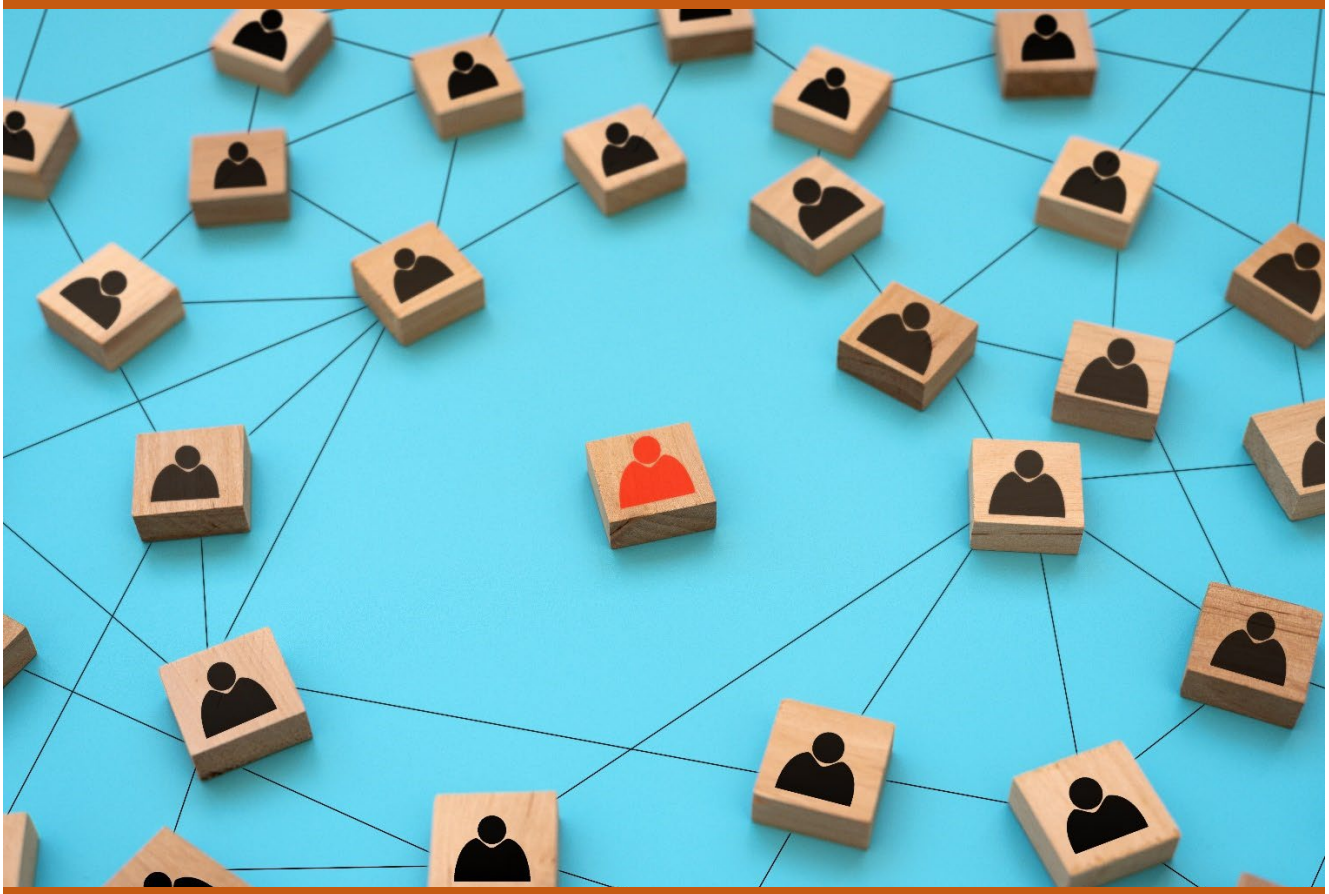
ROADMAP FOR IMPROVING THE FIDELITY OF ADMINISTRATIVE DATA LINKAGE IN SUPPORT OF EVIDENCE-BASED POLICY AND PRACTICE ON THE FOREIGN-BORN SCIENCE AND ENGINEERING WORKFORCE



