

UCLA

UCLA Previously Published Works

Title

How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data.

Permalink

<https://escholarship.org/uc/item/8737510q>

Journal

Journal of Economic Literature, 58(4)

ISSN

0022-0515

Authors

Bailey, Martha
Cole, Connor
Henderson, Morgan
[et al.](#)

Publication Date

2020-12-01

DOI

10.1257/jel.20191526

Peer reviewed



Published in final edited form as:

J Econ Lit. 2020 December ; 58(4): 997–1044. doi:10.1257/jel.20191526.

How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data

Martha Bailey^{1,2}, Connor Cole¹, Morgan Henderson¹, Catherine Massey¹

¹University of Michigan

²National Bureau of Economic Research

Abstract

This paper reviews the literature in historical record linkage in the U.S. and examines the performance of widely-used record linking algorithms and common variations in their assumptions. We use two high-quality, hand-linked datasets and one synthetic ground truth to examine the direct effects of linking algorithms on data quality. We find that (1) no algorithm (including hand-linking) consistently produces representative samples; (2) 15 to 37 percent of links chosen by widely-used algorithms are classified as errors by trained human reviewers; and (3) false links are systematically related to baseline sample characteristics, showing that some algorithms may induce systematic measurement error into analyses. A case study shows that the combined effects of (1)-(3) attenuate estimates of the intergenerational income elasticity by up to 20 percent, and common variations in algorithm assumptions result in greater attenuation. As current practice moves to automate linking and increase link rates, these results highlight the important potential consequences of linking errors on inferences with linked data. We conclude with constructive suggestions for reducing linking errors and directions for future research.

New large-scale linked data are revolutionizing empirical social science. Record linkage is increasingly popular as a tool to create or enhance data for observational studies, randomized control trials, and lab and field experiments. Examples abound across subfields in economics, including health economics and medicine, industrial organization, development economics, criminal justice, political economy, macroeconomics, and economic history. In addition, current U.S. data infrastructure projects are linking national surveys, administrative data, and research samples to recently digitized historical records, such as the full-count 1880 (Ruggles et al. 2015, Ruggles 2006) and 1940 U.S. Censuses (the first Census to ask about education and wage income).¹ These newly available “big data” have the potential to break new ground on old questions and open entirely novel areas of inquiry.

¹Many on-going initiatives link the 1940 U.S. Census to other datasets. The Census Bureau plans to link the 1940 Census to current administrative and Census data (Census Longitudinal Infrastructure Project, CLIP) and the Minnesota Population Center plans to link it to other historical censuses. The *Panel Survey of Income Dynamics* and the *Health and Retirement Survey* are linking their respondents to the 1940 Census. The Longitudinal, Intergenerational Family Electronic Micro-Database Project (LIFE-M) is linking vital records to the 1940 Census (Bailey 2018). Supplementing these public infrastructure projects, entrepreneurial researchers have also combined large datasets. See, for example, Abramitzky, Boustan, and Eriksson (2012a), Abramitzky, Boustan, and Eriksson (2013), Abramitzky, Boustan, and Eriksson (2014), Boustan, Kahn, and Rhode (2012), Mill (2013), Mill and Stein (2016), Hornbeck and Naidu (2014), Aizer et al. (2016), Bleakley and Ferrie (2013, 2017, 2016), Nix and Qian (2015), Collins and Wanamaker (2014, 2015, 2016), and Eli, Salisbury, and Shertzer (2018). This paper discusses many of the linking approaches used in these papers.

Machine-linking methods are critical to these projects, especially those linking U.S. Censuses. But outside of protected data enclaves, little is known about how machine algorithms influence data quality and inference, due both to false matches (Type I errors) and missed matches (Type II errors).² This gap in knowledge reflects the lack of “ground truth” data. Although some diagnostic exercises are suggestive, they typically rely on selected samples (genealogy), non-U.S. samples (Goeken et al. 2017, Christen and Goiser 2007, Eriksson 2016), or rich administrative data unavailable to most researchers (Scheuren and Winkler 1993, Winkler 2006, Massey 2017, Abowd 2017). Uncertainty about the quality of machine-linked data limits their value to social science and also the development of methods to improve them.

This paper reviews the literature in historical record linkage in the U.S. and evaluates the effects of different linking algorithms on data quality. Unlike contemporary data, historical data are public and contain identifiable information, allowing us to be fully transparent about our samples and assumptions in assessing algorithm performance. Our samples include the Longitudinal Intergenerational Family Electronic Micro-database’s (LIFE-M) sample of birth certificates linked to the 1940 Census (Bailey 2018) as well as a sample of Union Army veterans which the Early Indicators Project linked to the 1900 Census (Costa et al. 2017).

Even well-trained human linkers and genealogists make errors, so we also build a synthetic ground truth to validate our findings. The synthetic ground truth deliberately introduces common errors in recording, transcription and digitization of historical data. Although this synthetic ground truth is an imperfect representation of the more complicated errors in original records, the dataset’s construction means that there is complete certainty about the correct links. In all cases, the synthetic data produce very similar findings to the hand-linked records.

The results highlight how widely-used linking algorithms affect data quality and illustrate how different assumptions impact performance. First, we find that no linking method produces samples that are consistently representative of the linkable population, and the ways in which the data are not representative differ by algorithm. Second, widely used automated-linking algorithms produce large numbers of links that well-trained humans classify as incorrect, with rates ranging from 15 to 37 percent. Similar results in synthetic ground truth suggest that links rejected in human review are likely Type I errors. Third, false links produced by different algorithms tend to be strongly associated with baseline sample characteristics, suggesting that linking algorithms could induce systematic measurement error into analyses. In addition, the systematic measurement error varies across algorithms and records, suggesting that any bias may be difficult to predict and correct.

Our analysis also investigates how algorithm assumptions impact data quality, including phonetic name cleaning, linking more common names, and using methods to resolve ties. We find that common uses of spelling standardization in deterministic algorithms tend to increase both Type I errors, from 16 to 60 percent, as well as Type II errors. Linking more

²“Ground truth” is defined as data obtained by direct observation of the true link.

common names dramatically increases Type I errors although Type II errors fall. Lastly, including records with exact ties on name, age, and birth place (often used in conjunction with simple probability weights) increases error rates by an additional 55 to 79 percent.

After characterizing the theoretical implications of linking errors using a within-between decomposition framework, we link the same fathers and sons to the 1940 Census using different algorithms and examine the resulting estimates of intergenerational mobility. We find that some linking algorithms attenuate intergenerational income elasticity estimates by up to 20 percent. Frequently used variations in assumptions, such as including more common names and phonetic name cleaning, result in attenuation of more than 30 percent. Eliminating false matches, however, renders intergenerational income elasticities from different algorithms statistically indistinguishable. In our case study, false links appear to have a larger impact on inferences than sample composition—a finding that cautions against recent efforts to increase match rates at the expense of precision.

We conclude with easy-to-implement recommendations for improving machine linking and inference with linked samples. In particular, we recommend reweighting to address sample non-representativeness and using multiple linking algorithms and supervised learning methods (with training data) to identify and reduce false links, break ties for multiple matches, and better train machine algorithms.

I. The Evolution of U.S. Historical Record Linkage

Record linkage has been a mainstay of social science for over 80 years. The earliest methods used painstaking manual searches to link hand-written manuscripts, and recent developments in digitized records, computational speed, and probabilistic linking techniques have expanded the possibilities for automated, or machine-based, record linkage. We briefly summarize this early literature, focusing on the components of this history that laid the groundwork for current practice.³

One feature important to historical and modern linking is “blocking.” Blocking refers to the partition of a dataset into “blocks” (or clusters of records) using a record attribute (Michelson 2006). This technique limits the number of potential matches according to the blocking attribute, thereby improving computational efficiency while (ideally) maintaining accuracy. For instance, blocking on place of birth and sex means that a linking algorithm looking for Franklin Jones born in Kentucky would *only* search within the set of candidate matches of men born in that state.

Historical linkage has always used blocking techniques to increase the feasibility of manually linking samples across manuscripts. The earliest blocking methods involved identifying a group of individuals in a particular location (e.g., township, county, or state) in one census and manually searching for the same people within the same region (the block) in the subsequent census (Malin 1935, Curti 1959, Bogue 1963, Thernstrom 1964, Guest 1987). While making manual searching feasible, this blocking strategy missed those who

³See Ruggles, Fitch, and Roberts (2018) for a more detailed history of the findings in this early literature.

relocated or changed names between census years. The resulting linked samples omitted the geographically mobile population and were, therefore, unrepresentative (Ruggles 2006).

The creation of digitized state population indexes facilitated refinements in blocking.⁴ In one such approach that improved on previous methods, Steckel (1988) drew a random sample of households with children at least 10 years old in a historical census. He then searched for the same household in the previous census using the birth state of the child to narrow the search. This technique was able to locate individuals who moved between the census years, but it restricted the sample of linked households to those with children surviving to age ten.

