AN IDS RESOURCE FOR CUYAHOGA COUNTY:

CHILDHOOD INTEGRATED LONGITUDINAL DATA SYSTEM

Claudia Coulton

Robert Fischer

Nina Lalich

Elizabeth Anthony

Francisca Richter

Meghan Salas Atwell

Center on Urban Poverty and Community Development

Case Western Reserve University

December 2017

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 on-going 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.

The CHILD System began in the late 90's 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, non-profit 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 on their 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 7 to 35 administrative systems and has been extended to cover the population as it transitions to adulthood (see Figure 1).

---

[1] Fischer, R. L., Lalich, N., & Coulton, C. (2008). Taking it to scale: Evaluating the scope and reach of a community-wide initiative on early childhood. Evaluation and Program Planning, 31, 199-208.

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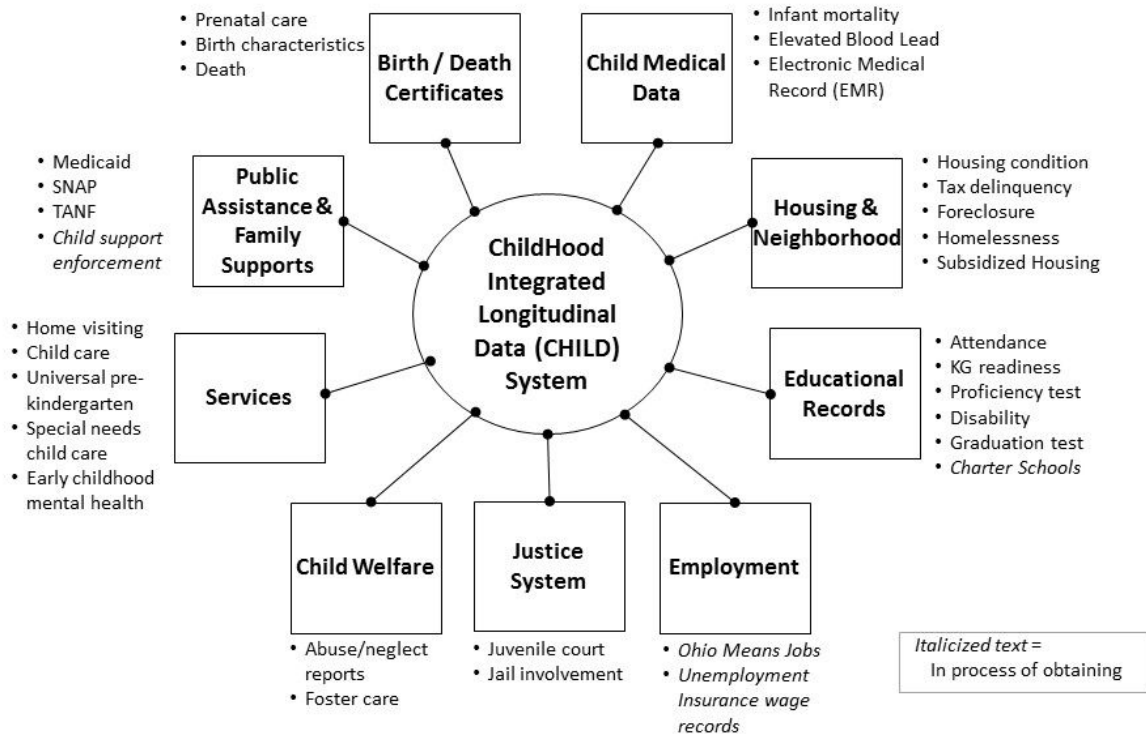

Figure 1. CHILD integrates data from 35 different administrative systems represented in nine broad categories above.

From the 35 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 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.

