THE CONSORTIUM
stimulating self-sufficiency & stability scholarship

USING & LINKING ADMINISTRATIVE DATASETS FOR FAMILY SELF-SUFFICIENCY PROGRAMS

EMILY R. WIEGAND and ROBERT M. GOERGE | Chapin Hall at the University of Chicago

Linking together administrative data sources from programs managed at the state and local levels is a prerequisite to many analyses of program effectiveness and family well-being. However, the availability and quality of data on individual and family characteristics vary between programs, with implications for how data sources can be linked and used. This paper documents a series of interviews with record linkage practitioners and data experts who routinely work with state and local administrative data in the human services. We provide detailed summaries of data sources commonly used in family self-sufficiency research and describe the strengths and weaknesses of data elements for identifying individuals and families across sources. We then highlight several overall characteristics unique to the human services record linkage use case. This report is intended for analysts and researchers looking to develop new sources of linked data to support family self-sufficiency research and analysis or to understand the particular challenges these data sources represent for record linkage. We explore the methodological implications of these challenges in more detail in a companion report (Wiegand & Goerge, 2019b) and statement of recommendations (Wiegand & Goerge, 2019a).

FSSDC
Family Self-Sufficiency Data Center

CHAPIN HALL
AT THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO
Harris Public Policy

This research was supported by the Family Self-Sufficiency Research Consortium, Grant Number #90PD0272, funded by the Office of Planning, Research, and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services. The Family Self-Sufficiency Data Center (FSSDC) facilitates the use of administrative data by researchers and administrators to improve understanding of and identify methods for increasing family well-being. The authors would like to thank the interview subjects for their time and consideration and David McQuown for research assistance. We also thank Nicole Deterding and Chris Kingsley for their review and insightful feedback on an early version of this report. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

Recommended citation:
Wiegand, E. R. & Goerge R. M. (2019). *Using and linking administrative datasets for family self-sufficiency research.* Washington, DC: Family Self-Sufficiency and Stability Research Consortium.

EXECUTIVE SUMMARY

Researchers and policymakers increasingly appreciate that individuals experiencing adversity seek help from or are served by multiple government entities. They also recognize that the long-term impacts of social service programs are not comprehensively visible in any single data source. Answering many research questions around family self-sufficiency or family well-being in the context of government programs requires access to linked administrative data sources from systems such as child welfare, most public benefit programs, education (early childhood, K–12, and higher), workforce development and job training programs, criminal justice at various levels, health care, and labor. Because of how various government programs are funded and structured, the disaggregated data on these topics are almost entirely collected at state and local levels, although sometimes state and local agencies submit these datasets to federal agencies.

The process of combining those datasets is called "record linkage." Record linkage is not a purely technical task. In particular, assumptions about comparability between and sometimes within datasets require careful examination. In this paper, we explore human services data sources and discuss the strengths and limitations of each source for linking populations. Our goal is to understand the characteristics of family self-sufficiency programs and datasets that impact the choice of record linkage approach and the corresponding results.

Methodology. We conducted semi-structured interviews with analysts and managers from four integrated data systems operated with a focus on human services. Additionally, we draw on the authors' own collective experience developing and maintaining the Integrated Database on Child and Family Programs in Illinois at Chapin Hall at the University of Chicago (Goerge & Lee, 2002; Goerge, Van Voorhis, & Lee, 1994; Kitzmiller, 2013).

Results. Key findings include the following:

- Most record linkage among family self-sufficiency datasets relies on names, birth dates, and Social Security numbers (SSNs).
- Even within the same datasets, the quality of identifiers is often inconsistent. The relative quality of this information varies based on who the subject is and how they are engaged with the system. For example, information about foster children is very strong while data on child recipients in public assistance cases are relatively weak. The age of the data in question also correlates with data quality. Where respondents used data sources stretching back decades, they often reported significantly greater quality concerns and limitations in older data.
- Address quality is best from data systems that mail checks (e.g., child support and public assistance), although the increasing prevalence of electronic funds transfers and debit cards raises questions about the long-term viability of this quality.
- Vital records are best for capturing relationships between parents and children (for biological parents). Another source for this information could be in systems for which the role of parent or guardian is central: child welfare, child support, and child care subsidies. Only child welfare data were reported as a quality source of information on (some) extended family relationships. However, public assistance or housing data may provide opportunities to connect individuals in families without detail about specific relationships.

