THE CONSORTIUM
stimulating self-sufficiency & stability scholarship

RECOMMENDATIONS FOR ENSURING THE QUALITY OF LINKED HUMAN SERVICES DATA SOURCES

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

There is a great deal of interest in the public sector and among researchers in linking public sector administrative datasets, particularly from state and local agencies, to better understand the impacts of human service programs, to target services, and to assess questions of family well-being. We lay out several threats to the quality of those data sources and the validity of associated research. First, the record linkage research literature and existing best practices do not discuss this use case, which has important, unique methodological characteristics. Second, it is difficult to assess match quality and even harder to compare quality across different matches. Finally, the current trajectory of research and software development is toward expanding the use of complex or proprietary methods at a cost to transparency and without specific consideration of these quality questions. We lay out recommendations for practitioners and the field for paying more attention to linkage quality and the ultimate accuracy of linked data sources.

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 Julia Lane, Steven T. Cook, and Nick Mader for their review and insightful feedback on an early version of this work. 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). *Recommendations for ensuring the quality of linked human services data sources.* Washington, DC: Family Self-Sufficiency and Stability Research Consortium.

EXECUTIVE SUMMARY

The data systems that track families' experiences with human services programs and the data sources that speak to those families' barriers and outcomes are fragmented across agencies, jurisdictions, and technologies. Researchers and policymakers seek to integrate data across these systems to understand and design policies that serve families holistically.

Despite significant interest and investment in increasing the prevalence of these linkages (what we call the human services use case for record linkage), we identify a series of threats to the continued development of high-quality matched data. Critically, poor match quality can create systematic biases in the linked data, undermining the rigor of any analyses conducted on those data.

Rationale. We identified three features of the human services record linkage use case that threaten the quality of the resulting linked data:

1. The human services use case for record linkage is unique in several important methodological ways, including variable data quality, unknown rates of overlap, and limited identifiers. However, it is not a source of current methodological research or conversation.

2. There are no defined standards to assess match quality or best practices with regard to methodology.

3. Current research and commercialization trends encourage the adoption of increasingly complex methodologies and black box tools, reducing transparency with regard to how linked datasets are created.

Recommendations. Ultimately, we put forward two recommendations.

In the short term, we recommend that practitioners adopt simpler and more standardized record linkage techniques that scale well. Practitioners should emphasize using these more scalable methods to introduce routine use of sensitivity testing in the analysis of integrated data.

We also encourage the active development of a research space specific to the human services record linkage use case. Practitioners and the policy community often expect a silver bullet and need to better understand the limitations of any algorithmic solution. All parties must discuss and keep front of mind the significant ramifications for policy and the public if a record linkage solution perpetuates biases. Therefore, the field needs to develop best practices with regard to methodological transparency.

INTRODUCTION

Government services, particularly in the human services, are notorious for being managed and assessed in silos (Farhang and Yaver, 2016). Families receiving services such as medical assistance, cash assistance, food assistance, housing subsidies, and childcare subsidies must balance divergent, sometimes contradictory, eligibility and verification processes. As families are referred for additional supports (e.g., medical treatment, counselling, etc.) these referrals are similarly decentralized. Supports do not reflect other ways the family may engage with government, as in the child welfare or criminal justice systems. These programs are, almost universally, administered at the state or local level, leading to further fragmentation across jurisdictions.

Data practices in the human services are an important component of this separation, since assistance programs and other family services frequently maintain distinct data systems and routinely do not share unique identifiers for individuals across systems. This situation severely limits opportunities to use program data to better understand and serve the families who engage with public services (Commission on Evidence-Based Policymaking, 2017).

Jurisdictions are frequently looking to serve families holistically, avoid duplication or gaps in service delivery, and understand the impact of policies on long-term well-being. These jurisdictions usually identify data integration as a goal. Data integration, or record linkage, is the process of identifying common entities (individuals, families) across those data systems. At the Family Self-Sufficiency Data Center, our goal is to foster data use so that agencies can truly address these questions of well-being. Central to that process is linking state and local administrative data sources on program participation, criminal justice and child welfare involvement, education, and employment.