Table 1. CHILD Data Sources, Years, Number of Records

| Data Source	Data Years	Number of Records	Notes
Vital Records			
Ohio Births	1992-2017	3,790,000	Vital Statistics
Ohio Deaths	1992-2017	2,600,000	Vital Statistics
Human Services			
DCFS Child Abuse and Neglect Reports	1989-2017	29,770,000	Demographics, abuse/neglect, placements, service use
Public Assistance	1992-2017	125,050,000	Demographics, service use (Medicaid, SNAP, TANF)
MomsFirst Prenatal Home Visiting	2007-2017	3,790,000	Demographics, service use, assessments
Help Me Grow Home Visiting	1999-2014	1,660,000	Demographics, service use, assessments
Early Childhood Mental Health	2008-2012	2,000	Demographics, service use
County Newborn Home Visiting	2010-2017	160,000	Demographics, service use, assessments
Homeless Services	2009-2017	12,780,000	Homeless clients and entry/exit information
Cuyahoga Metropolitan Housing Authority	1989-2016	130,000	Public Housing application data and dates of use
Education/Child Care			
Universal Pre-Kindergarten (UPK)	2007-2017	240,000	Demographics, attendance, assessments, providers
Special Needs Child Care	2001-2017	160,000	Demographics, assessments, service use, providers
Cleveland Metropolitan School District	2005-2017	2,450,000	Demographics, enrollment, attendance and assessment data
Suburban School Districts	2005-2017	5,500,000	Demographics, enrollment, attendance and assessment data for 10 suburban school districts
Subsidized Child Care	1997-2017	7,850,000	Demographics, service use, providers
Starting Point Child Care Provider Data	2002-2017	740,000	Provider type and address, child care slots, enrollment, quality ratings
Corrections			
Juvenile Court Data	2000-2016	1,990,000	Demographics, cases and charges, and hearing and disposition information
Cuyahoga County Jail	2002-2016	400,000	Demographics, entry/exit dates
Health			
Lead	2000-2015	410,000	Demographics, blood lead test results
Data Sources Pending			
Breakthrough Charter Schools			
Child Support Enforcement			
Ohio Means Jobs			
Unemployment Insurance Wage Records			
Electronic Medical Records			

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.[2]

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).

---

[2] Goldberg Arthur, Borthwick Andrew, ChoiceMaker Technologies, Inc. (2004). The ChoiceMaker 2 Record Matching System. Choicemaker Technologies, Inc.

Figure 2. ChoiceMaker matching overview
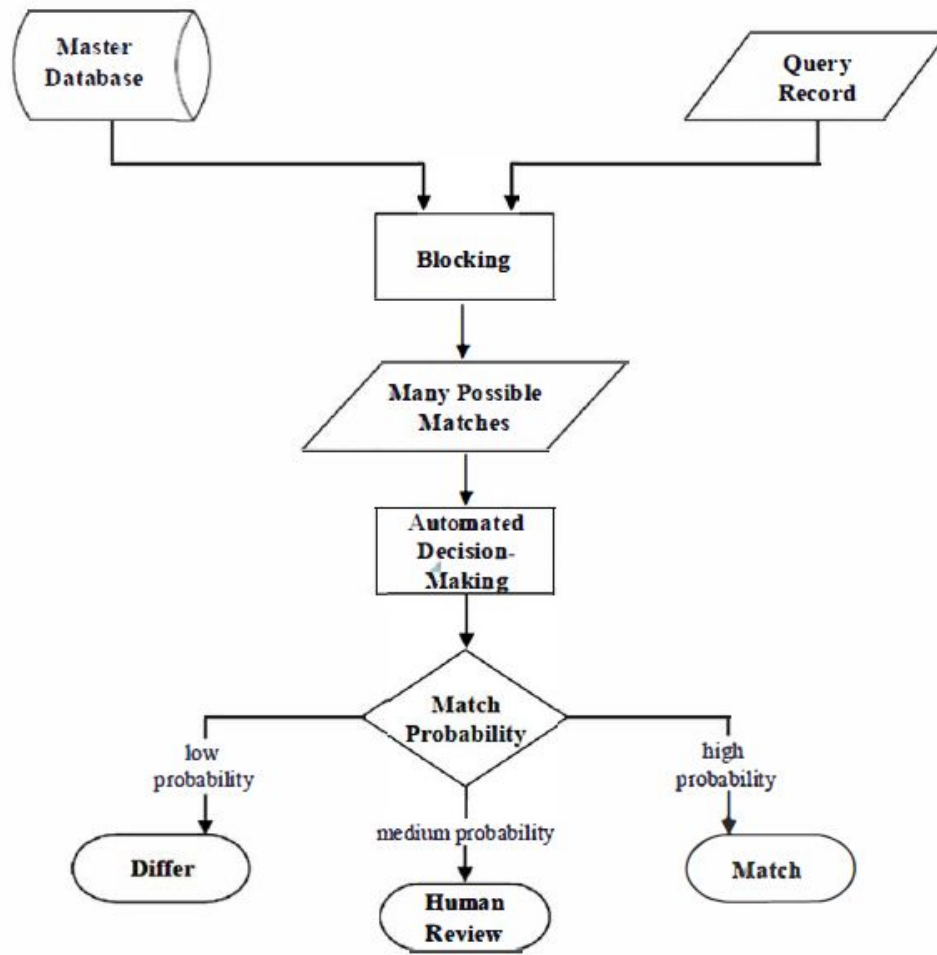
ChoiceMaker has a component called Analyzer that permits us to review the system's accuracy (including false positives and false negatives) based on a set of data that we have reviewed manually. Another advantage of using ChoiceMaker is that there is a growing national and international community of 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 one-time 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 they have not experienced any of the events captured in the administrative records. [3] 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've begun to link children to their parents or guardians and are following them into adulthood as well as including adult public assistance users born prior to 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)