Veronika Gudipati, Priyanka Bejugam, Heather Saco, and Robert McGough
Arkansas Department of Transformation and Shared Services,
Office of the Chief Data Officer (ARData)

Prepared under the Foreign-Born Scientists and Engineers and the U.S. Workforce (FBSE) project of the America's DataHub Consortium

ARDATA ardata.arkansas.gov

COLERIDGE INITIATIVE

NSF National Science Foundation

Executive Summary

This paper summarizes research conducted under the America's DataHub Consortium to assess record linkage approaches for the foreign-born population and to recommend a roadmap for improving the fidelity of administrative data linkage on the foreign-born science and engineering workforce.

America's Data Hub facilitates coordinated research to drive both infrastructure recommendations and response to relevant evaluation and research needs for current policy priorities through a series of research efforts for the development of a national secure data service. The Foreign-Born Scientists and Engineers and the U.S. Workforce (FBSE) project is focused on improving the evidence base for measuring the economic impact of foreign-born scientists and engineers. This is a particularly relevant topic for record linkage research, because current administrative record linkage approaches are highly dependent upon Social Security Number (SSN), which is often missing for the foreign-born population. This can lead to unique record linkage challenges that can introduce coverage gaps and biases in data and evidence on the foreign-born population and the science and engineering workforce.

As part of a multi-state collaborative effort coordinated by the Coleridge Initiative, researchers from the Office of the Chief Data Officer for the State of Arkansas (ARData):

- Surveyed the predominant approaches, challenges, and recommended best practices for linking administrative data through a comprehensive literature review
- Profiled the quality and availability of identifying attributes in statewide administrative data and assessed the impact on record linkage performance for the population of interest
- Assessed and compared the performance of deterministic, probabilistic, and machine learning-based record linkage approaches on synthetic truth set data and statewide administrative data
- Surveyed and assessed approaches and recommendations for assessing, mitigating, and communicating record linkage bias through literature reviews and testing with statewide administrative data

Assessment of record linkage performance on statewide administrative data (as an analogue for national administrative data) found that 34% of postsecondary completion records for the population of interest could not be successfully linked to employment and earnings records without overmatching due to insufficient individual identifiers.

Mitigation of record linkage bias through improved record linkage transparency led to a 47% change in post-completion employment statistics for the population of interest, suggesting a material impact to data and evidence on the foreign-born population, the programs from which they graduate, and the science and engineering and STEM workforce.

Based on findings and consultation with a panel of experts, the research team developed a roadmap of actionable recommendations for improving record linkage fidelity for the foreign-born population by:

- Raising awareness of record linkage bias
- Improving mitigation and communication of record linkage bias through education
- Increasing transparency of record linkage approaches, performance, and data quality
- Enhancing the collection and sharing of individual identifiers to improve record linkage performance

Introduction

This paper summarizes research conducted under the America's DataHub consortium to assess record linkage approaches, performance, and challenges for administrative data on foreign born populations.

America's DataHub is an initiative of the National Science Foundation (NSF) and the National Center for Science and Engineering Statistics (NCSES) to contribute to evidence building in critical areas such as data access, the use of administrative and other data sources, data linking, data security and privacy, and analysis and dissemination. To start solving the problem of how to access and link myriad sets of statistical data, NCSES has identified some complex questions that simultaneously address relevant evidence building needs while informing the development of a cutting-edge national data infrastructure.

The first task is to analyze the availability of and demand for scientists and engineers on a global scale. That includes building evidence to fully understand the public value of recruiting scientists and engineers from other countries and training them in U.S. universities and labs. Record linkage for foreign born populations poses some unique challenges, and this research, conducted as part of a multi-state collaborative effort coordinated by the Coleridge Initiative, seeks to propose an actionable roadmap for assessing and improving record linkage performance and bias for the foreign-born population, which also has much broader potential benefits to other populations and to administrative record linkage in general.