Discussion. Our survey of state and local administrative data sources that are frequently linked to assess questions of family self-sufficiency highlighted several characteristics of this record linkage use case. We explore the methodological implications of these characteristics in more detail in our companion report and recommendations statement (Wiegand & Goerge, 2019b, 2019a).

These datasets focus on specific subpopulations with unknown rates of overlap; the links rely on a small subset of limited identifiers. These realities increase the challenge of assessing linkage quality. Furthermore, the quality and availability of particular data elements vary markedly within and across data sources. This variation increases the challenge of using matches on these elements to identify pairs of records representing the same individual.

Recommendations. Although there are no easy ways to mitigate these challenges, they highlight the importance of understanding the data sources that form component parts of the link to anticipate quality concerns as much as possible.

Additionally, links between state and local administrative data sources can be used to identify and address limitations in the source datasets, either by comparing the same data points across systems or by using a data source to bridge the gap between two sources that do not share any identifiers in common.

INTRODUCTION

Concepts such as family self-sufficiency or family well-being are innately broader than any single public assistance program or social service. As a result, the research to build evidence about self-sufficiency needs and the effectiveness of interventions has relied on information from not only an array of public benefits (such as TANF, SNAP, Medicaid, child care subsidies, disability, housing assistance, unemployment insurance, and public education) but also data from public systems affecting family composition and stability such as child welfare, homeless shelters, and criminal justice as well as private sources of economic support such as employment and child support.

The good news for the family self-sufficiency evidence base is that relatively high-quality administrative data sources exist for all of these topics. However, without integrating data across sources, it is impossible to embark on research or analysis to approximate or understand family self-sufficiency. Data that are linked or integrated, with records for the same individual identified and connected across data sources, unlock important contextual and outcome measures. As Penner and Dodge write, “the potential for insight grows exponentially as data are integrated” (2019, p.8). Record linkage was identified as a key theme in our prior research on improving the use of data around family self-sufficiency (Weigensberg et al., 2014).

Record linkage is a broad methodological category that includes an array of tasks and decisions: for example, what data elements are compared, how these data are standardized, and what level of certainty is placed on the presence of a shared identifier (e.g., name, SSN). Furthermore, while record linkage commonly refers to identifying multiple records for the same individual across data systems, studying *family* self-sufficiency necessitates another kind of linking between records—identifying parents and family members.[1]

The process of combining administrative datasets should not be viewed as a purely technical task; any assumptions about comparability between and sometimes even within datasets require careful examination. Loukissas writes, “All data are entangled with places, institutions, processes, and people that fundamentally shape their significance and use. If we haven’t understood the data’s setting, we haven’t understood the data” (2019, p.189). This statement is particularly relevant to the use of administrative data (which are, by definition, collected for nonresearch purposes) for research and analysis. Both the integration process and the subsequent research should be rooted in an understanding of the context of individual datasets and records, with implications for quality, usefulness, and interpretation.

We wanted to understand the characteristics of family self-sufficiency programs and datasets that impact record linkage methodological choices and the outcome of a match. To this end, we interviewed analysts and managers from four integrated data systems operated with a focus on human services. We incorporated our own experiences creating and maintaining the Integrated Database on Child and Family Programs in Illinois (Goerge & Lee, 2002; Goerge et al., 1994; Kitzmiller, 2013). Across the integrated data systems we included in our sample, all of the family self-sufficiency data sources described in the first paragraph of this section were represented at least once. Most were cited frequently. Additionally, some respondents mentioned using vital records such as birth certificates to identify individuals and relationships.

In this paper, we discuss the strengths and limitations of each data source for record linkage. We then summarize key findings and conclusions across the range of family self-sufficiency data sources. We present this report together with two complementary publications. The first summarizes the record linkage methodological literature, with particular attention to approaches that may be valuable for analysts linking human services data sources (Wiegand & Goerge, 2019b). The second lays out a series of threats to the generation of high-quality linked human services data and the rigor of research conducted on those data, along with recommendations to address these issues (Wiegand & Goerge, 2019a).