---

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

at Case Western Reserve University has a free and publicly accessible longitudinal database of social, economic, and property data called NEO CANDO.[4] 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, or 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).[5] 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 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.[6] 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

---

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

[5] Hirsh, A., Schramm, M., & Coulton, C. (August 2012). Neighborhood Stabilization Team Web Application. Briefly Stated No. 12-04. Center on Urban Poverty and Community Development. Retrieved from: http://blog.case.edu/msass/2012/09/13/Briefly_Stated_No_12-04_NST_Web_App.pdf

[6] Petrila, John (2011). Legal Issues in the Use of Electronic Data Systems for Social Science Research. Actionable Intelligence for Social Policy (AISP), University of Pennsylvania. Retrieved from: http://www.sp2.upenn.edu/aisp_test/wp-content/uploads/2012/12/0033_12_SP2_Legal_Issues_Data_Systems_000.pdf

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 multi-dimensional 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 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 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 non-profits 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. Examples of the CHILD System in use

The CHILD System can be used to help Cuyahoga County monitor and comprehensively explore strategic outcomes. The following selected examples illustrate how the CHILD System can be used for program planning, improvement, evaluation and policy innovation.

Program Planning

Magnitude of Need: In 2011, the Sisters of Charity Foundation of Cleveland engaged the Poverty Center in a planning process to create a Promise Neighborhood in Central, one of Cleveland's most economically distressed neighborhoods. Data from the CHILD System were used to describe the magnitude of need in the community. This descriptive information informed the identification of partners and service providers who could participate in revitalizing Central. With a collaborative team in place, data from the CHILD System have been used to monitor and describe demographic, educational, and social service involvement changes in the neighborhood over time.

Cross Service Use: Recent analyses using data from CHILD have examined the cross-service use of adult-headed SNAP assistance groups as well as the use of public assistance by families experiencing homelessness. These descriptive analyses represented an initial step toward identifying the extent to which adult-headed SNAP assistance groups experience homelessness, jail stays, and reliance upon public housing and other forms of public assistance. In addition, our analysis explored the extent to which families experiencing homelessness relied upon public assistance in the years leading up to and after shelter entry. Future analyses will be more predictive in nature, identifying risk and protective factors that are associated with adverse events such as homelessness, child maltreatment, joblessness, and criminal justice involvement across households and families to aid in the design of effective interventions. An October 2017 report prepared by the Seattle Jobs Initiative for Cuyahoga County recognized the unique

contribution that the CHILD System could make to these efforts by recommending that the community, "explore the option of using CHILD and other data systems to better understand those who are 'at-risk' of becoming homeless and their experiences as well as understanding how the data may help predict system utilization."

Program Improvement and Evaluation

Medical Home Study: Ongoing monitoring of health insurance coverage showed nearly universal coverage of the child population under age 6 in Cuyahoga County. But, using CHILD to review data on receipt of well-child visits showed that only 50% of infants on Medicaid were receiving the recommended number of well-child visits in the first year of life. Thus, access to medical care did not translate to utilization of medical care. As a result, the County developed a medical home model in which patient advocates in health clinic settings conducted outreach to pregnant women. The patient advocates also served as liaisons between healthcare providers and families. Initially, patient advocates were embedded in two health care clinics via a pilot project. CHILD was used to evaluate the effectiveness of the medical home pilot. The completion of scheduled well-child visits in the pilot sites was about double the rate found in similar populations in Cuyahoga County (86% v. 40%).