Population of Interest: United States foreign-born labor force, particularly in science, engineering, and other STEM occupations

For Arkansas higher education administrative records, inclusion in the population of interest (foreign-born) was determined by a Non-US Resident value of "Yes" or a County or Origin other than "USA".

Importance of Foreign-Born STEM Workers to the United States Economy

"Foreign-born STEM workers have made important contributions to the U.S. economy in terms of productivity and innovation. Research has found that immigrants are more likely than the U.S.-born to obtain a patent, and immigrants account for rising shares of U.S. patents in computing, electronics, medical devices, and pharmaceuticals. Immigrants are also more likely to start their own businesses, many of which go on to be major companies.

In 2021, 44 percent of Fortune 500 companies in the United States were founded by an immigrant or the child of an immigrant.

As the demand for STEM workers continues to increase, foreign-born STEM workers will likely continue to complement U.S. workers and play a key role in U.S. productivity and innovation. The number of STEM jobs is projected to increase by 10.5 percent (to almost 11.3 million jobs) between 2020 and 2030. This growth rate is greater than the 7.7 percent growth projected for all occupations during the same period."
(American Immigration Council, 2022)

Problem

The key problem this research seeks to explore is that:

- Key methods for generating data and evidence on education-to-workforce pipelines rely upon the ability to successfully link data on higher education completion, employment, and earnings.
- The predominant methods for linking administrative education and workforce data have a high dependency upon every record having a valid Social Security Number (SSN).
- A large percentage of individuals in the population of interest do not have valid Social Security Numbers in higher education completion records, which precludes successful record linkage.

Hypothesis: A large percentage of the foreign-born population are likely missing from current data and evidence used to inform policy and practice due to unsuccessful record linkage.

Potential Implications: Immigrants made up almost one-fourth of all STEM workers in the United States in 2019, so exclusion of this population can represent a material lack of coverage. (American Immigration Council, 2022)

Broader Implications: Many additional populations often lack coverage of unique interoperable identifiers such as SSN in administrative data, including K-12 students, justice-involved individuals, private and noncredit training program students, and social benefit program participants.

23% of all STEM workers in the United States in 2019 were immigrants.
(American Immigration Council, 2022)

Administrative Data

Information collected and managed by organizations through the administration of programs and regular operations. Data include records of citizens' interactions with government entities, which are crucial for enabling informed decision-making and shaping effective policies and practices.

Higher Education

Data collected and utilized by institutions of higher education serve multiple purposes such as compliance with federal requirements, informed decision-making, and improving student outcomes. Data systems collect a wide range of information about students including demographics, enrollment, program of study, credential attainment, and more.

Workforce

Data systems collect information about employment and earnings, including unemployment claims, wage records, and employment statistics. Information is also collected on education and training programs funded by the federal Workforce Innovation and Opportunity Act (WIOA). Workforce data is collected to support labor market analysis, facilitate workforce planning, evaluate employment programs, and provide valuable insights for policy and decision-making.

Linking Data

Linked data between workforce and higher education systems plays a crucial role in understanding the relationship between education, training, and employment outcomes. By integrating and analyzing data from both systems, policymakers, researchers, and practitioners gain valuable insights into the effectiveness of educational programs and their impact on workforce outcomes.

Individual Identifier Assessment

Record linkage is dependent upon the availability of a common set of individual identifiers, so the availability, completeness, and validity of identifiers was profiled for higher education and wage data.

Higher Education Data and Individual Identifiers

The predominant source of administrative data for higher education completion are the records collected for completion of the Integrated Postsecondary Education Data System (IPEDS) surveys conducted annually by the National Center for Education Statistics (NCES), a part of the Institute for Education Sciences (IES) within the United States Department of Education. IPEDS survey completion is required for all institutions that participate in any federal financial assistance program authorized by Title IV of the Higher Education Act of 1965. These data are typically available for all public postsecondary institutions and some private institutions.