Advances in computing allowed improvements in automated matching, effectively replacing time-intensive human search with computer queries. Leveraging newly created national population indexes and Public Use Microdata Samples (PUMS), automated matching began incorporating more data elements in the linking process. An early example of this strategy was Attack, Bateman, and Gregson (1992)'s probabilistic matching software, called "PC Matchmaker." PC Matchmaker transformed names using phonetic codes and allowed for user-specified blocking and weighting schemes. Attack (2004) used this software to create a linked sample between the agricultural and population censuses between 1850 and 1880. Ferrie (1996)'s approach, which we describe in more detail in the following sections, aimed to create large, representative linked U.S. Census samples and has since been embraced by the literature, forming the basis of prominent methods in use today. Before summarizing more modern methods, we present an example linking problem to illustrate common challenges in historical linking.

II. Current Approaches to Linking Historical Data

Matching records across sources requires choosing linking variables, also called "features" in the computer science and statistics literatures. Modern administrative records typically have multiple, high-quality features (e.g., full legal name, Social Security Number, exact date of birth, address of residence). Outside of restricted administrative enclaves, data typically contain a limited set of noisy linking features. Historical data have the advantage of containing identifiable information, allowing transparent study of how limited data and data errors affect the quality of linked samples. Like many modern linking problems, historical data have limited information that is often measured with error.

As an example, consider the challenge of linking birth certificates to the 1940 U.S. Census. Researchers typically use "time-invariant" features to do linking in order to minimize concerns about selection bias and non-representativeness in linked samples (Ruggles, 2006). For U.S. Census linking, these variables typically include first name, last name, age, birth state, race, and sex.⁵ In practice, names may vary over time, either because Census enumerators misspelled names, the individual reported incorrectly, or the individual changed names (perhaps using a middle name or nickname in place of the given name). Goeken et al.

⁴A state "index" is a list of individuals living in a state at a point in time.

⁵Matching in historical settings in other countries often makes greater use of characteristics not available in U.S. data. Modalsli (2017) notes that in Norway before 1910 there is less first name variation and more flexible surname traditions than in the U.S. In addition, Norwegian censuses use 500 birthplaces (municipalities) for a population of under 2 million, whereas the U.S. Censuses identify birthplaces as 48 states and foreign countries for a much larger population (~132 million residents in the 1940).

(2017) document that in two enumerations of St. Louis in the 1880 Census, nearly 46 percent of first names are not exact matches. Similarly, the Early Indicators project notes that 11.5 percent of individuals in the Oldest Old sample have a shorter first name in pension records than in the original Civil War enlistment records (Costa et al. 2017).

Similar problems arise in reported age and birth state. The recording of age in the Census tends to reflect “age heaping,” the common practice of rounding ages to the nearest multiple of five (A’Hearn, Baten, and Crayen 2009, Hacker 2013). In addition, birthplaces are often inaccurately recorded. Goeken et al. (2017) show that 8 percent of reported birthplaces do not match across the two enumerations of St. Louis. In addition, rates of disagreement for mothers’ and fathers’ birthplaces for the same individuals *average* 19 and 18 percent, respectively.

The digitization of hand-written manuscripts compounds errors in recording. Our comparison of two independently digitized versions of the 1940 Census by [Ancestry.com](https://www.ancestry.com) and [FamilySearch.org](https://www.familysearch.org) shows that 25 percent of records have different transcriptions of last name due to digitization alone.

These data quality issues are well known, and linking algorithms account for them by allowing for variation in age and name spellings. To deal with differences in age, researchers typically search over a range of ages. Researchers account for orthographic differences by using metrics such as Jaro-Winkler or Levenshtein to quantify the dissimilarity of two name strings. In some cases, researchers use phonetic string cleaning algorithms to help account for spelling differences, name Anglicization, and transcription errors. Soundex, for example, was developed in the early 20th century to help create Census links, simplifying names into phonetic codes to facilitate record searches. For example, Soundex assigns the same code (S530) to similar sounding names like “Smith,” “Smyth” and “Smythe.” Another cleaning algorithm, NYSIIS, the New York State Identification and Intelligence System, was developed as an improvement to Soundex in 1970. NYSIIS transforms the same root name to a common string, making names like “Wilhem” and “William” into “WALAN.” While phonetically cleaned strings allow researchers to identify more candidate links, matching on them deterministically treats distinct names as the same. One implementation of NYSIIS, for instance, categorizes John and James as perfect matches (Ruggles, Fitch, and Roberts 2018).

Figure 1 illustrates how limited information and measurement error create challenges for matching records. The linking problem is depicted as two-dimensional scatter plot after blocking on birth state and sex, as is common in the literature. The x-axis captures the similarity between the name on record to be linked and the names of candidate links in the 1940 Census using the Jaro-Winkler similarity score, which will equal 1 if the names are identical and is less than 1 otherwise.⁶ The y-axis captures the difference in the age implied by the birth certificate (which contains exact date of birth) and the reported age in the 1940 Census. A perfect match in ages occurs when the age difference is zero.

⁶Jaro-Winkler similarity score adapts the Jaro (1989) string score, the minimum number of single-character transpositions required to change one string into another, to up-weight differences that occur at the beginning of the string. See Winkler (2006) for an overview.

In this two-dimensional space, candidate links fall into one of four categories:

- (M1) A perfect (1,0), unique match in terms of name and age similarity (Figure 1A).
- (M2) A single, similar match that is slightly different in terms of age, name, or both (Figure 1B).
- (M3) Many *perfect* (1,0) matches, leading to problems with multiple matches (Figure 1C).
- (M4) Multiple similar matches that are slightly different in terms of age, name, or both (Figure 1D).

As we discuss, historical linking algorithms generally treat M1 cases as matches. However, methods differ in their treatment of candidates in the M2, M3, and M4 categories. To account for differences in age as in M2, researchers typically search within a band of ± 3 or ± 5 years. Prominent approaches to dealing with ties in categories M3 or M4 include random selection among equally likely (tied) candidates (Nix and Qian 2015), equal probability weighting of tied candidates (Bleakley and Ferrie 2016), or the use of a weighted combination of linking features to classify true matches (Feigenbaum 2016, Abramitzky, Mill, and Pérez 2018). The next sections describe how commonly used linking algorithms work and ultimately classify records in cases such as M2, M3, and M4.

A. Ferrie (1996)

Ferrie's (1996) path-breaking approach links men in the 1850 U.S. Census to men who were 10 years and older in the 1860 U.S. Census. Ferrie (1996) begins by selecting a sample of uncommon names from the 1850 Census.⁷ To correct for minor orthographic differences (category M2 above), his algorithm transforms last names using NYSIIS codes and also truncates the untransformed first name at the fourth letter. The algorithm then links his sample to the 1860 Census and eliminates candidate links that were not born in the same state and not living with the same family. The algorithm keeps all candidate links within a ± 5 year difference in age (or "age band") and, if more than two links remain, chooses the link with the smallest age difference. At the end of this process, the algorithm drops cases where two individuals from 1850 link to the same observation in 1860.⁸ This process produces a linked sample of 4,938 men—9 percent of the male population in 1850, and 19 percent of the population of men with uncommon names. Ferrie has used different approaches in more recent work, including smaller age ranges, different ways of parameterizing name dissimilarities like SPEDIS, or altered restrictions on common names. More than 20 years later, Ferrie's approach has become the foundation for much of the historical linking literature.

Two features of this algorithm are especially worth noting. First, the decision to make links among observations with uncommon names reduces both the computational burden and the number of candidate matches of the M3 variety. Consequently, this method never attempts to link common names like "John Smith." Second, the decision to use NYSIIS and truncate

⁷Ferrie (1996) searched for 25,586 men in the 1860 Census whose surname and first name appeared ten or fewer times in 1850.

⁸Ferrie (1996) does not specify a process for multiple match disambiguation – in his linking from 1850-1860, there were no ties after minimizing the difference in age.

first name reduces problems associated with minor orthographic differences, but it may also increase ties of the M3 variety and, therefore, the number of records the algorithm will not link. The independent effects of both of these choices are considered in our subsequent analysis.

B. Abramitzky, Boustan, and Eriksson (2012 and 2014)

Abramitzky, Boustan, and Eriksson's (2012, 2014) "Iterative Method" scale up Ferrie (1996) to use the full-count census. This procedure relaxes Ferrie's (1996) uncommon name restriction to the extent that the combination of name *and* age provide distinctive information. Summarized in a detailed web appendix, Abramitzky, Boustan, and Eriksson (2012b) select a sample of boys ages 3 to 15 with unique name-age combinations in the 1865 Norwegian Census, standardize first and last names using NYSIIS codes, and look for exact, unique matches in U.S. and Norwegian Censuses. For the observations in the 1865 Census without an exact, unique link (M1), the algorithm then searches for a name match within a ± 1 age band and, if there is no match in this band, the algorithm searches within a ± 2 age band. The algorithm does not link a record if more than one candidate match exists within an age band. The algorithm ultimately links a sample of 2,613 migrants and 17,833 non-migrants from a primary sample of 71,644 individuals for a match rate of 29 percent. Abramitzky, Boustan, and Eriksson (2014) use the same procedure to link men ages 18 to 35 with unique age-name combinations from the 1900 U.S. Census to the 1910 and 1920 Censuses, producing a sample of 20,225 immigrant and 1,650 native-born men for match rates of 12 percent and 16 percent.