There is significant interest and investment in what we call the human services use case for record linkage. Despite this interest, we have heard repeatedly from researchers and organizations seeking to complete these matches about their challenges understanding what was necessary to create high-quality, linked data sources. We embarked on a course of research to address these challenges, including practitioner interviews (Wiegand & Goerge, 2019b) and a literature review on record linkage methodology (Wiegand & Goerge, 2019a).

These recommendations are grounded in that research and stem from a series of threats we perceive to the continued development of high-quality matched data from state and local administrative data sources.

The potential cost of poor match quality is great. There is a very real risk that inappropriately applied matching algorithms can yield incorrect results with systematic biases. These results can then undermine any downstream analyses. Analyses of record linkage in epidemiological studies have captured and highlighted these risks (Brenner, Schmidtmann, & Stegmaier, 1997; Harron, Wade, Gilbert, Muller-Pebody, & Goldstein, 2014; Krewski et al., 2005). The research and policy communities are currently focused on developing more sources of linked data. However, they are not facing head-on questions of how to define and characterize rigor. Record linkage is often considered a technical or operational question, but the reality

is that research is only as good as the data upon which it rests. In the rush to develop the policy evidence base, we risk doing more harm than good by sidestepping the question of rigor or quality measurement in record linkage.[1]

RATIONALE

We identified three features of the human services record linkage use case that threaten the quality of these linked data (see "Threats to the Integrity of Linked Data Undermine Research Quality").

Unique Characteristics of the Human Services Use Case

The human services use case for record linkage is unique in several important methodological ways, but it is not a source of current methodological research or conversation.

State and local data sources are characterized by variable data quality and contents across—and sometimes even within—datasets. The linking process is frequently complicated by the lack of at least one population-level data source (i.e., one that contains all possible unique records). Most record linkage done with these datasets relies on limited identifiers, particularly names, birth dates, and Social Security numbers (SSNs). These constraints increase the difficulty not only of linking the records, but also of assessing the accuracy of the matched data that result (Wiegand & Goerge, 2019b). In fact, characteristics of the data routinely violate the mathematical assumptions underpinning classical record linkage theory (Wiegand & Goerge, 2019a).

The intended uses for these linked data—which include not only policy research but also analytics for program management—generally require linking across multiple (i.e., more than two) data sources and regularly updating those matches. These uses require a process that is not only rigorous but also at least somewhat scalable. The methodological questions that arise from linking 2–3 comprehensive national data sources differ notably from those arising from linking 10-20 small data sources that are scattered across overlapping jurisdictions and that contain populations for whom the level of intersection is not known.

Furthermore, the margins of error on any statistical technique increase significantly as population sizes decrease. Yet researchers, policymakers, and advocates often most want to understand the relatively small populations that sit where multiple systems overlap, such as the population of youths dually involved with both child welfare and juvenile justice systems. Methodological questions are thus most important in exactly the contexts that are most significant to researchers and policymakers.

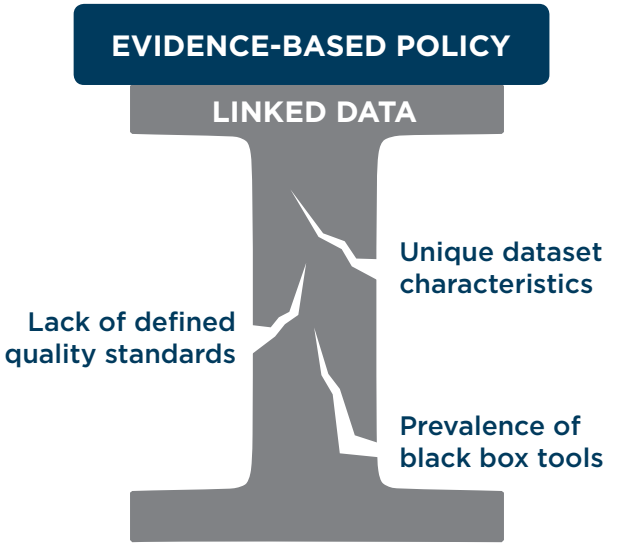
The record linkage methodological literature is rampant with new approaches to an old problem. However despite its breadth, the literature does not offer clear recommendations for which methods best suit which purposes. Elmagarmid and colleagues summarized this problem in a 2007 review:

> A question that is unlikely to be resolved soon is the question of which of the presented methods should be used for a given duplicate detection task. Unfortunately, there is no clear answer to this question. The duplicate record detection task is highly data-dependent, and it is unclear if we will ever see a technique dominating all others across all datasets. . . it is currently unclear which metrics and techniques are the current state-of-the-art. (Elmagarmid, Ipeirotis, & Verykios, 2007, p. 11, p. 13)

The efficiency and accuracy of any one method is heavily influenced by underlying data characteristics, including data quality, the expected rate of overlap between the datasets being matched, data preparation and standardization decisions, and overall file size. To date, there has been no attempt to rigorously control for these factors in testing methods. Furthermore, there are dozens of different ways to implement any record linkage approach. Even when a study compares a "machine learning classifier" to "traditional probabilistic methods," those results are able to tell us little about how the types of methods compare.

Because of what Elmagarid and colleagues describe as the "data-dependent" nature of a match methodology decision, the lack of record linkage research specifically tailored to the human services use case is particularly problematic. Issues of data quality and completeness in human services administrative data look different from problems in health and survey data, two domains that have been key drivers of record linkage research and innovation to date.

THREATS TO THE INTEGRITY OF LINKED DATA UNDERMINE RESEARCH QUALITY



Lack of Standards to Assess Record Linkage Quality

A lack of independent standards with regard to what constitutes a high-quality match creates a major potential challenge to the scientific credibility of policy research and analysis completed using linked data sources.

There is almost never a gold standard truth for these cases—even a human reviewer sometimes cannot tell if two records represent the same person if information is scarce and overlap uncertain. As a result, short of having frontline practitioners confirm linked data with the subjects themselves, actual validation of match accuracy is in and of itself subjective. Studies that have crowdsourced the creation of training data for supervised match methods, using websites like Mechanical Turk, have found that human workers agree on the truth of what is and is not a likely match only about 90% of the time (Getoor & Machanavajjhala, 2013). While this rate will vary with the richness and provenance of the component datasets, this finding is a good reminder that while record linkage methods are attempts to duplicate human logic mathematically at scale, the logic human reviewers use in making a match determination can itself be unreliable.

[1] In our discussions of record linkage in the human services, we concentrate on linking data sources for analytical purposes. While some of our findings are transferable to the challenge of linking data systems in real-time for case management purposes, that use case is not the focus of our research.

The inability to assess the accuracy of any specific match could be mitigated by a clear body of research on what match methodologies are most appropriate to what circumstances. Crucially, except in the case of data created for academic papers to test record linkage approaches, there are no external data sources that can compare the performance of different match methodologies. This lack of an agreed-upon standard makes it difficult for researchers to contextualize the accuracy of an approach or understand what is lost when a detail is ignored or a less manual process is adopted.

Finally, match quality must not be considered in a vacuum, but also as a part of any specific analysis. A match that is rigorous and accurate for one analytic question may not be appropriate for another. For example, where identifiers are weak in administrative data, many practitioners seek to use more and varied contextual fields in record linkage—fields like family members and addresses. While including these elements may improve match accuracy in a vacuum, there are potential biases for certain research questions. Goeken, Huynh, Lenius, and Vick describe a similar scenario in matching historical census data. Ultimately, the authors opted not to include geography or household composition in their matches because the resulting data are frequently used for studies about mobility and household change. This research would be biased if the same elements were components of the match, for example decreasing the relatively likelihood of a match for more mobile households (Goeken, Huynh, Lenius, & Vick, 2011).

Risk of Decreased Transparency

We have written in more detail about approaches to record linkage, particularly new research and innovations and potential applications to the human services use case (Wiegand & Goerge, 2019a). In that work, we identified record linkage approaches with potential to address analysts' desire to incorporate more and more complex features in linkage models in less manual ways. Analysts would like to incorporate items such as values that change over time, addresses, and relationships. We also identified new areas of research with implications for scalability, addressing the challenges of linking more than two datasets at a time and frequently updating existing links.