The primary individual identifiers collected for IPEDS reporting include First Name, Middle Name, Last Name, Date of Birth, and a Unique Identification Code within the institution. Social Security Number (SSN) is commonly used for Unique Identification Code values when available, but institutions may not require students to provide an SSN, and some states prohibit the collection of SSN. Profiling of Arkansas Higher Education data found that Middle Name is only present on 48.2% of records for the population of interest.

Workforce Data and Individual Identifiers

The predominant source of administrative data for employment and earnings are the records collected for management of Unemployment Insurance (UI) programs within federal guidelines. Most states mandate the reporting of basic information on employer UI wage records on a quarterly basis. These data are frequently used for administrative data products by states, the Department of Labor, the Census Bureau, and other statistical agencies.

The primary individual identifiers collected for UI Wage reporting include First Name, Middle Name, Last Name, and Social Security Number (SSN). Date of Birth is not commonly required for UI Wage reporting. Profiling of Arkansas UI Wage data found that Middle Name is only present on 55% of records.

Higher Education Social Security Number Completeness

One of the only common identifiers (and the only unique identifier) common to both Higher Education and UI Wage data is Social Security Number. While SSN is frequently used for the Unique Identification Code value in Higher Education Records, it is not required to be collected. Foreign-born students often have alternate identifiers assigned because they do not have an SSN assigned during their postsecondary education unless they are receiving wages through student work programs.

To assess SSN completeness in Higher Education records, the Social Security Administration's validation criteria were applied to the Unique Identification Code values in Arkansas administrative data. While 98.62% of total Arkansas Higher Education records were found to have Unique Identification Code values with a valid SSN format, only 66% of foreign-born postsecondary graduates were found to have a valid SSN format. Valid SSN format does not guarantee that the value present is actually an SSN, but invalid SSN format does indicate that the value present is not an SSN and not a candidate for SSN-based record linkage.

Individual Identifier Assessment Findings

Assessment of individual identifiers for Higher Education and UI Wage administrative data found that:

- SSN is available on all UI Wage records but only valid on 66% of Higher Education records.
- First Name and Last Name are available across both sources with a high level of completeness.
- Middle Name is only complete across 55% of UI Wage and 48.2% of Higher Education records.
- Date of Birth is available on Higher Education records but not UI Wage records.
- Additional demographic identifiers (gender, race/ethnicity) are available on Higher Education records, but not on UI Wage records.

The only individual identifiers present across both sources are SSN, First Name, Last Name, and Middle Name. These are the only candidate attributes currently available for use in record linkage.

Record Linkage Approaches

There are multiple possible approaches for record linkage, each with different advantages depending on characteristics of the source data and the intended use of the linked data.

To identify the most relevant record linkage approaches for the population of interest:

- The predominant record linkage approaches and their respective benefits and applicability were surveyed through a comprehensive literature review.
- The performance (accuracy) of representative algorithms for each type of record linkage was assessed through testing with synthetic truth data sets.
- Record linkage approaches supported by the available identifying attributes were assessed for performance (accuracy) against a curated truth set constructed using actual administrative data.

Record Linkage Accuracy

Record linkage is the process of comparing pairs of records (pairwise comparison) to evaluate equivalency and determine if they refer to (or the probability that they refer to) the same real-world entity. Accuracy is a measurement of how closely the record linkage process can result in linkages that match the real world, which for individual record linkage means that both records refer to the same real-world person and were successfully identified as equivalent. Assessing record linkage accuracy typically requires comparison against a gold standard or truth set for which the ground truth is known. Manual review is also possible but not practical for large volumes of data.

There are multiple measurements of record linkage accuracy representing different combinations of:

- True Positives – Pairs evaluated as equivalent that represent the same entity.
- True Negatives – Pairs evaluated as not equivalent that do not represent the same entity
- False Positives – Pairs evaluated as equivalent that do not represent the same entity.
    - This is also known as **overmatching** because the process is matching too many records, including those that are not equivalent in the real world.
- False Negatives – Pairs evaluated as not equivalent that do represent the same entity.
    - This is also known as **undermatching** because the process is failing to match records that are equivalent in the real world.