Limitations

This paper does not discuss legality, privacy, or ethics regarding using and linking administrative datasets. However, these are important topics. They become only more pressing when we recognize that record linkage requires access to personal identifiers such as name, birthdate, and SSN, and that record linkage in the human services is often focused around vulnerable populations, such as children and poor families. There are a number of industry best practices to reduce both ethical and data security risks in record linkage, including routinely separating personal identifiers from other data so that no analyst views identifiers and program or outcome data at the same time. We suggest Actionable Intelligence for Social Policy as a source for introductory resources on these topics.[2] The 2017 report from the federal Commission on Evidence-Based Policymaking also focuses on many of these issues (Commission on Evidence-Based Policymaking, 2017).

This report and the linkage processes discussed in our interviews focus on record linkage to create datasets for research and analysis as distinct from real-time data integration to support case management.

METHODOLOGY

We contacted a number of organizations that are well-known industry leaders in linking human services data. We conducted semi-structured interviews with individuals from each organization who had detailed familiarity with the organization’s record linkage process and at least some of the human services datasets used by the organization. The primary focus of these interviews was the strengths and weaknesses of particular human services datasets for record linkage purposes. However, we also discussed various technical and methodological approaches the interviewees had explored or considered to address these challenges.[3]

Additionally, we drew on the authors’ own collective experience developing and maintaining the Integrated Database on Child and Family Programs in Illinois at Chapin Hall at the University of Chicago (Goerge & Lee, 2002; Goerge et al., 1994; Kitzmiller, 2013).

Our focus in this report is on the usability and quality for record linkage of various datasets, not on dataset availability. What data are available for linkage and analysis vary tremendously across jurisdictions and over time; our respondents were not sampled in any way to attempt to adequately represent questions of availability at the national scale. However, by concentrating on organizations with years or even decades of experience linking human service data, we aim to identify prevailing quality and usability patterns that appear durable over time and across a varied set of contexts. To the extent that we comment on the availability of datasets, we do so to contextualize our respondents’ familiarity with particular data sources.

RESULTS

The following discussion of data sources is primarily based on data sources collected by state and local government agencies or private agencies under contract to government. Individuals at these organizations collect the data—often capturing it in a computer application—and submit it to a central electronic repository. In some cases, extracts of the state and local government agency data are sent to a corresponding federal agency for compliance purposes. In a subset of these, identifiers are included and the opportunity exists to link these extracts to other datasets held by the federal agencies. However, we focus on these data in the state and local context.

Public Assistance Data

For most respondents, public assistance data generally comprise data on TANF, SNAP, and Medicaid eligibility and enrollment. They may also include disability assistance or other state assistance

[1] This is a complex question in and of itself. Research often focuses solely on connecting parents with children. But whether a family is defined to include siblings, other relatives who share a household, relatives who do not share a household, or nonrelatives (such as unmarried partners) in the same household impacts the available administrative data sources to document these relationships. See Goerge & Wiegand, 2019 for more discussion of options and challenges identifying families in administrative data.

[2] See www.aisp.upenn.edu.

[3] See Appendix 1 for full list of interviewees and Appendix 2 for our interview protocol.

programs, such as refugee assistance. Importantly, these data usually come from an integrated eligibility system (i.e., a common database for data entry and management); no respondents discussed record linkage among public assistance datasets, even though many were using public assistance data stretching back more than a decade. These data systems generally maintain integrated eligibility and enrollment information but may not include details of participation in work activities or specific services received.

Respondents agreed that public assistance data are generally a good source for core identifiers such as names, birthdates, and SSNs. Participants in assistance programs are usually required to demonstrate their identity, and SSNs are often externally validated. However, the quality of these data are not uniform: challenges mentioned include missing or incomplete information for children, especially when first entered (e.g., records for "Baby Smith") and quality problems with SSNs in older records (this was reported by respondents working with several decades of data and likely reflects old systems or practices).