Current applied record linkage methodologies have limited ability to incorporate these complexities, and lack of resources for practitioners forces them to ignore some of the potential richness of the data. There is a pronounced desire for computational or statistical solutions that can incorporate more complexity without requiring more manual labor. This demand will only increase as more jurisdictions, agencies, and universities emphasize the value of integrating administrative data sources for research and analytic purposes. The demand will also increase as the public and policymakers put greater focus on measuring program outcomes and impacts through administrative data sources.

However, as record linkage methods become more complex and sophisticated, they also risk becoming less transparent and interpretable. Currently, matches are completed without any independent standards for quality or rigor, and researchers and analysts seeking to use linked data continue to disregard match decisions or view them as tangential to analytic decisions. If these patterns do not change soon, it will be much more difficult to change them, as matching becomes even more of a black box operation.

Record linkage approaches are being commercialized, and transparency is further degraded by this process. Unlike the sciences, where open source methods and software are increasingly the norm, translation from theoretical research to practice in record linkage is frequently left to vendors and private companies. These entities have incentives to sell comprehensive solutions that they claim can integrate datasets from any domain, including the human services. These solutions ignore variation across use cases and downplay the limitations of their own methods. (These entities also often have minimal familiarity with the state and local administrative data context.) Even where vendors disregard these incentives and clearly state the purposes and limitations of their tools, users face a lack of tailored options.

RECOMMENDATIONS

We make two recommendations to address these threats. One is intended for practitioners with concerns about maximizing quality in the near-term; the other is our recommendation for the full community of stakeholders (researchers, policymakers, and practitioners) as we look to the future.

We make these recommendations under the assumption that the specific characteristics and challenges of the human services record linkage use case are unchangeable. This assumption is not entirely true. Potential ways to reduce the challenges of record linkage for these datasets include (a) increasing the overall quality of the data and the breadth of identifiers collected or (b) expanding access to population-level datasets (such as vital records or tax data) for record linkage purposes. Today, administrative data are collected for both administrative and analytic purposes. Policymakers and government leaders could adapt data collection and management practices in ways that better facilitate record linkage. However, given the sheer scale of this use case—the number of jurisdictions and agencies involved—this kind of policy change would take significant time. Even then, challenges will remain where, in historical datasets or due to security, ethical, legislative, or political reasons, data collection or management practices cannot be changed.

In the short term, we recommend that analysts and organizations seeking to conduct rigorous, but financially feasible, record linkage step away from the notion of one-size-fits-all record linkage. There is no single solution, model, or algorithm that will best fit all needs. By trying to develop one (especially in the absence of any true gold standard against which to validate our processes), we risk investing our resources in addressing the wrong problems and overlooking the gaps that truly bias results.

In making this recommendation, we are not saying that every question or problem requires a totally new approach. Instead, we recommend that organizations adopt record linkage techniques with an emphasis on efficiency and scale (requiring a minimum of manual review and adjustment). Adopted techniques should have adjustable parameters that can be used for sensitivity testing. Most probabilistic and classification methods fit this approach well, but even deterministic logic can be designed to generate results with different tradeoffs between false positives and missed matches when applied hierarchically. (For example, a "level one" match requires an exact match on name, birthdate, and SSN; a "level two" match matches on Soundex code of name, birthdate, and SSN; and a "level three" match matches on two out of those three characteristics).

A match process that, by default, yields results at approximately three different levels of precision allows analysts to specifically test the sensitivity of a given question to match methodology. If results are largely the same across the different match levels, than the routine match logic is likely appropriate for this question. If sensitivity testing produces significant swings in the results, however, it is a sign that match logic is extremely important in the final outcome. For example, an analysis of employment outcomes for high school graduates in a city by race, age, and gender is likely robust to some errors in matching, because mismatches are likely to be among people with similar characteristics who will

ultimately be aggregated in the same subpopulation. An evaluation of employment outcomes conducted on a small population of individuals enrolled in a job training program would potentially be more sensitive to the match methodology.