Different measures exist because different uses of data may dictate a preference for overmatching, undermatching, or a more balanced assessment of accuracy. For example, healthcare use cases often have a higher tolerance for false positives but very low tolerance for false negatives because missing an early diagnosis represents a greater risk and impact to the patient than identifying a false potential risk that can be ruled out through additional testing. For the purposes of this assessment, a balanced measurement of accuracy was used in which the score is improved for true positives and true negatives and penalized for false positive and false negatives.

A key concern of note with overmatching for evaluation and research purposes is that it not only represents an inaccurate linkage. Overmatching can also result in more linkages than the total number of candidate records to be matched. When these additional linkages are used for analysis of linked data sets, the size of the entire result set can be increased, sometimes substantially, which can have a material impact on the accuracy of resulting data and evidence.

Record linkage performance is affected by not only the type(s) of record linkage used and their respective parameter, but largely by the quality of the identifying data available to the process.

Record Linkage Computational Performance

Record linkage approaches can vary in the computational expense required to perform the pairwise comparison, which is particularly important when considering approaches for deployment at national scale. Since pairwise comparison compares every record with every other record, the number of potential comparisons grows quadratically (not linearly) with an increase in record volume. Some methods use an approach called "blocking" to reduce the number of comparisons by splitting out the record set into groups more likely to be equivalent due to similarity tests that can be performed at low computational expense.

Key factors in computational expense can include:

- The number of records to be compared
- The number of attributes to be compared
- The data types of the attributes
    - Comparisons of numeric data is less computationally expensive than string comparison
    - "Fuzzy" comparison of string similarity can be very computationally expensive
- Whether the comparison is performed synchronously (all records compared at once) or asynchronously (only new or changed records compared).
    - Asynchronous record linkage can be less computationally expensive but requires that linkages be persisted versus "match and destroy" approaches that start over each time

Record Linkage Transparency

The transparency of record linkage approaches refers to how much access data users have to:

- The type(s) of record linkage used
- The strength or confidence of a linkage
- Characteristics of the source data included in the match

Record linkage transparency is important because source characteristics and record linkage decision can have a material impact on data and evidence produced from the linked dataset.

Deterministic Record Linkage

Deterministic record linkage is an approach in which a set of rules comparing one or more attributes makes an all-or-nothing determination on whether a pair of records are found to be equivalent.

The simplest implementation of deterministic record linkage is an exact match on a single unique identifier, such as a Social Security Number. This is the type of record linkage most common on the population of interest and in much administrative data record linkage due to insufficient availability of alternative identifiers.

The incumbent record linkage solution for the population of interest, deterministic linkage on Social Security Number, cannot yield higher than 66% successful record linkage on the representative administrative data due to the lack of valid SSNs for 34% of the population of interest.

Deterministic record linkage can also be implemented using matches across multiple identifying attributes such as name, date of birth, demographic identifiers, or contact mechanisms (phone, email, address) in the absence of globally unique identifiers such as SSN.

The only common identifier currently available between Higher Education and UI Wage administrative data is Name. To assess the feasibility of deterministic record linkage on name, the uniqueness of name values was assessed across Arkansas administrative data. This was done by measuring the number of distinct SSN values for each unique Name.

15% of names in Arkansas UI Wage data were found to belong to two or more individuals (with a maximum of almost 600). This percentage is likely to be higher in larger states and significantly higher on a national scale.

Deterministic record linkage on Name alone was not found to be a viable record linkage approach.

The remaining identifier available with high completeness in Higher Education administrative data but not currently available in Arkansas UI Wage data is Date of Birth. To assess the feasibility of deterministic record linkage on the combination of Name and Date of Birth, the number of distinct SSN values was measured for each unique combination of Name and Date of Birth in Arkansas Higher Education administrative data.

Only 484 distinct combinations of Name and Date of Birth (out of over 515K) were found to belong to two or more individuals (with a maximum of 3). This represents only 0.09% of the total population.

Deterministic record linkage on Name and Date of Birth was found to be a viable record linkage approach in absence of SSN if Date of Birth were available on UI Wage records.