Public assistance was frequently cited as a good data source for addresses. Historically, participants received benefits by mail, so address information was complete and regularly updated. However, the increasing prevalence of electronic benefit receipt reduces this incentive and raises questions about whether public assistance data will continue to have strong address information moving forward.

Because some benefits are received at the family (TANF) or household (SNAP) level, case identifiers in public assistance data can sometimes be used to connect family members (Goerge & Wiegand, 2019). At least one respondent indicated that the public assistance data they used also listed specific relationships among individuals in the family, but these data were often incorrect (for example, an aunt listed as a mother). A number of other respondents cited the lack of coded relationship information as a limitation in using public assistance data to characterize families. In some cases, gender and birthdate had been used to identify likely parents within households. Where Medicaid claims are available, Medicaid birth records could also be used to define relationships for the subset of individuals involved in births paid for by Medicaid.

Housing and Homelessness

Data from homeless shelters and services were mentioned multiple times as a potential source of information about family units. They are a good source since individuals entering the shelter together are often recorded together and relationship information may also be recorded. Any births occurring while the family is in the shelter will be recorded. However, comparisons between shelter data and child welfare data in one jurisdiction highlighted a potential limitation to data on these family units: mothers who go to a shelter may leave some or all of their children with other family members or friends. These children are not captured in homelessness data.

Housing subsidy data can be a good source of addresses for current participants, but not for applicants, who are likely to lack stable housing. This divergence creates research challenges; applicants who do not receive subsidies represent an ideal control group but may be difficult to place geographically unless they can be tracked through another data source.

Education

Among our interviewees, most discussed K-12 education data, though access to these datasets was often particularly limited geographically, temporally, or in the breadth of data elements available for research. The school data used by respondents only infrequently contained SSNs. Overall quality of personal identifiers in the data was characterized as average across datasets. In a few areas, respondents thought schools could be a data source for information that would be hard to find in other systems, such as nicknames or parent–child relationships. However, these data were often not stored or shared in such a way that they could be used for record linkage.

School data have obvious geographic connotations. While they may contain addresses, those addresses may or may not be routinely updated. In some jurisdictions, school enrollment serves as a proxy for location. However, in jurisdictions where school choice is prominent, this connection is lost.

Our respondents did not speak to higher education data or job training data. Although these data sources are potentially important for understanding pathways to self-sufficiency, they are not usually managed by state or local jurisdictions. As a result, these pathways are difficult to systematically incorporate. Because young adults may attend a college or university or get a job outside of the city where they attended high school, analyzing postsecondary outcomes comprehensively requires national data sources, such as the National Student Clearinghouse.

Public data systems frequently capture self-reported educational attainment information for participants, but these data are generally suspect. There are few incentives for most systems to capture this information correctly or keep it updated. The exceptions are workforce development systems, where these data become a key characteristic of the participant.

We did not explore use of datasets specific to early childhood education, such as data from Head Start programs or preschools. Where these programs are funded through school districts, preschool or prekindergarten data may be included in K-12 education datasets. However, the landscape of early childhood education provision is even more fragmented than that of K-12 education (not only are services managed at the local level, but funding sources also vary meaning the same locality is usually served by a number of distinct entities). As a result, systematically collecting and aggregating these data sources, even for small geographies, can represent a significant data acquisition and integration challenge. For more information about efforts to integrate early childhood education data, see the U.S. Department of Health and Human Services and the U.S. Department of Education's 2016 report, *The Integration of Early Childhood Data* (U.S. Department of Health and Human Services & U.S. Department of Education, 2016).

Child Care Subsidies

Child care subsidy data were used by only a handful of the sites we spoke to. While these data have value to link parents and children and frequently include SSNs, at least two sites mentioned high levels of duplicated individuals over time in these data. In one location, child care subsidies were tracked in an integrated eligibility system with the other public assistance programs.

Child Welfare

Respondents emphasized the variability of quality in child welfare data perhaps more than any other data source. Specifically, the child welfare system includes data on a heterogeneous population, and data quality is dependent on the individual's level of engagement with the system. Demographics and personal identifiers, often, but not always, including SSNs, tend to be quite good for children and parents who engage significantly with the system, such as children placed in foster care. Information on individuals involved with investigations, especially short, unsubstantiated investigations, are missing much more often and are more likely to be of poor quality.