The authors provide the most recent version of their code for our analysis, which has been used for record linkage in a number of high profile papers.⁹ A variation on this approach is also reported in Abramitzky, Boustan, and Eriksson (2014)'s appendix as a robustness check. Similar to Ferrie's (1996) uncommon name restriction, this robustness check requires that names be unique within a five-year age band (a ± 2 year difference).

Two differences to Ferrie (1996) are worth noting. First, Abramitzky, Boustan and Erickson (2012, 2014) link more common names, while Ferrie's (1996) algorithm does not. Second, Abramitzky, Boustan and Erickson (2012, 2014) use a narrower age band than Ferrie (1996).

⁹Many other papers have used variations on Ferrie (1996) and Abramitzky et al. (2014). These variations are similar in that they require matches to match completely on a cleaned name variable. For example, Abramitzky et al. (2013) use the Abramitzky et al. (2014) algorithm to match men aged 3 to 15 in the 1865 Norwegian Census to the 1880 U.S. and Norwegian Censuses, and match 26 percent of records unique by name and birth year from the 1865 Census. Boustan et al. (2012) link the IPUMS sample of the 1920 U.S. Census to the 1930 U.S. Census using a link uniqueness age band that functions similar to an uncommon names restriction, and match 24 percent of men unique by name, age and birthplace in their 1920 U.S. sample. This same dataset was used in Hornbeck and Naidu (2014). Collins and Wanmaker (2015) use a variation on the Ferrie (1996) method with an alternate name uniqueness requirement to match southern men younger than 40 in the 1910 Census to the 1930 Census. They match 24 percent of their records from the 1910 Census. Other papers use name similarity measures. Eli, Salisbury, and Shertzer (2018) match Civil War recruitment records from Kentucky to U.S. Censuses before and after the Civil War using a variation of Ferrie (1996) without an uncommon names restriction, and impose an additional restriction on Jaro-Winkler string dissimilarity after generating candidate matches with NYSIIS. They match 30 percent of selected records of recruits from the 1860 U.S. Census to the 1880 U.S. Census. Aizer et al. (2016) match state mother's pension records to other data sources, including the Social Security Death Master File, using a variation on Ferrie (1996) without an uncommon names restriction and Soundex, but allow some additional matches to differ in exact name but have low SPEDIS and Levenstein dissimilarity measures. Aizer et al. (2016) match 48 percent of their records to the Social Security Death Master File, but this high match rate reflects the fact that they can use exact date of birth to match their observations.

C. Feigenbaum (2016) and IPUMS (2015)

A common feature of Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014) is that they search among identical, phonetically cleaned names for a match that *uniquely* has the minimum age difference. This restriction reduces computational burden, but comes at the cost of excluding very similar (though not exact) names with exact or very close age matches.

New methods in probabilistic linking relax these assumptions and allow machine models to weight different kinds of disagreements in names and ages.¹⁰ The key insight is that the best link may not *exactly* match on name (or phonetically cleaned name) or age as in Figure 1B and Figure 1D but it may dominate other candidates when simultaneously considering *both* age and name differences.

One class of machine-learning algorithms are known as supervised learning methods and use “training data” to classify matches. Training data may be a subset of data coded by humans (sometimes genealogists and sometimes by others) or result from the observation of true links (called ground truth). If the training data are ground truth and the model is well specified, the computer will learn how to classify links to approximate this truth. However, if the training data are of limited quality, the computer algorithm will replicate these incorrect decisions. Another potential limitation is that if the training data have little in common with the records to be linked, the supervised-learning algorithm will have unpredictable performance. Consequently, the advantage and disadvantage of supervised-learning algorithms is that they depend heavily on the quality of the training data and its similarity with the data to be linked.

The Minnesota Population Center (MPC) uses a supervised learning method to create the Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS), a set of links between the 1850 to 1930 one-percent samples and the 1880 Census (Ruggles et al. 2015). Using clerically reviewed data, MPC trains a Support Vector Machine (SVM) using features of matches that, after specifying a few tuning parameter choices, classifies links as true or false (Goeken et al. 2011). Illustrating the conservative nature of this approach, MPC produced final match rates of 12 percent for native-born whites, 3 percent for foreign-born whites, and 6 percent for African Americans for the 1870-1880 links.¹¹ Unfortunately, this model is proprietary and we cannot use it in our analysis.

In a similar spirit, Feigenbaum (2016) uses a supervised-learning technique to link the 1915 Iowa state Census to the 1940 U.S. Census. After creating training data by hand, he estimates a probit model to quantify the joint importance of different record feature in determining a link, including name Jaro-Winkler scores, differences in age, indicators for Soundex matches of first or last name, indicators for matches in letters or names, and indicators for matching truncated first or last names. He then tunes his model so that a link is only chosen if its probability of being a match is sufficiently high and sufficiently greater

¹⁰See Mullainathan and Spiess (2017) for a useful primer on machine learning for economists.

¹¹Researchers have used the IPUMS-LRS for a variety of research questions, including the economic effects of racial fluidity (Saperstein and Gullickson 2013), long-term differences in black and white women’s labor-force participation (Boustan and Collins 2014), and intergenerational co-residency (Ruggles 2011).

than the second-best candidate's match probability (if a second-best candidate exists). These cutoffs are derived from training data to assess the Type I and Type II errors of different choices. Feigenbaum (2016) achieves a match rate of 57 percent.¹²

D. Abramitzky, Mill, and Pérez (2018)

An alternative to supervised learning is unsupervised learning, an approach which evaluates the quality of different links without training data. These algorithms depend on using observed patterns in the data to classify the data by quality of potential link. Similar to other deterministic methods like Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014), an advantage of unsupervised learning is that it creates links without training data. Moreover, if training data are error ridden or too different from the dataset of interest, the lack of reliance on them is a feature. Not relying on human decisions may also be a limitation, because the algorithm's performance depends on difficult-to-validate modelling assumptions.

Abramitzky, Mill, and Pérez (2018) use unsupervised learning in the form of the expectation-maximization algorithm (Fellegi and Sunter 1969, Winkler 2006, Dempster, Laird, and Rubin 1977) to link Censuses in the U.S. and Norway. Building on Mill (2013), they fit a mixture model that allows for conditionally independent multinomial probabilities of specific age distances and discretized Jaro-Winkler scores for two distributions. They then fit this model with observed data using the expectation-maximization algorithm. Then, using their results, they calculate the estimated probability that a given potential match is a correct match conditional on the Jaro-Winkler score and age distance of the match. Like Feigenbaum (2016), they create a final set of matches by applying cut-offs to these estimated probabilities so that links are chosen that reach a sufficiently high estimated probability that is sufficiently greater than the second-best candidate match. However, unlike Feigenbaum (2016), these cutoffs are not guided by training data.

The approach is not completely automated, because it requires the user to define tuning parameters, including discretization thresholds of Jaro-Winkler scores and probability cut-offs for classifying a link.¹³ With regards to the latter, using lower cut-offs will create more matches but potentially include more marginal matches less likely to be correct. Conversely, higher cut-offs will create fewer matches but create matches that have a higher estimated probability of being correct. To address this trade-off, Abramitzky, Mill, and Pérez (2018) use two cut-offs, a more conservative and less conservative choice. Using these two cut-offs for their algorithm, they achieve match rates of 5 percent and 15 percent in their Census

¹²We focus on Feigenbaum (2016) in this analysis because it was developed for U.S. data, and is transparent and easy to replicate. Many other researchers have also incorporated probabilistic and machine learning. Mill (2013) and Mill and Stein (2016) use an expectation maximization method. Similar to the IPUMS-LRS, Wisselgren et al. (2014) use a support vector machine and link the 1890 Swedish Census to the 1900 Swedish Census in a few select parishes for match rates ranging from 25 to 72 percent. Antonie et al. (2014) link across historical Canadian census data and achieve linkage rates from 17.5 percent (Quebec) to 25.5 percent (New Brunswick). Other work uses [Ancestry.com](https://www.ancestry.com)'s search algorithm to link records. Bailey et al. (2011) link records of lynching to the 1900 to 1930 U.S. Censuses using the [Ancestry.com](https://www.ancestry.com)'s algorithm, linking 45 percent of their lynching records. Collins and Wanamaker (2014) and (2015) search Ancestry using Soundex for names as well as age and place of birth for white and black men ages 0 to 40 resident in Southern states in the 1910 Census. They match 19 and 24 percent of men, respectively, to a unique person in 1930.

¹³These choices may be consequential. For example, setting a high Jaro-Winkler similarity threshold corresponding to (0.92,1] assigns the same estimated match probability to a pair with first names, Katherine/Catherine, as to a pair with an exact match on first name, all else equal.

data. Our analysis implements these cutoffs and considers the effects of alternate cut-offs in an appendix.

In summary, existing linking methods involve a variety of modelling choices with unknown effects on data quality. Which set of assumptions should researchers use in different contexts? What are the implications of different assumptions for error rates? The next sections answers these questions by presenting a systematic comparison of methods in different records.