Because deterministic rules can also include aggregations, it is possible to implement a rule establishing equivalency if Name and Date of Birth match and the count of distinct SSNs on the UI Wage side of the match is not greater than one. This could mitigate the chance of overmatch for the 0.09% of the population (or greater in larger geographies) who share the same Name and Date of Birth. This step would be more computationally expensive but only needs to be performed against records that could not be matched through more efficient rules.

Probabilistic Record Linkage

Probabilistic record linkage primarily differs from deterministic record linkage in that it assesses the similarity of pairs or the probability of equivalence on a continuous scale (0% to 100%) versus a hard determination (Y or N). This increased granularity can be more finely tuned than is possible with the finite and coarse rules available through deterministic approaches.

Another key difference is that probabilistic approaches can take into account the specificity of each identifying attribute value and give them more or less weight in assigning probability of equivalency accordingly. For example, a very rare or uniquely spelled name can contribute more weight to a match probability than a very common name.

The continuous match threshold can also be used to specify multiple states such as probable positive, probable negative, and a range of uncertainty that is flagged for review.

Probabilistic record linkage was assessed for the population of interest but was not found to result in significant improvement in accuracy to warrant the additional implementation complexity given available identifying attributes.

Machine Learning Approaches

Multiple machine learning-based record linkage approaches were also assessed, including neural networks and transfer learning. While these approaches demonstrated impressive accuracy and computational performance on synthetic truth data sets, there was insufficient data available for the population of interest for these approaches to be immediately applicable or beneficial.

Comparison of Record Linkage Approaches

The key differences between record linkage approaches can be summarized as:

- Probabilistic linkage can yield improved accuracy for data with poor data quality due to the ability to better facilitate similarity metrics. (Zhu et al., 2015)
- Deterministic linkage can yield comparable accuracy with lower computational expense on high quality data compared to probabilistic methods. (Zhu et al., 2015)
- Both deterministic and probabilistic linkage perform poorly if rules and data offer low discriminative power (fewer rules, lower cardinality data, low data quality). (Zhu et al., 2015)
- When working with large record sets, deterministic linkage shows greater advantage over probabilistic linkage in terms of computational efficiency and simplicity of implementation. (Zhu et al., 2015)

The choice between the two methods depends on the data quality, available identifiers, and the desired trade-off between precision and recall. In some cases, a combination of both approaches can be used, leveraging deterministic linkage as a pre-processing step to reduce the search space for probabilistic linkage. (Harron et al., 2017)

Facilitation of both deterministic and probabilistic record linkage approaches is recommended for achieving a balance of record linkage fidelity and computational efficiency while affording data analysts more versatile linkage options based on data use and tolerance for false negatives, false positives, and overall predictive performance.

Assessing and Mitigating Record Linkage Bias

Due to the identification of unsuccessful record linkage for the population of interest, a literature review was conducted on approaches and best practices for assessing, mitigating, and communicating record linkage bias.

Key findings from the literature review include:

- There should be awareness and education efforts to train users of linked administrative data on the existence, impact, measurement, and mitigation of record linkage error and bias as well as how to communicate record linkage methods, performance, and bias. (Wiegand et al., 2019)
- Data analysts should assess and report on the quality of linked data used for analysis, including how analyses took linkage error and bias into account. (Harron et al., 2020) (Wiegand et al.,2019)
- Measuring and mitigating the presence and impact of record linkage error and bias requires infrastructure design considerations to allow for more transparency into record linkage processing and performance. (Ruth et al., 2018)
- Data providers should make details available on the population included in the data set, the coverage, and the data generation or collection mechanism. (Harron et al., 2020)

Recommended approaches for assessing and mitigating record linkage bias were tested with statewide administrative data to determine feasibility, implementation requirements, and impact on results.

One of the most common metrics included in federal reporting and consumer information products leveraging linked Higher Education and UI Wage administrative data is the percentage of graduates who are employed one year (or other intervals) post completion. This metric is calculated as the number of Higher Education completers found in UI Wage records at the interval of interest divided by the total number of Higher Education completers in the period being assessed. Failed record linkage due to insufficient identifiers essentially removes completers from the numerator, artificially lowering post completion employment due to record linkage bias.