Similarly, data on siblings, other family members, or other individuals involved with an investigation will be much less complete than on children who are alleged victims and on their parents.

Respondents frequently cited child welfare data as one of the strongest sources for understanding parent/child relationships, as well as potentially identifying other relationships, such as siblings and extended family. These data are also a potential source for addresses, though addresses are only updated for children and families who remain engaged with the system.

Criminal and Juvenile Justice

The criminal justice system is highly fragmented across jurisdictions. This reality translates into data coming from a range of different systems: local police arrest records, county jail and court records (both adult and juvenile), and data from state prison and parole systems (both adult and juvenile). Interviewees had experience with different parts of this system. However, our small pool of respondents did not show any patterns specifically related to level or type of criminal or juvenile justice data.

Respondents frequently mentioned data challenges with criminal and juvenile justice data, including the presence of large quantities of unverified, self-reported information on identifiers such as name, birthdate, or SSN (also frequently missing). The self-reported information often did not match values in other systems. Within the criminal justice system, jurisdictions rely on fingerprints to uniquely identify individuals and may place less weight on traditional identifiers. While system identifiers that correspond to fingerprints are available across jurisdictions (such as between the local police and county court), these identifiers can be used to link records but cannot be independently validated. In addition, these identifiers have no value for linkage to records outside the criminal justice system.

Several respondents noted that criminal justice data can track and store a history of known aliases. This is a potentially important element in catching nicknames or name changes across other systems.

Even for juvenile offenders, juvenile justice data do not seem to be a good source for parent and other relationship information.

Employment and Wages

Our respondents consistently used unemployment insurance wage records to track employment and wages. These records are typically maintained by states on a quarterly basis. They do not include out-of-state employment, self-employment, or certain job types (such as farm workers, domestic workers, and federal employees).

Access to wage data is often heavily curtailed. In almost all cases, respondents did not link directly to wage datasets, but instead sent SSNs, and sometimes names, to the state employment agency to perform this link. SSN is generally very important in linking wage data since the presence and quality of other identifiers in these datasets are very limited.

Despite all of these limitations, it is important to capture a measure of unsubsidized employment in studying family self-sufficiency. Therefore, unemployment insurance wage data remains a key data source. It is also the only somewhat comprehensive source of this information with analytic availability for most jurisdictions.

Other data sources sometimes discussed for employment and wage information are the National Directory of New Hires (NDNH) database[4] and tax records. However, there are strict legislative limits on how and by whom these data may be used. None of our interview respondents had experience with or access to NDNH or tax data.

Child Support

Respondents did not frequently use child support data. Where they are used, these data provide a particularly strong source of parental relationship information. In general, SSNs in child support data are reliable, and where payments are still received via check, addresses for custodial parents are particularly likely to be high quality.

Vital Records

Although birth and death records do not have a direct family self-sufficiency connection, they can be useful for several aspects of understanding trajectories. Birth records provide a canonical source connecting parents and children and for children's birth dates; death records are one of the only available sources of death information and dates. For record linkage purposes, vital records can provide a backbone: a core cohort to follow over time or a gold standard for an individual's name.

Summary of Strengths and Weaknesses across Data Sources

Most record linkage between family self-sufficiency datasets relies on names, birth dates, and SSNs. Quality SSNs are not consistently available and are particularly likely to be missing in data systems managed at the local level—such as education or county courts—compared to state-level systems. SSNs are typically more accurate in the Social Security Act Title programs (Titles IV and XIX, in particular). However, SSNs are generally required for linking to wage data, since these datasets include minimal other identifiers.

Even within the same datasets, the quality of identifiers is often inconsistent. The relative quality of this information varies based on who the subject is and how they are engaged with the system. For example, information about foster children is very strong while data on child recipients in public assistance cases are relatively weak. The age of the data in question also correlates with data quality. Where respondents used data sources stretching back decades, they often reported significantly greater quality concerns and limitations in older data. This pattern is not necessarily due to inherent changes in accuracy over time but may reflect a lack of knowledge about how data elements were collected historically or the impact of conversions from legacy data systems.