III. Data and Metrics of Automated Method Performance

Our analysis considers four different linking algorithms: Ferrie (1996); Abramitzky (Abramitzky, Boustan, and Eriksson 2014); regression-based, supervised learning (Feigenbaum 2016); and unsupervised machine learning (Abramitzky, Mill, and Pérez 2018). Detailed web appendices, published articles, and posted code make replicating these methods straightforward. Ferrie (1996) and Feigenbaum (2016) describe their methods step by step, which we implement exactly.¹⁴ We present Feigenbaum (2016) using both his regression coefficients for the Iowa Census-1940 training data (labeled “Iowa coef.”) as well as coefficients estimated using hand-linked samples (called “Estimated coef.”; see Appendix A for details and coefficient estimates). To implement Abramitzky, Boustan, and Eriksson (2014) and Abramitzky, Mill, and Pérez (2018), we use the code provided by the authors (see Appendix A) and report two cutoff implementations per the latter’s recommendation, “less conservative” and “more conservative.” For interested readers, we created a public Stata ado-file that implements these methods and the variations we consider in this paper (Bailey and Cole 2019).

A. Hand-linked and Synthetic Data

We examine the performance of each algorithm in two high-quality, hand-linked historical samples: the LIFE-M sample of birth certificates linked to the 1940 Census (Bailey 2018) and the Early Indicators Project’s genealogically linked sample of Union Army veterans (Costa et al. 2017).

The LIFE-M sample is based on a random draw from birth certificates from Ohio and North Carolina. These birth certificates are then linked to siblings’ birth certificates using parents’ names. We exclude girls because they typically changed their name at marriage in this era, making them hard to find as adults in the Census (see Appendix B). The LIFE-M sample consists of 42,869 boys born from 1881 to 1940, 24,408 of whom were born in North Carolina and 18,461 born in Ohio.

The LIFE-M sample of boys is then linked to the 1940 full-count U.S. Census using a semi-automated process, making use of both computer programming and human input. Our linking variables include first, middle (when available), and last name, birth state, and age.

¹⁴Unlike Ferrie (1996), we do not limit links based on family continuity. In addition, we treat records with multiple matches after the last step as having no link, although Ferrie reports having none of these instances and, therefore, does not indicate how he would have dealt with them.

We do not use race, because it is not available on all birth certificates (see section V.B for an analysis of this limitation).

After cleaning and standardizing the data, we use bi-gram matching on name and age similarity within a birth state to generate candidate links (Wasi 2014).¹⁵ Each candidate is independently reviewed by two “data trainers” who choose a correct link (or no link) from the set of candidates. If the two trainers agree, we treat their choice (link or no link) as the truth. In cases where the two trainers disagree, the records are independently *re*-reviewed by three new trainers to resolve these discrepancies (see section III.B.3).¹⁶ LIFE-M data trainers are instructed to reject links if they are not *completely* certain the links are correct.

LIFE-M trainers also work under a senior data trainer and receive multiple rounds of feedback across approximately 30 hours of work. Before trainers are allowed to work on the LIFE-M team, their decisions must reach a 0.95 correlation with a truth dataset.¹⁷ After trainers achieve this threshold, they begin receiving training batches from an automated distribution system, which guarantees that links are reviewed initially by two different trainers and that discrepancies are reviewed by three additional trainers. This automated system also distributes audit batches at least once per week to provide weekly feedback to trainers about their accuracy. Trainers meet weekly to discuss their mistakes, difficult cases, and learn about historical-contextual factors affecting the quality of the data.

The Family History and Technology Lab at Brigham Young University (BYU) performed two independent quality checks of the LIFE-M links. First, BYU research assistants used genealogical methods and multiple data sources to hand link a random sample of 543 of the 18,461 Ohio boys, 241 of which had been linked by LIFE-M. The BYU team had no knowledge of LIFE-M’s links. Among links made by both LIFE-M and BYU, BYU agreed with LIFE-M matches 93.4 percent of the time (16/241 matches were discordant). Second, BYU compared 1,043 LIFE-M links to those already on the [FamilySearch.org](https://www.familysearch.org) “Tree.” ([FamilySearch.org](https://www.familysearch.org) tree links are created by genealogists and users of [FamilySearch.org](https://www.familysearch.org), who are independent of the LIFE-M process.) For 1,043 birth certificates linked to the 1940 Census by LIFE-M and [FamilySearch.org](https://www.familysearch.org) users, the LIFE-M links agreed with [FamilySearch.org](https://www.familysearch.org) users 96.7 percent of the time. A link-weighted average of the two exercises implies that LIFE-M’s false link rate is around 3.9 percent. To account for potential errors in the LIFE-M data, we additionally require all links that differ from the LIFE-M sample to be re-reviewed using the “police line-up” process described in section III.B.

Our second sample is the Oldest Old sample of Union Army veterans from the Early Indicators Project. Costa et al. (2017) created this sample of 2,076 individuals at least 95 years old linked to the 1900 complete-count U.S. Census using genealogical methods and a

¹⁵We generate a set of candidate links using “relink.ado”, an algorithm that uses bi-gram comparisons of name strings. We also block on the first letter of last names to reduce computation time.

¹⁶“Data trainers” participate in a rigorous orientation process where they receive detailed feedback on their accuracy relative to an answer key. They continue this process for 10 to 20 hours per week until their matches agree with the truth dataset 95 percent of the time. After completing this orientation, trainers become part of the larger team that conducts independent clerical review.

¹⁷This truth dataset has been vetted by multiple individuals for accuracy. The cases for this truth dataset are selected to test the trainers’ knowledge and decision-making for a variety of linking problems.

rich set of supplementary information. These veterans tended to report complete and accurate information to ensure they would receive their army pensions and benefits. Moreover, sources such as gravestone databases, obituaries, newspaper accounts, veterans associations and pension files allow multiple cross-validation exercises, ultimately resulting in a high match rate of 90 percent among men confirmed to live beyond the 1900 Census. The Early Indicators Project scores matches on a scale of 1 to 4 to indicate their confidence in a match. We use 1,887 matches coded as the highest quality (1 and 2) as the hand-linked sample. Importantly, we do not use all possible records for which matches were attempted.

Because these hand-linked data may contain errors, we validate our conclusions by building a third sample: a synthetic ground truth. This synthetic ground truth adds noise to true links to mimic common errors in historical data while ensuring complete certainty about correct and incorrect links. That is, this synthetic dataset characterizes the performance of each matching algorithm relative to an *objective* truth, which shares important commonalities with the LIFE-M sample.

We construct the synthetic ground truth in two steps. First, we take all of our Ohio and North Carolina born boys' birth certificates, randomly drop 10 percent to reflect mortality and emigration and drop another 5 percent to reflect under-enumeration.¹⁸ Using the LIFE-M records as a basis allows us to retain sample name characteristics (e.g., ethnic origin and other conventions and name commonness). To account for orthographic differences in enumeration or transcription errors, we add noise to names and ages to reflect age heaping and transcription or digitization errors (Goeken et al. 2017, Hacker 2013, 2010).¹⁹ One limitation of this approach is that the true error structure in names and ages is unknown, so our decisions about how to simulate error may be simplistic and incomplete.

The resulting synthetic truth dataset is a noisy version of the truth for 85 percent of the Ohio and North Carolina boys. Then, we append to a random sample of boys from the 1940 Census who were born in Michigan, Indiana, Tennessee and South Carolina. Because these states neighbor Ohio and North Carolina, these individuals are incorrect links by construction. We chose these states because they share regional naming conventions and have similar demographic and economic characteristics. The size of our random sample of boys from neighboring states ensures that our set of candidates for each state has the same number of observations as in the LIFE-M linking exercise: 3,133,982 boys from the relevant age ranges born in Michigan and Indiana for Ohio and 1,904,592 boys born in Tennessee or

¹⁸Based on life tables from 1939 to 1941, we calculate that 8 percent of our sample should be un-linkable due to death prior to 1940 (National Office of Vital Statistics 1948). Moreover, Census analyses estimate that around 5.4 percent of individuals were missed in 1940 (West and Robinson 1999). This calculation leaves some scope (about 1.5 percentage points) for emigration, which reflects the fact that we think emigration for native-born boys would have been much lower than for those born abroad. To the extent that our approximation of emigration is too low, the actual Type II errors should be adjusted accordingly.

¹⁹To mimic age-heaping, 25 percent of ages are rounded to the closest multiple of 5. We introduce orthographic and transcription errors as follows. In 10 percent of cases, the first and middle names are transposed (if a middle name exists) and, in 5 percent of cases, the first and last names are transposed. In 5 percent of cases each, the first character of the first name or last name is randomly changed. In 5 percent of cases, each second character of the first name or last name is randomly changed. In 5 percent of cases, each third character of the first or last name is randomly changed. In 5 percent of cases each, we add a repeated letter to first names (e.g., "James" → "Jamees") or last names. In 5 percent of cases each, a random letter is dropped or two letters are transposed in the first or last name (e.g., "Matthew" → "Mathew" or "William" → "Willaim"). In 5 percent of cases, we replace the first name with an initial. In 50 percent of cases, we drop middle names (resulting in the same share of observations having middle names as is observed in the 1940 Census).

South Carolina for North Carolina. When linking to this synthetic dataset, we emulate the common process of blocking on birthplace and consider only the synthetic Ohio data as candidate matches for the Ohio boys and only the synthetic North Carolina data as candidate matches for the North Carolina boys.