Policy Innovation

Partnering for Family Success: The CHILD System was used to inform the County's decision to explore performance-based contracting (Pay for Success or PFS) for financing human services. Through a partnership between government, nonprofit, philanthropic organizations, and private investors, PFS provides immediate capital to implement innovative social programs targeting entrenched social issues. If successful, governments see cashable savings and a return to investors. Using the CHILD System, the Poverty Center identified several high risk groups known to engender high costs for the County. Ultimately, after thorough exploration of service receipt patterns, the County launched Partnering for Family Success to more quickly reunite parents experiencing homelessness with their children who were placed in out-of-home foster care.

9. Examples of Innovation from other IDS

Below are a few select examples highlighting the unique contributions of other IDSs around the country. These examples are offered to further demonstrate what is possible here in Cuyahoga County using CHILD. The use of predictive modeling and in-depth exploration into complex family structures and intergenerational poverty could be replicated or expanded here.

Allegheny County - Predictive Risk Modeling: The Allegheny County Department of Human Services (DHS) in Allegheny County has built its own IDS. Recently, it began mining the data to develop a Family Screening Tool (AFST), a predictive risk modeling tool designed to improve child welfare call screening decisions. The objective was to develop a decision aid to support hotline screeners in determining whether a maltreatment referral is of sufficient concern to warrant an in-person investigation. Researchers carefully evaluated the accuracy of the underlying data and the precision of the estimates. They are now implementing an algorithm to process many integrated, longitudinal data records to predict the risk of maltreatment in a more consistent and informed way that can be used in practice. Here's a link to the full report.

Utah's - Intergenerational Poverty: The State of Utah's Intergenerational Poverty Initiative used linked administrative records to identify four groups among the population receiving public assistance: 1) those adults using public assistance that had also relied upon public assistance as children (intergenerational poverty or IGP adults) and 2) their children (IGP children) and 3) those adults for whom they did not have a record of childhood use of public assistance (non-IGP adults) and 4) their children (non-IGP children). They compare these four groups against state averages looking at indicators of child well-being (health, education, early childhood and financial stability) and workforce-related variables (educational attainment, wages, employment). The rate of substantiated cases of abuse/neglect was much higher among IGP groups than non-IGP groups. Here's a video describing what they did.

Wisconsin - Modeling Complex Family Structures: Multiple partner fertility leads to increased family complexity, which can have various policy implications. Of particular relevance are determinations about the amount of child support to be provided when children are spread across multiple families. Researchers used linked administrative records data from Wisconsin to document the incidence and evolution of family complexity from the perspective of children using a cohort of first born children whose mothers were not married at the time of their birth. Here's a link to the report.

10. Future Directions

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 close to $2 million in its IDS efforts, with the majority of funds committed on a project-by-project basis to support targeted planning or evaluation efforts. This piecemeal approach has garnered impressive results, though the Poverty Center's leadership recognizes 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.

In the near term, the Center has identified two areas that require the investment of significant time and resources: 1. Improving the efficiency and functionality of the system; and 2. Enhancing the impact and value proposition of CHILD for the community.

1. **System improvements**

- Fully integrate and transition into using the ChoiceMaker software;
- Improve ETL processes to enhance the ability of researchers to select study populations and create analysis files.
- Improve integration to increase the number of data sources, track multiple generations/families
- Develop a system of metadata documentation (data codebook); and
- Develop a business model and governance structure that maximizes sustainability and excellent stewardship.

2. **Value Proposition**

- Engage the institutions, agencies and programs that can benefit from analysis of CHILD system data;
- Support decision-making on the impact of social programs to ensure effective funding allocations by local, state and federal government;
- Contribute to the understanding of best practices in government and human services programming;
- Enhance existing, and enable the creation of new community partnerships and solutions;
- Inform philanthropy in its efforts to make program-related and mission-driven investments; and
- Encourage multi-sector collaborations and investments through social innovation financing models.