Testing was performed to mitigate this bias in analyses by:

- Adding an SSN validity indicator to the source data prior to deidentification.
    - This is important because validity rules cannot be applied to hashed identifiers.
- Making the full population of source records available to analysts with transparency into which records were successfully or unsuccessfully linked.
- Incorporating additional data quality and linkage metadata into the analysis in order to remove records that could not be linked due to invalid SSNs from the denominator since they were already being removed from the numerator.
    - Removal of these records essentially treats this as a sample statistic versus a population statistic, and the increased transparency allows for communication of the confidence.

Mitigation of record linkage bias through improved record linkage transparency led to a 47% change in post-completion employment statistics for the population of interest, suggesting a material impact to data and evidence on the foreign-born population, the programs from which they graduate, and the science and engineering and STEM workforce of which they constitute a significant percentage.

Recommendations

Recommendations for improving the fidelity of administrative data linkage in support of evidence-based policy and practice include:

Awareness, Measurement and Mitigation

- There should be awareness and education efforts to train users of linked administrative data on the existence, impact, measurement, and mitigation of record linkage error and bias as well as how to communicate record linkage methods, performance, and bias.
- Data analysts should assess and report on the quality of linked data used for analysis, including how analyses took linkage error and bias into account.

Record Linkage Approaches

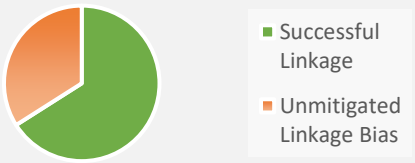
- Measuring and mitigating the presence and impact of record linkage error and bias requires infrastructure design considerations to allow for more transparency into record linkage processing and performance.
- Facilitation of both deterministic and probabilistic record linkage approaches is recommended for achieving a balance of record linkage fidelity and computational efficiency while affording data analysts more versatile linkage options based on data use and tolerance for false negatives, false positives, and overall predictive performance.

Data Collection and Preparation

- Data providers should make details available on the population included in the data set, the coverage, and the data generation or collection mechanism.
- A key limiting factor to record linkage fidelity is the lack of identifying attributes on some key administrative data sources. The lack of identifying attributes beyond Social Security Number and Name on UI Wage data is particularly limiting due to the broad use and relevance of administrative data on employment and earnings.
    - Efforts to enhance the collection of individual labor market information data should consider not only information gain from additional observational attributes (occupation, hours worked) but also enhanced collection of individual identifiers to reduce information loss from record linkage error.
    - Government and employer participation in the Jobs and Employment Data Exchange (JEDx) initiative has the potential to not only provide more timely, detailed, and relevant administrative data, but also improved record linkage fidelity through the inclusion of numerous (14) individual identifiers and descriptors in the JEDx schema.

Roadmap

Current State

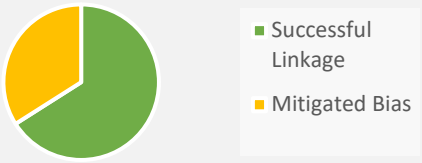


- **Poor Linkage**: Record linkage is unsuccessful for 34% of the population of interest.
- **Poor Transparency**: Transparency on data and linkage quality is extremely limited.
- **Poor Mitigation**: Awareness, mitigation, and communication of record linkage bias for the population of interest is extremely limited.

Recommended Gap Plan: Awareness, Mitigation, Communication, and Transparency

- Raise **awareness** of record linkage bias through presentations, briefings, and other communications across the Multi-State Data Collaboratives, State Chief Data Officers Network, Workforce Data Quality Initiative, and other groups of administrative data users.
- Develop and deliver curricula on **assessment**, **mitigation**, and **communication** of record linkage bias through Applied Data Analytics training programs, educational materials, and other supports for effective use of administrative data for evidence-based policy.
- Increase **transparency** of record linkage approaches, record linkage performance, and source data quality in the Administrative Data Research Facility and other administrative data linkage environments through increased communication, documentation, metadata (such as SSN validity), and access to all linked and unlinked source records.