B. Performance Criteria

We use four main criteria to measure performance. The first two are almost universally reported in papers using linked samples.

(1) Match rate: We calculate the match rate as the share of records that were linked of the sample that we attempted to link. Even if matching were perfect, this rate is expected to be less than 100 percent due to emigration and mortality. Notably, emigration and mortality are not expected to have different impacts by method, so they should not impact the relative performance of methods.

(2) Representativeness: We compare characteristics for the linked sample to the same characteristics for the unlinked sample. In a multivariate, linear probability model with Huber-White standard errors (Huber 1967, White 1980). A heteroskedasticity-robust Wald test of model significance tests the null hypothesis that the covariates are jointly related to successful linkage.²⁰ This straightforward, single summary metric and the regression coefficients describe the extent of non-representativeness as well as the subgroups that are under-represented.

These measures alone are inadequate to assess link quality. This fact is easily illustrated in an example. Consider a matching algorithm that *randomly* links individuals between two datasets. This algorithm would perform very well in terms of the first two criteria, because the entire sample would be matched and identical to the baseline sample in observed characteristics (and, therefore, representative). Few researchers, however, would want to work with these data, because—with large enough datasets—the incidence of false links would approach 100 percent.

We, therefore, use two more criteria to assess link performance (Abowd and Vilhuber 2005, Kim and Chambers 2012).

(3) False link rate (Type I error rate)²¹: We compare links for each automated method to a measure of the truth. We treat the high-quality, hand-linked Early Indicators dataset as the “truth,” given that genealogists have used multiple data sources to confirm each link. In the synthetic data, we know the true link, so we code differences in links between an algorithm and the synthetic data as Type I errors.

For the LIFE-M data, we subject discrepancies between the hand-links and the algorithm to an *additional* blind review. Similar to a “police line-up,” two reviewers independently review

²⁰We implement this in Stata by multiplying the F-statistic reported in Stata following a regression with robust standard errors by the relevant degrees of freedom parameter. Note that this test could be very conservative in the sense that it would reject the null hypothesis due to one variable’s significance in the regression and does not weight for the ‘importance’ of different covariates.

²¹Computer scientists focus on precision, or 1-T1 error rate presented here.

the LIFE-M link (made by hand), the link made by the automated method, and close candidate links. Reviewers may choose to code any of these links as correct or incorrect. This process gives the links from the hand-match and the automated method an equal shot at being chosen to avoid preferential treatment. For the LIFE-M data, only links that are rejected in clerical review as part of the police line-up are treated as Type I errors. This analysis may understate the true Type I error rate if the hand-links are incorrect and agree with the automated method.

(4) False negative rate (Type II error rate)²²: This metric captures the fraction of true links that are not found, or $1 - \text{Match Rate} * (1 - \text{Type I Error Rate})$. With this definition, the false negative rate can never be zero, because mortality and emigration mean that many individuals cannot be linked even with perfect data.²³

IV. The Performance of Prominent Automated Matching Methods

Because a central focus of a growing literature is linking to the newly available 1940 Census, we begin our analysis linking birth certificates to the 1940 Census. We then corroborate our findings using our synthetic ground truth and the Oldest Old sample from the Early Indicators Project.

A. Evaluating Algorithms Using the LIFE-M Data

Figure 2 compares the performance of selected, prominent automated linking methods to the hand-linked LIFE-M data, where each of these methods uses the *same* information to create links-name, age, and birth state. The length of each bar represents the match rate, computed as the share of the baseline sample of 42,869 boys who were successfully matched to the 1940 complete count Census. LIFE-M hand-review matched 45 percent of the baseline sample. Ferrie's (1996) method matched 28 percent of the baseline sample, and Abramitzky, Boustan, and Eriksson (2014) achieve a higher link rate of 42 percent. This result makes sense because Abramitzky, Boustan, and Eriksson (2014) do not impose Ferrie's (1996) uncommon name restriction. Feigenbaum's (2016) regression-based machine learning method matches 52 percent of the baseline sample both when using Iowa coefficients and when we estimate the coefficients using a random sample of the LIFE-M links. Abramitzky, Mill, and Pérez (2018)'s expectation-maximization method links 46 percent of the sample when using less conservative cutoffs and 28 percent of the sample with more conservative cutoffs.

Across the board, these match rates are higher than in the original studies. For instance, the Ferrie (1996) method matches 28 percent of the LIFE-M data versus his published figure of 9 percent of all men between 1850 and 1860 Censuses. Similarly, Abramitzky, Boustan, and Eriksson (2014) link 40 percent of the LIFE-M sample, whereas the same method links only

²²Computer science focuses on a similar statistic, "recall." This measure is defined as the number of true links found by the algorithm divided by number of linkable observations, or those linked by the data trainers.

²³Note also that, if the marginal link is more likely to be incorrect, an increase in the match rate within a specific algorithm has a weakly negative effect on Type II error rates and a weakly positive effect on Type I error rates. If the marginal link is wrong, then the Type II error rate would not change but the Type I error rate would increase. However, if the marginal link is correct, the Type II error rate would fall and the Type I error rate would decrease.

29 percent in Abramitzky, Boustan, and Eriksson (2012b) and 16 percent of native-born men in Abramitzky, Boustan, and Eriksson (2014). These higher match rates likely reflect two factors: the LIFE-M boys are on average 24 years old in the 1940 Census, so mortality and outmigration are likely lower for them than in other studies. In addition, birth certificate data quality is higher compared to other sources. Birth certificates (1) contain a complete and correct *full* name, often including middle names omitted in the Census; (2) record the exact date of birth rather than age in years;²⁴ and (3) capture the birth state by construction (it is issued by the birth state and so should not have reporting error like the Census).

Figure 2 also summarizes the share of links that human reviewers rejected in a blinded review using the “police line-up” method. These rejected links are presented in two ways. First, the share of the entire sample determined to be wrong for each method is displayed in red. For less than 2 percent of original sample, trainers reversed LIFE-M decisions upon re-review in favor of the link chosen by one of the automated methods. Consistent with genealogical validation by BYU, these reversals are rare. Second, the column on the far right in Figure 2 presents that share of *links* that were rejected by human reviewers (the estimated Type I error). We compute this share by dividing the share of the total sample that is incorrect by the match rate. Because the LIFE-M match rate is 45 percent, this implies a Type I error rate of 4 percent (approximately $0.017/0.445$). As shown in section VI, the implications of measurement error for inference is linked to the share of incorrect links, so our discussion focuses on this second metric.

Relative to clerical review, the share of false links for automated methods is higher across the board. The lowest Type I error rate occurs in the more conservative version of Abramitzky, Mill, and Pérez (2018) at 15 percent. Ferrie’s (1996) method of selecting uncommon names achieves the second lowest Type I error rate at 25 percent. These error rates are consistent with Massey (2017) who uses contemporary administrative data linked by Social Security Number as the ground truth. She finds that methods similar to Ferrie (1996) are associated with 19 to 23 percent Type I error rates. Abramitzky, Boustan, and Eriksson (2014)’s refinement of Ferrie (1996) increases match rates to 40 percent, but only half of the added links appear to be correct, and the Type I error rate increases to 32 percent. Feigenbaum’s (2016) supervised, regression-based machine learning model produces a Type I error rate of 34 percent when using the Iowa coefficients, and the Type I error rate decreases to 29 percent when estimated using LIFE-M data. Finally, Abramitzky, Mill, and Pérez (2018)’s less conservative cut-off results in the highest error rate at 37 percent. The difference between the conservative and less conservative versions of Abramitzky, Mill, and Pérez (2018) highlights the sensitivity of performance to the tuning parameters.

In terms of missed links, Ferrie (1996) correctly linked the lowest share of the sample without error at 21 percent, and Feigenbaum (2016)’s algorithm estimated with the LIFE-M data correctly linked the largest share of the sample without error at 37 percent. It is worth noting that Feigenbaum (2016) and Abramitzky, Mill, and Pérez (2018) allow for a variety of different choices of sample restrictions within their linking methods that alter the trade-off between Type I and Type II errors in their matches. We implemented versions of these

²⁴Massey (2017) shows that decreasing the noise in age results in higher match rates and lower Type I error rates.

methods that reflected how they were implemented in each. Our Appendix Figures A1 and A5-A7 shows how altering these restrictions impacts results in both cases.

Table 2 describes the representativeness of the linked sample. Because birth certificates do not contain socio-demographic measures found in the Census (race, age, or incomes of the parents), we regress a binary dependent variable (1= linked records) on a variety of covariates from the birth certificates. These variables include the individual's exact date of birth;²⁵ the number of siblings in the family; the number of characters in the infants' (boys'), mothers', and fathers' names—a characteristic which is strongly positively correlated with years of schooling and income from wages in the 1940 Census; and the share of family records with a misspelled mother's or father's name, which we expect to be negatively correlated with years of schooling and income (Aizer et al. 2016).²⁶ Table 2 presents the Wald-statistic for tests of whether these covariates are jointly associated with an observation being linked (p-value beneath). If a representative set of birth certificates were linked, then these characteristics would not be jointly related to whether an observation was linked. However, Wald-statistics for the joint test of the association of these characteristics with linking show a persistent association. For all methods, including LIFE-M's clerical review, we reject representativeness at the 1-percent level.