Address quality is best from data systems that mail checks. In the case of child support and public assistance, recipients have historically been likely to regularly update their addresses for the purposes of receiving checks by mail. However, the increasing prevalence of electronic funds transfers and debit cards puts this quality in doubt for the long term.

Relationships between parents and children are captured best in vital records and in systems for which the role of parent or guardian is central: child welfare, child support, and child care subsidies. Only child welfare data were reported as a quality source of information on extended family relationships, but public assistance or housing data may provide opportunities to connect individuals in families without detail about specific relationships.

DISCUSSION

Our survey of state and local administrative data sources that are frequently linked to assess questions of family self-sufficiency highlights the following characteristics of this record linkage use case. We explore the methodological implications of these characteristics in more detail in our companion report and recommendations (Wiegand & Goerge, 2019a, 2019b).

First, it is particularly difficult to assess the quality of a link between these data sources because there are few sources of "truth" against which match results can be compared, either individually or in the aggregate.

[4] NDNH is operated by the federal Department of Health and Human Services Administration for Children and Families Office of Child Support Enforcement.

Second, most of these links rely on limited identifiers, chiefly names, birthdates, and SSNs. SSNs are inconsistently collected across and within systems. The likelihood of even an exact name and birthdate match representing a single person depends on the geographic scale, the frequency of the name and birthdate within the population in question, and the likelihood of overlap between the data sources. It is entirely possible to have matches of this type that even a human reviewer cannot assess with any certainty.

Furthermore, these datasets are focused on specific subpopulations with unknown rates of overlap. Addressing questions of family well-being often involves linking specific subpopulations contained in agency databases (for instance, program recipients, incarcerated individuals, employed individuals, etc.) across agency lines and sometimes across jurisdictions. Often there are few existing numbers to validate the expected rate of overlap between two datasets; unless a project is able to access vital records data, it is relatively uncommon for an analyst to expect that one dataset will contain all or even most of the records in a second dataset.

In addition to challenges assessing linkage quality, the variability in quality and availability of particular datasets within and across data sources increases the challenges of rigorous linkage (see "Patterns of Data Quality Variation within Datasets"). Quality tends to be better for:

- individuals who are more closely engaged with the system, such as foster children, as compared with children investigated for a single, unsubstantiated allegation of abuse or neglect;
- more recently collected data; and
- data points that are central to the business use of the data, such as relationship data in vital records or address data in systems that mail checks.