Mitigation Stage

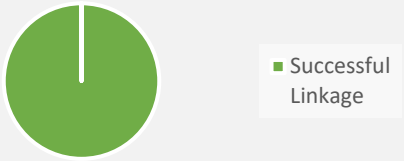


- **Poor Linkage**: Record linkage is unsuccessful for 34% of the population of interest.
- **Improved Transparency:** Transparency on data and linkage quality is included in record linkage approaches and communication across roles.
- **Improved Mitigation:** Awareness, mitigation, and communication of record linkage bias for the population of interest is common.

Recommended Gap Plan: Improved Collection and Sharing of Individual Identifiers

- Enhance the **collection** and **sharing** of **individual identifiers** (particularly Date of Birth) in administrative data on employment and earnings to facilitate improved record linkage.
- Contribute to discussions by the Workforce Information Advisory Council on enhanced collection and sharing of wage data to communicate the importance of individual identifiers.
- Assess and communicate the potential record linkage improvement that might be realized through participation in the Jobs and Employment Data Exchange (JEDx) initiative.
- Incorporate additional identifiers into record linkage approaches for improved performance.

Improvement Stage



- **Improved Linkage:** Record linkage is successful for almost the entire population of interest with immaterial bias to evidence.
- **Improved Transparency:** Transparency on data and linkage quality is included in record linkage approaches and communication across roles.
- **Improved Mitigation:** Awareness, mitigation, and communication of record linkage bias for the population of interest is common.

Expert Panel Participants

We acknowledge with gratitude the following members of the expert review panel:

- Dr. John Talburt – University of Arkansas at Little Rock Center for Advanced Research in Entity Resolution and Information Quality
- Dr. Nathan Barrett– Coleridge Initiative
- Dr. Annelies Goger – Brookings Institution
- Jason Tyszko – United States Chamber of Commerce Foundation

References

Ali, M. Sanni, Maria Yury Ichihara, Luciane Cruz Lopes, George CG Barbosa, Robespierre Pita, Roberto Perez Carreiro, Djanilson Barbosa Dos Santos et al. "Administrative data linkage in Brazil: potentials for health technology assessment." Frontiers in pharmacology 10 (2019): 984.

American Immigration Council. (2022). Foreign-Born STEM Workers in the United States. https://www.americanimmigrationcouncil.org/sites/default/files/research/foreign-born_stem_workers_in_the_united_states_final_0.pdf

Barlaug, Nils, and Jon Atle Gulla. "Neural networks for entity matching: A survey." ACM Transactions on Knowledge Discovery from Data (TKDD) 15, no. 3 (2021): 1-37.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. Big Data & Society, 4(2).

Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Ann Hum Biol. 2020 Mar;47(2):218-226. doi: 10.1080/03014460.2020.1742379. PMID: 32429765; PMCID: PMC7261400.

Kirielle, Nishadi, Peter Christen, and Thilina Ranbaduge. "TransER: Homogeneous Transfer Learning for Entity Resolution." In *EDBT*, pp. 2-118. 2022.

Ruth Gilbert, Rosemary Lafferty, Gareth Hagger-Johnson, Katie Harron, Li-Chun Zhang, Peter Smith, Chris Dibben, Harvey Goldstein, GUILD: GUidance for Information about Linking Data sets, *Journal of Public Health*, Volume 40, Issue 1, March 2018, Pages 191–198, https://doi.org/10.1093/pubmed/fdx037

Shlomo, Natalie. "Overview of data linkage methods for policy design and evaluation." Data-driven policy impact evaluation: how access to microdata is transforming policy design (2019): 47-65.

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.

Winkler, William E. Record linkage software and methods for merging administrative lists. US Bureau of the Census, 2001.

Zhu, Ying, Yutaka Matsuyama, Yasuo Ohashi, and Soko Setoguchi. "When to conduct probabilistic linkage vs. deterministic linkage? A simulation study." Journal of Biomedical Informatics 56 (2015): 80-86.