The signs and magnitudes of the regression results provide clues about the individuals easier to link (see the full set of regression results in Appendix C). Many automated methods are more likely to link boys with higher incidence of misspelled father's last name, and more likely to link boys with a longer mother's name. All methods except Feigenbaum (2016) with estimated coefficients are more likely to link children with longer names. Based on the correlation of name length with wage income in the 1940 Census, this finding indicates that linked records are drawn from more affluent families. Some methods are more likely to link individuals with more siblings, while other methods are more likely to link individuals with fewer siblings. In short, even though no linking algorithm generates representative samples, different algorithms yield samples that are non-representative in different ways.

Finally, Table 3 tests for the systematic correlation of links rejected in hand review with baseline characteristics. The method is identical to what is presented in Table 2 but that the dependent variable is equal to 1 if the link was rejected in a blind review. If the rejected links are systematically related to baseline characteristics, this suggests that the algorithm introduces systematic measurement error in variables of interest. Column 1 of Table 3 reports the heteroskedasticity-robust Wald-statistic (p-value beneath) by method for the LIFE-M data (see the full set of regression results are in Appendix D). For each algorithm, we reject the null hypothesis that errors in linking are unrelated to baseline characteristics at the 1-percent level. False links are significantly negatively associated with the length of a mother's name and length of a father's name in nearly all samples, suggesting that being falsely linked is negatively associated with affluence. Patterns across other variables are more varied. For example, the number of siblings is positively associated with the

²⁵Exact day of birth (1-366, due to leap years) is as close to a continuous measure as we can get in historical records, and season of birth is strongly correlated with socio-economic characteristics in modern data (Buckles and Hungerman 2013).

²⁶We measure misspellings in father and mothers' names as the number of name spellings in the birth certificates of all siblings that differ from the modal spelling divided by the total number of children in a family.

probability that a link is incorrect for the Feigenbaum (2016) algorithm, but the number of siblings is negatively associated with the probability that a link is incorrect in the Abramitzky, Boustan, and Eriksson (2014) sample. In short, different algorithms appear to induce different types of systematic measurement error.

B. Evaluating Algorithms Using the Synthetic Ground Truth and Early Indicators Samples

One critique of these findings is that human errors survive even the blind review process. This could lead the *incidence* of Type I errors in the LIFE-M analysis to be too high or too low relative to the truth. To address this potential issue, we reevaluate algorithm performance in synthetic data (where the truth is known). Because this objective truth is not influenced by human reviewers at all, this exercise validates those obtained from human review. We also evaluate the same algorithms using the Early Indicators, data, proving a complimentary perspective using a sample that was linked by genealogists and is known to be highly accurate.

For both the synthetic and Early Indicators data, Table 1 compares the match rates and error rates for each prominent algorithm. Recall, for the synthetic data, a perfect match rate is 85 percent, because 15 percent of the original links are absent by design. Patterns in match rates across methods are slightly higher in the synthetic data but generally within a few percentage points of the LIFE-M match rates, with the exception of Feigenbaum (2016) and Abramitzky, Mill, and Pérez (2018). Notably, both methods perform substantially better in the synthetic data than in the LIFE-M data with match rates of 56 and 57 percent for Feigenbaum (2016) with the Iowa and estimated coefficients and 52 and 32 percent for Abramitzky, Mill, and Pérez (2018) with the less and more conservative cutoffs. The match rates for Early Indicators' veterans linked to the 1900 complete count Census are generally higher than in the LIFE-M sample, which reflects the fact that all individuals in these data are known to be linkable. The LIFE-M data, however, contains both individuals who can be linked and those who cannot.

Table 1 also shows that patterns of error rates in the synthetic and Early Indicators data are similar to those in LIFE-M. Importantly, the best performing algorithms in LIFE-M continue to perform the best in the synthetic and Early Indicators data. Figure 4 describes patterns of error rates graphically across algorithms and datasets. In most cases, the error rates are slightly lower in the synthetic data and Early Indicators data relative to the LIFE-M records, ranging from 11 to 33 percent. Because there was no hand linking involved in producing the synthetic data, similar error levels suggest that the results of the LIFE-M hand-linking process and blind review reflect *true* errors in the automated linking algorithms. Larger reductions in error rates for machine-learning algorithms like Feigenbaum (2016) and Abramitzky, Mill, and Pérez (2018) suggest that these methods may be effective at detecting the simple errors we simulated. Because the Early Indicators data contain only individuals who have been successfully linked by genealogists, Type I error rates are lower, ranging from 10 to 24 percent versus 15 to 37 percent in the LIFE-M data. The fact that the *patterns* of error rates are the similar in all datasets, however, provides strong support for the notion that prominent machine linking methods in current practice make considerable errors.

The findings for representativeness in the synthetic and Early Indicators data are also similar, with Table 2 suggesting that the linked samples are unrepresentative. For the synthetic data, this exercise allows a particularly strong test of the hypothesis that the non-representativeness of linked samples reflects the linking algorithm *per se*. Because we randomly dropped 15 percent of individuals, non-random attrition due to differential death, enumeration, or emigration is ruled out by construction. The only reason that the linked synthetic sample would not be representative is that the methods link certain groups more systematically than others. Consistent with this hypothesis, the Wald-statistics and p-values in column 2 reject representativeness for all methods in the synthetic data at the 1-percent level. Most methods are less likely to link individuals with more siblings. In the Early Indicators data, nearly all methods are more likely to link individuals with U.S.-born mothers; some methods are more likely to link individuals with longer first or last names, while others exhibit the reverse correlation.²⁷ (See Appendix C for the full set of regression results.)

Table 3 underscores the finding that false are systematically related to baseline sample characteristics. For all methods in both the synthetic and Early Indicators data, we reject the null hypothesis that false links are unrelated to baseline covariates—a pattern that may complicate inference by introducing systematic measurement error. (See Appendix D for the full set of regression results.)

C. Summary of Findings

Prominent algorithms yield widely varying results—even using the same data and linking variables. In the LIFE-M data, match rates range from 28 to 52 percent, while the share of links rejected in the police line-up ranges from 15 to 37 percent and the associated Type II error rate ranges from 63 to 79 percent. A synthetic ground truth dataset confirms these patterns and also suggests that machine-learning algorithms like Feigenbaum (2016) and Abramitzky, Mill, and Pérez (2018) are effective at detecting and correcting for synthetic errors, which speaks to their potential value in improving the quality of linked data.

An equally important finding is that error rates vary across datasets—even when links are created using the same algorithm and linking variables. The share of links rejected by humans in the Early Indicators data is slightly lower and the match rates are higher, possibly owing to the fact that the data consist of individuals selected on having been linked by genealogists (i.e., living in the U.S. and also less likely to have changed a name or its spelling). However, error rates likely differ across the datasets due to differences in data quality that are difficult to easily measure and diagnose. This variation across datasets cautions against generalizing this paper’s findings and recommends that researchers examine their linked samples for clues about error rates, representativeness, and systematic measurement error.

²⁷For the synthetic dataset, we use the same covariates as in the LIFE-M data when considering representativeness. For the Early Indicators data, we use continuous variables in age and length of first and last names and dummy variables for speaks English, owns a farm, currently married, foreign born, day of birth by year, literacy, and foreign born status of parents.

V. How Variations in Algorithms Alter Method Performance

Understanding what drives differences in algorithm performance is key to improving existing methods and current practice. This section considers how the performance of these algorithms changes when varying key features of their set-ups. First, we examine the role of different phonetic name cleaning strategies for algorithms that require agreement in cleaned names. Second, we extend the Ferrie (1996) algorithm to include more common names or eliminate them using a narrow age-band as in Abramitzky, Boustan, and Eriksson (2014)'s robustness test. Third, we examine equal weighting of exact ties (i.e., multiple, exact matches). A final section examines the robustness of these findings across all methods to using middle names, information on race, and extensions to population-to-population linking.

A. Phonetic Name Cleaning, Common Names, and Ties

Phonetic string cleaning algorithms account for orthographic differences that could lead a true match to be missed, such as minor spelling differences, name Anglicization, and transcription/digitization errors. Figure 3 and Table 4 show how the performance of the Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014) algorithms vary with three types of phonetic name cleaning: no cleaning (labeled “Name”), Soundex (labeled “SDX”), and NYSIIS. Interestingly, although name cleaning is intended to increase match rates, it can also *decrease* match rates if it increases ties (by removing meaningful spelling variations). This interaction is important for the Ferrie (1996) method, which matches between 20 and 33 percent of baseline sample depending on the phonetic name cleaning used. Because this cleaning creates more common name strings, and Ferrie's algorithm restricts the sample to uncommon names, the algorithm discards more links due to ties when cleaning is used: the match rate falling from 33 (Name) to 28 (NYSIIS), to 20 percent (Soundex). Because Abramitzky, Boustan, and Eriksson (2014) does not restrict to uncommon names, it does not show reductions.

