in its current IDS infrastructure is needed to enable further expansion of its innovative efforts to inform effective and efficient decision making.