In cases where varying the match criteria impacts the analytic results, it may make sense to design a custom match. This match will likely involve a combination of probabilistic or classification and deterministic methods, specific to the question at hand. These more sensitive match results are likely to occur for research questions about small populations or for subpopulations where there are particularly concentrated data quality concerns. For example, most match logic around names is designed with Anglo names in mind. For a match particularly focused on a Chinese population—where names are generally more common and Soundex codes are not a good representation of phonetics—a custom match that accounts for cultural differences would be ideal. A custom match might also be more effective for a Hispanic population, where individuals may routinely have multiple last names. Few match methods are designed to fully capture this richness. Alternately, a research question focused on outcomes for a population of a few hundred children enrolled in an experiment may warrant high levels of clerical review, because even a handful of mismatches could impact results in a small population, and the resources needed for manual review are not that great.

Introducing sensitivity testing as part of the core process of using integrated data for analyses also emphasizes the existence of variation and error in match processing to the research and policy communities that consume these data. By making the existence of uncertainty more transparent, we can educate these audiences and demonstrate the kinds of questions that are and are not susceptible to assumptions in record linkage methodology.

We encourage the development of an active community of partnership and improved communication among researchers in record linkage theory (across several mathematical and computational domains), the analysts who apply those models to practical administrative data challenges, and the researchers and policy leaders who hope to interpret the integrated data. It is important that all three parties come to the table.

In an era where data science and evidence-based policy are ever more prominent topics of media attention, funding, and public/private/university partnerships and initiatives, the popular discourse is oddly silent on record linkage. The integration of datasets to inform policy decisions represents a key area where statistics and computer science significantly impact our ability to build evidence. It is also an area in need of significant applied research. But, to the extent that these questions are noticed, they are currently treated as isolated or unusual problems in data cleaning or data management and addressed with custom, one-off algorithms. While these solutions meet current needs, they often duplicate efforts. They do not look beyond the current situation to future opportunities to scale either within or across jurisdictions.

At present, the record linkage research community is focused on linking census/population files, health records, and public data sources (such as academic citations). There are important differences in the challenges in linking smaller administrative datasets; there is a need for research that speaks specifically to this use case. Growing and expanding this research area is particularly difficult because the data needed to develop and test these methodologies are extremely sensitive and not publicly available.

Practitioners and the policy community often expect a silver bullet. They need to better understand the limitations of any algorithmic solution. There is a real need for translational materials that explain, in plain language, the limitations of different technical solutions, the kinds of data and questions for which they may not work, and how to handle those circumstances.

All parties need to discuss and keep front of mind the significant ramifications for policy and the public if a record linkage solution perpetuates biases. The increasing number of conferences, committees, and papers around data science ethics and algorithmic bias provide a model for this conversation. Until the research and policy communities are more informed and transparent about data integration assumptions and their impact on results (including potential for bias), it will be too easy for analysts to extend basic record linkage techniques beyond the purposes for which they have been designed and honed.

Finally, it is important that the full community of stakeholders define best practices with regard to transparency of match methodology. Commercial vendors occupy a significant portion of the record linkage space and are often inclined to treat their methodologies in proprietary ways. However, this approach is problematic for a process that directly feeds policy research and decision making.

REFERENCES

Brenner, H., Schmidtmann, I., & Stegmaier, C. (1997). Effects of record linkage errors on registry-based follow-up studies. *Statistics in Medicine, 16*(23), 2633–2643. doi:10.1002/(SICI)1097-0258(19971215)16:23<2633::AID-SIM702>3.0.CO;2-1

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

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16. Retrieved from https://www.cs.purdue.edu/homes/ake/pub/TKDE-0240-0605-1.pdf

Farhang, S., and Yaver, M. (2016). Divided government and the fragmentation of American law. *American Journal of Political Science, 60*(2), 401–17. doi:10.1111/ajps.12188.

Getoor, L., & Machanavajjhala, A. (2013). *Entity resolution for big data*. Retrieved from http://users.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf

Goeken, R., Huynh, L., Lenius, T., & Vick, R. (2011). New methods of census record linking. *Historical Methods, 44*(1), 7–14. doi:10.1080/01615440.2010.517152

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., & Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology, 14*(36): 1-10. doi:10.1186/1471-2288-14-36

Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinski, J. M., & Mallick, R. (2005). The effect of record linkage errors on risk estimates in cohort mortality studies. *Statistics, 31*(1), 13–21.

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

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