The likelihood of Type I errors increases with the use of phonetic name cleaning in Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014). As shown in Table 4 and Figure 3, using NYSIIS rather than uncleaned names increases Type I error rates by an average of 18 percent, ranging from as little as 5 to as much as 25 percent across datasets. Using Soundex rather than uncleaned names increases Type I error rates by an average of 36 percent, ranging from 14 to as much as 64 percent. These increases occur because, in addition to orthographic and transcription errors, phonetic codes may remove *meaningful* spelling variation. For example, both Soundex and NYSIIS would code “Meyer” and “Moore” as the same name, whereas reviewers tend to treat these as different names. Counterintuitively, the increase in error rates induced by phonetic name cleaning universally *decreases* the share of the sample that is correctly linked.

Another modification to Ferrie (1996) is to link more common names. Recall that—for computational reasons—Ferrie (1996) discarded matches if there were 10 or more candidate matches, regardless of age differences. We relax this assumption and include records with 10 or more candidates when we find links (labeled “Ferrie 1996 + common names”). Table 4 and Figure 3 show that including common names results in significantly higher match rates,

including higher shares of true links than in the original method. The results are almost identical to those of Abramitzky, Boustan, and Eriksson (2014), which is because this method's key deviation from Ferrie's (1996) is attempting to link more common names. The inclusion of common names roughly doubles the share of incorrect links but it also decreases Type II error rates.

Similar in spirit to the common names restriction, Abramitzky, Boustan, and Eriksson (2014) implement a robustness check that discards links if a name tie occurs within a two-year age band. As reported in Figure 4 under Abramitzky, Boustan, and Eriksson 2014 (NYSIIS, Robustness), this restriction lowers the match rate to 24 percent but it also halves the share of the sample that is incorrectly linked, from 14 percent to 6 percent, and changes the Type I error rate from 32 percent to 23 percent. This robustness check is very similar to the uncommon names restriction and, therefore, performs almost identically to Ferrie (NYSIIS) in terms of match rates and Type I errors. Relatedly, the Abramitzky, Mill, and Pérez (2018) algorithm with more conservative cut-offs is also similar to an uncommon names restriction. By requiring a high threshold of the probability of a candidate match being a correct match, and requiring the second-best options for the observations in the match to be much lower, this restriction ensures that the links made have no close analogues either due to differences in spelling or age. This unsupervised approach slightly outperforms Ferrie (NYSIIS) in terms of Type I and Type II error rates.

A third variation on prominent algorithms relates to how ties are handled (e.g., cases like M3 in Figure 1C and M4 in Figure 1D). Ties are very common in contexts with limited information (such as matching between U.S. Censuses). If one could break *exact* ties or use ties in the analysis, researchers could match the majority of the sample, raising match rates substantially. For the Ferrie (1996) algorithm, using both common names and ties raises the match rates from 20 to 33 percent to 69 to 86 percent.

Two main approaches to using exact ties have been suggested by the literature. First, the statistics literature offers an alternative to tie-breaking by using probability weighting. For instance, one could use a weight that is the conditional probability that the match is correct (Scheuren and Winkler 1993, Lahiri and Larsen 2005). In the absence of other data features, this suggestion simplifies to weighting by $1/J_r$, where J_r is the number of exact ties for record r .²⁸ Nix and Qian (2015)'s *random* selection among ties is similar in spirit. Their process draws *one* of the candidate matches with probability $1/J_r$. Importantly, simple probability weighting and random selection among ties have the same expected performance in certain contexts such as those we consider later for our case study.²⁹ We label results that include ties as "Ferrie 1996 + ties." Table 4 shows that including ties may dramatically increase

²⁸This simple probability weighting differs from Lahiri and Larsen (2005), because match probabilities vary in their data due to a specifically defined data generating process, and they are able to trim candidate links with lower match probabilities. We do not assume a specific data generating process. Furthermore trimming is not possible when all records are equally tied.

²⁹We explain this result later in more detail, but the intuition is straightforward. Let N =the number of observations, M =the number of primary records with multiple exact ties as their best matches, J_r the number of ties for a primary record $r=1,2, \dots, M$. Assuming that one of the ties is the correct link, the expected number of false matches for records with ties is

$\sum_{r=1}^M (J_r - 1) / J_r = M - \sum_{r=1}^M 1 / J_r$ for both random selection and simple probability weighting. As the number of multiples increases for a given record, the probability weight on a false match gets smaller as does the weight on the true match. The results from probability weighting may differ slightly due to sampling variation.

match rates, but Figure 4 shows that this substantially increases the share of observations that are incorrectly matched in every sample. This makes sense, because at most one of the candidate links can be correct. For instance, if there are ten candidate “John Smith” links and only one of these is the correct link, nine out of ten of these links are incorrect. Notably, the Type I error rate is higher in the Early Indicators data, reflects the fact that they are selected upon being successfully linked and have fewer close ties.

Figure 5 describes the mixed progress in historical automated linking since 1996. As the literature has moved from the use of Ferrie’s (1996) uncommon name sample and increased match rates, some methods have also increased Type I errors (and decreased precision). The hope of researchers using these methods is that—on net—they are increasing the share of true links in their sample as well as sample representativeness. However, for the synthetic and Early Indicators data, the pattern of Type I and Type II errors suggest that there may be scope to improve in both dimensions by leveraging the strengths of different algorithms.³⁰ Similar to the findings for prominent methods, each of these variations produces samples that are unrepresentative (Table 5). Moreover, these variations produce false links that are systematically related to baseline sample characteristics in all datasets (Table 6).

B. Robustness: Middle Names, Race, and Population-to-Population Linking

How much should we expect these results to change with the addition of information commonly available in historical datasets? A first robustness check considers how the addition of middle name or race could reduce Type I error rates. For middle names, we examine a subsample of cases where middle name or initial was available for both the birth certificate and the linked Census record. Then, we calculate the number of false links that would have been eliminated had the automated method required that middle initial match for all potential matches *after* running a matching algorithm. We apply this restriction *ex post*, but it would also be possible to include middle name agreement in the matching process as a feature considered by an algorithm in the process of making matches.

Table 7 shows that the availability of middle initials may reduce match rates but also the reduce rate of false matches. For comparison, columns 1 and 2 reprint the information on match rates and Type I error rates from Table 4. Column 3 shows the share of matched observations that have information on middle initial in the birth certificate and the Census record, which range from a quarter to a third of matches. Column 4 reports the share of matches that have discordant middle initials among the subset of matches that have middle names in both records, ranging from 20 percent to 57 percent. Column 5 reports the Type I error rate among the matches with discordant middle initials. What is clear is that the Type I error rate is always above 87 percent within this subset. Presumably, these error rates are high because disagreements among middle initials mattered to the trainers considering these observations when making matches. Finally, columns 6 and 7 recalculate the match rate and Type I error rates after dropping observations that have discordant middle initials. Match

³⁰Of course, the *level* of Type II errors in the LIFE-M and synthetic data is overstated, because some infants did not survive until the 1940 Census, emigrated, or were missed by enumerators in the 1940 Census. We estimate that these factors likely account for around 15 percent of missed links (see footnote 16). A linking method that linked all LIFE-M or synthetic data individuals correctly would locate at (0.15, 0), missing only the 15 percent individuals who are unlinkable and making no errors. Because these sources of attrition affect all methods equally, these factors do not influence our comparisons across methods.

rates tend to drop by several percentage points, but Type I error rates drop by more. For example, Ferrie (1996) with NYSIIS drops from a match rate of 28 percent and an associated Type I error rate of 25 percent to a match rate of 26 percent and an associated Type I error rate of 20 percent. Type II error rates are nearly unchanged despite the drop in match rates, with all changes in Type II error rates never exceeding one percentage point. This result suggests that the addition of more information contained in middle names can substantially reduce Type I linking errors with minimal changes in Type II errors, at least among observations that have middle initials in the LIFE-M data.

A second robustness check compares “race” indicated on the 1940 Census for LIFE-M linked records to the race reported on an individual link.³¹ Column 3 of Table 8 shows the share of linked birth certificates that would *not have been erroneously linked* by an algorithm that blocked on race. Interestingly, only a small share of incorrect links have discordant races, ranging from 0 to 5 percent. When we omit incorrect links with discordant races in column 5, we find that the match rate drops slightly and the Type I error rates decrease by no more than two percentage points across methods. This is consistent with Massey (2017), who finds that errors in linking the 2005 *Current Population Survey* to the Numident only decreased by 0.07 percentage points when blocking on race. In contrast to using middle initial, race does not appear to add much information to reduce errors. Note, however, that including race may result in better link disambiguation among links that appear to be close substitutes.

A third robustness check examines to the implications of linking a sample (rather than a population) to a Census. The critical difference in these two settings is that an observation could appear to be unique in a sample while having duplicates or near-duplicates in the population. This is important because many algorithms drop a match if (1) it occurs more than once in the set of records to be linked or (2) more than one observation links to the same record. Both are more likely to occur for a population than a sample. Sample-to-population linkage may, therefore, result in a higher share of incorrect matches than population-to-population linkage.

To quantify the importance of linking a sample to a population as we do here, we compare our results to matching the *universe* (e.g., the population) of birth certificates to the 1940 Census. First, we match the universe of Ohio and North Carolina birth certificates to the 1940 Census using each automated method, including adjustments described above. Then, we isolate attention to the subset of records in the LIFE-M sample to assess performance. Since the LIFE-M sample is a random subsample of birth certificates, we expect the results to generalize to the population.