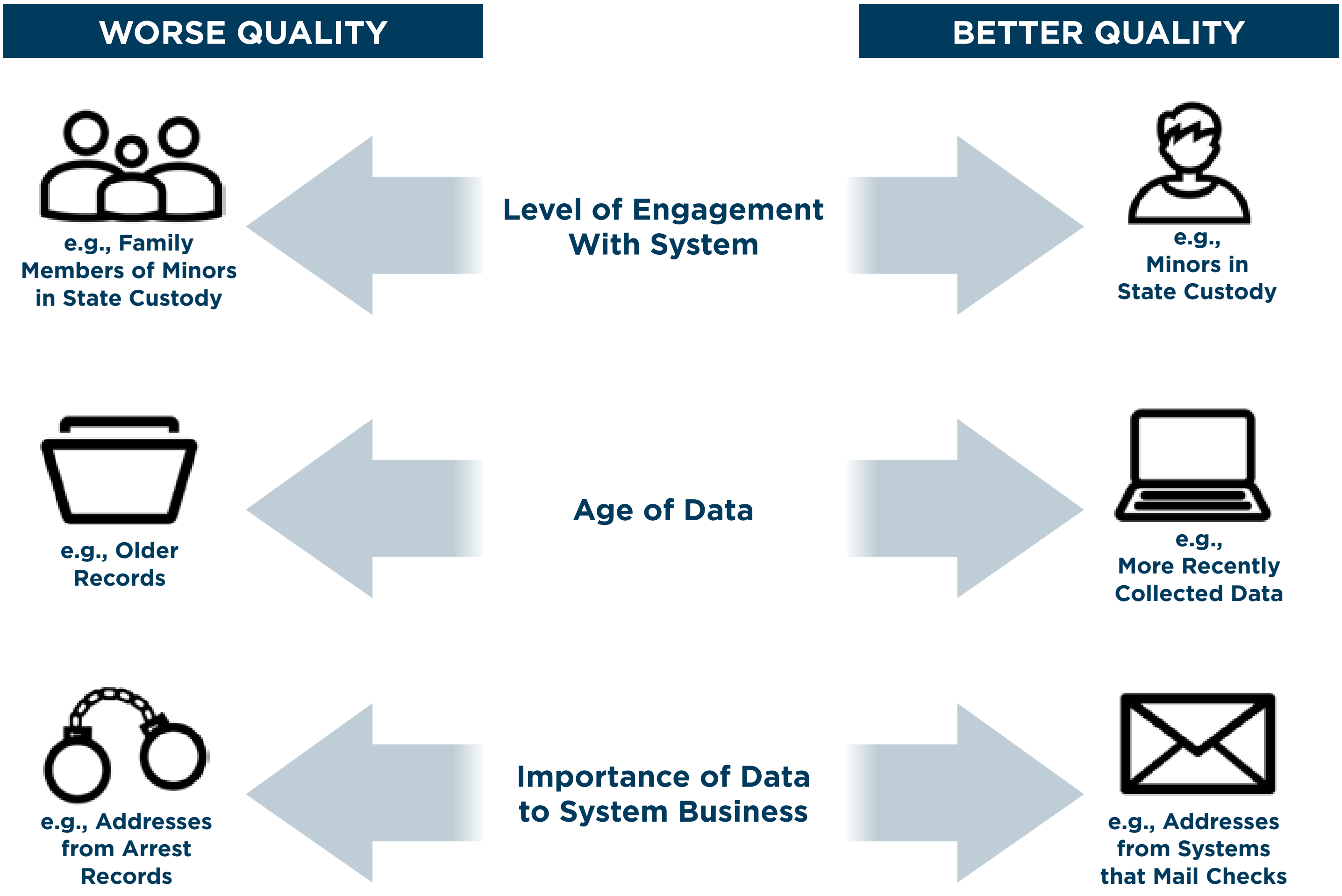
Record linkage relies on treating all matches or non-matches on a given data element according to consistent logic; variation in quality across records for the same data element challenges a basic methodological assumption.

RECOMMENDATIONS

Links between state and local administrative data sources can also be used to identify and address limitations in the source datasets. Where data from one system are known or expected to be accurate, they can be used to understand the opportunities and limitations of another data source. Comparing the same information across two systems can highlight unconsidered data challenges, as in the study that found homelessness records may not include all of the children in a family.

Data sources can also be used to bridge gaps where two datasets do not contain identifiers in common. For example, a large state

PATTERNS OF DATA QUALITY VARIATION WITHIN DATASETS



WORSE QUALITY

BETTER QUALITY

e.g., Family Members of Minors in State Custody

Level of Engagement With System

e.g., Minors in State Custody

e.g., Older Records

Age of Data

e.g., More Recently Collected Data

e.g., Addresses from Arrest Records

Importance of Data to System Business

e.g., Addresses from Systems that Mail Checks dataset like public assistance can provide a source of SSNs for individuals in a county jail (where SSN is not tracked). This allows wage data, only available by SSN, to be appended. However, in this approach the intermediate link creates a subset of the initial population in a very specific way (in this example, excluding from the analysis inmates who have never received public assistance).

Although there are no easy ways to mitigate the challenges of linking human services datasets, those challenges highlight the importance of understanding the data sources that form component parts of the link. To the extent an analyst linking two data sources is able to contextualize the expected overlap between those sources, anticipate and address possible variations in data quality for particular elements, and creatively maximize the diversity of identifiers available to define a match, the analyst is well-positioned to develop a high-quality result.

APPENDIX 1: INTERVIEW PARTICIPANTS

We acknowledge the following interviewees with gratitude:

Case Western Reserve University Center on Urban Poverty and Community Development
Nina Lalich

Allegheny County Department of Human Services
Seth Chizeck
Erin Dalton
Catherine Jensen
John Shantz

University of Wisconsin-Madison Institute for Research on Poverty
Steven T. Cook

New York City Center for Innovation through Data Intelligence
Andy Martens

APPENDIX 2: SEMI-STRUCTURE INTERVIEW PROTOCOL

General Background:

What family self-sufficiency (FSS) datasets do you link? *FSS datasets may include SNAP, TANF, Medicaid, child welfare, criminal justice, education, employment, and any other datasets that reflect aspects of family wellbeing.*

How long have you been doing this/how far back does the data go?

Operations:

What's your general process – rounds, real time links, a hash, etc.?

What technologies or software do you use? What is your linking methodology?

Does your linking approach vary depending on your research question(s)?

How do you staff/manage/fund this work?

How do you identify an individual?

Family Self-Sufficiency (FSS) Data Experiences:

For each FSS dataset linked, what are the strengths and weaknesses of the linking fields? What data do you particularly trust or not trust? What are the limitations?

Thinking about timeliness of data quality – are there datasets that are or are not particularly conducive to real-time linking?

What datasets are particularly valuable in identifying families, households, or connecting parents and children?

What are some of the major weaknesses of individual datasets for research and analysis purposes that you address by linking? (For example, adding wage data to get employment outcomes for former recipients, or adding another dataset to pull in family information, etc.)

Thinking about each of the datasets, are there datasets where you think your linking methodology is more or less important and why?

In general, what makes a dataset a good fit for your linking methodology? When do you question your ability to link a dataset?

REFERENCES

Commission on Evidence-Based Policymaking. (2017). *The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking*. Retrieved from https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

Goerge, R. M., & Lee, B. J. (2002). Matching and cleaning administrative data. In M. Ver Ploeg, R. A. Moffitt, C. F. Citro (Eds.), *Studies of welfare populations: Data collection and research issues* (pp. 197–219). Retrieved from https://www.nap.edu/read/10206/chapter/9

Goerge, R. M., & Wiegand, E. R. (2019). Understanding vulnerable families in multiple service systems. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 5*(2), 86–104.

Goerge, R., Van Voorhis, J., & Lee, B. J. (1994). Illinois's Longitudinal and Relational Child and Family Research Database. *Social Science Computer Review, 12*(3), 351–365. https://doi.org/10.1177/089443939401200302

Kitzmiller, E. (2013). *IDS case study: Chapin Hall - Leveraging Chapin Hall's mission to enhance child well-being*. Retrieved from http://www.aisp.upenn.edu/wp-content/uploads/2015/08/ChapinHall_CaseStudy.pdf

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. Cambridge, MA: The MIT Press.

Penner, A. M., & Dodge, K. A. (2019). Using administrative data for social science and policy. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 5*(2), 1–18. https://doi.org/10.7758/RSF.2019.5.2.01

U.S. Department of Health and Human Services, & U.S. Department of Education. (2016). *The integration of early childhood data: State profiles and a report from the U.S. Department of Health and Human Services and the U.S. Department of Education*. Retrieved from https://www.acf.hhs.gov/sites/default/files/ecd/intergration_of_early_childhood_data_final.pdf

Weigensberg, E., Schlecht, C., Wiegand, E., Farris, S., Hafford, C., Goerge, R., & Allard, S. (2014). *Family Self-Sufficiency Data Center: Needs assessment report*. Chicago, IL: Chapin Hall at the University of Chicago. Retrieved from https://www.chapinhall.org/wp-content/uploads/FSS_Data_Center_Needs_Assessment_Report_Final_0.pdf

Wiegand, E. R., & Goerge, R. M. (2019a). *Recommendations for ensuring the quality of linked human services data sources*. Washington, DC: Family Self-Sufficiency and Stability Research Consortium.

Wiegand, E. R., & Goerge, R. M. (2019b). *Record linkage innovations for the human services*. Washington, DC: Family Self-Sufficiency and Stability Research Consortium.