Figure 6 displays the results, with the horizontal axis depicting Type I error rates in matching between the *sample* and the 1940 Census and the vertical axis displaying the results matching the *population* of birth certificates to the 1940 Census. Results along the 45-degree line indicate perfect agreement in the two procedures. As expected, false link rates

³¹To the extent that some individuals “pass” for other races, this robustness check may eliminate true links (Nix and Qian 2015, Mill 2013). Note that race is only observable for the observations that LIFE-M successfully linked, as we infer race from the 1940 Census given that it is not reported on birth certificates.

fall below the 45-degree line for all methods that use the post-linking adjustments, suggesting that Type I error rates are somewhat higher for sample-to-population linkage. The Type I error rate in Ferrie (1996) with exact names is 20 percent for a sample but 17 percent for the population; for Abramitzky, Boustan, and Eriksson (2014) with exact names, the Type I error rate is 25 percent for the sample versus 20 percent for the population. In short, population-to-population linking may reduce errors but the improvements are not large, as no method achieves a Type I error rate lower than 15 percent.

C. Summary of Findings

Variations on machine-linking algorithms may improve or worsen performance. Deterministic algorithms that clean names using Soundex or NYSIIS perform worse than using raw name strings. Similarly, trying to link common names (especially in conjunction with phonetic name algorithms) tends to increase error rates *and* the share of the initial sample correctly linked. Tie breaking or weighting ties equally could dramatically increase sample sizes but may have the unintended effect of using more incorrect matches in analyses. Including middle names as a linking criterion appears to have large effects on Type I error rates. However, using race information or using population-to-population linking appears to alter data quality only modestly.

VI. How Automated Methods Affect Inferences

Our final analysis explores the consequences of Type I and Type II errors for inferences about historical rates of intergenerational mobility. Following the intergenerational literature (Solon 1999, Black and Devereux 2011), we consider the following benchmark specification,

$$\log(y) = \pi \log(x) + \varepsilon, \quad (1)$$

where the dependent and independent variables have been rescaled to capture only individual deviations from population means. The dependent variable, $\log(y)$, refers to the log of son's wage income in adulthood in the 1940 Census. The key independent variable, $\log(x)$, refers to the parent's log wage income in the 1940 Census. Within this framework, we interpret π as the intergenerational income elasticity. The magnitude of π is an important indicator of the role that parents' wage incomes play in determining their children's wage income. Intergenerational mobility is measured as $1 - \pi$, which is often regarded as a metric of economic opportunity.

Our analysis uses the LIFE-M sample of 19,486 boys (43 percent of the 45,442 that were linked to the 1940 Census) and samples linked using different automated methods to estimate intergenerational mobility. Unlike other analyses using the Census and *Panel Survey of Income Dynamics*, we must link fathers from birth certificates to the 1940 Census to obtain their income information. Links for fathers are obtained using only the LIFE-M clerical review method, so that father links remain constant in all regressions. By using the same links for fathers and different methods to link sons, our analysis describes differences in the estimates that are driven by differences in methods used to link sons.

A. How Type I Errors Affect Inferences

Different kinds of Type I errors could have different implications for inferences about intergenerational mobility. Within the regression framework in equation (1), measurement error in son's income (the dependent variable in the regression) that is uncorrelated with father's income will still allow us to estimate π consistently using OLS, though the estimates will be less precise. However, measurement error on the right-hand side in father's income (the independent variable in the regression) is more consequential. At first glance, considering measurement error in father's income seems counter to our problem of using different linking methods to link sons. Note, however, that linking a boy to the wrong man in 1940 is equivalent to assigning the *wrong father's income* to that man.

Our conceptual framework for thinking about linking-induced measurement error is similar to Horowitz and Manski (1995). We assume that a linking method, ℓ induces Type I error in matches by erroneously linking a father to a son (we do not derive bounds here, but that is a useful avenue for future research). The presence of this measurement error allows us to divide the sample into two groups, g : one for which the links are correct, denoted with a *, and another for which the link is imputed (or incorrectly classified), i . Following Greene (2008) and Stephens and Unayama (2017), we decompose the OLS estimate of π for a sample linked with method, ℓ into the sum of within and between covariance for the correct, *, and imputed groups, i . b denotes the between component. Let

$$s_{xy}^{\ell*} + s_{xy}^{\ell i} = \sum_g s_{xy}^{\ell g} = \sum_g \sum_k (\log(x_{kg}) - \overline{\log(x_{kg}))} (\log(y_{kg}) - \overline{\log(y_{kg}))},$$

$$s_{xy}^{\ell b} = \sum_g N_g (\overline{\log(x_{kg})} - \overline{\overline{\log(x_{kg}))}}) (\overline{\log(y_{kg})} - \overline{\overline{\log(y_{kg}))}}) \text{ where group means are defined with a single bar and overall means are defined by two bars, such that,}$$

$$\hat{\pi}^{\ell} = \frac{s_{xy}^{\ell}}{s_{xx}^{\ell}} = \frac{s_{xy}^{\ell*} + s_{xy}^{\ell i} + s_{xy}^{\ell b}}{s_{xx}^{\ell}} = \frac{s_{xx}^{\ell*}}{s_{xx}^{\ell}} \hat{\pi}^{\ell*} + \frac{s_{xx}^{\ell i}}{s_{xx}^{\ell}} \hat{\pi}^{\ell i} + \frac{s_{xx}^{\ell b}}{s_{xx}^{\ell}} \hat{\pi}^{\ell b}. \quad (2)$$

Equation (2) shows that an OLS estimator converges in probability to a weighted average of the plim for the correct links, $\hat{\pi}^{\ell*}$, imputed links, $\hat{\pi}^{\ell i}$ and the between group term (* versus i), $\hat{\pi}^{\ell b}$, Where the weights on each term reflect the share of variance due to each component, θ :

$$\text{plim } \hat{\pi}^{\ell} = \theta^{\ell*} \text{plim } \hat{\pi}^{\ell*} + \theta^{\ell i} \text{plim } \hat{\pi}^{\ell i} + \theta^{\ell b} \text{plim } \hat{\pi}^{\ell b}. \quad (3)$$

The between group component can be thought of as the “selection” term. In some cases, we expect that the plim of the between term is zero (e.g., if the means of son's income or father's income are the same for the imputed and correctly linked groups). This pattern could happen in practice if errors (e.g., enumeration or transcription error) randomly assign records to these groups. Initially, we assume this term is zero to simplify exposition but later relax this assumption. Note also that, if the variances of father income are equal across all groups, the weights θ become the share of the sample in each category.

Now, consider the probability limit of the two remaining non-weight terms, $\hat{\pi}^{\ell*}$ and $\hat{\pi}^{\ell i}$. The first term represents the elasticity for the linked subsample, $\text{plim } \hat{\pi}^{\ell*} = \pi$. The second term is an estimated elasticity for the imputed observations. If we assume $\text{cov}(\varepsilon, \log(x^{\ell}))=0$. then

$$\begin{aligned} \text{plim } \hat{\pi}^{\ell i} &= \frac{\text{cov}(\log(y^*), \log(x^{\ell i}))}{\text{var}(\log(x^{\ell i}))} = \frac{\text{cov}(\pi \log(x^*) + \varepsilon, \log(x^{\ell i}))}{\text{var}(\log(x^{\ell i}))} \\ &= \pi \frac{\text{cov}(\log(x^*), \log(x^{\ell i}))}{\text{var}(\log(x^{\ell i}))} \end{aligned} \tag{4}$$

If the imputed father’s income is the same as the true father’s income, $\log(x^*) = \log(x^{\ell i})$, then $\text{plim } \hat{\pi}^{\ell*} = \text{plim } \hat{\pi}^{\ell i}$. However, if $\frac{\text{cov}(\log(x^*), \log(x^{\ell i}))}{\text{var}(\log(x^{\ell i}))} \neq 1$, then $\text{plim } \hat{\pi}^{\ell i} \neq \pi$ and the degree of the inconsistency depends on the relationship between the true and imputed father’s income.

There are several special cases of interest. First, suppose that there is no relationship between the true father’s income and the imputed father log income, or that $\frac{\text{cov}(\log(x^*), \log(x^{\ell i}))}{\text{var}(\log(x^{\ell i}))} = 0$. Then, the $\text{plim } \hat{\pi}^{\ell i} = 0$ and the estimator is inconsistent in proportion to the share of imputed links, $\hat{\pi}^{\ell} = \theta^{\ell*} \pi$. Second, consider the case where imputed father’s income equals the true father’s income plus noise, or $\log(x^{\ell i}) = \log(x^*) + u$. Under the assumptions of the classical errors in variables model ($\text{plim}(ue) = 0$, $\text{plim}(u \log(x^*)) = 0$, and $\text{plim}(u \log(y)) = 0$), then $\text{plim } \hat{\pi}^{\ell i} = \theta^{\ell i} \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \right) \pi$. Moreover, plim

$\hat{\pi}^{\ell} = (1 - \theta^{\ell*})\pi + \theta^{\ell*} \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \right) \pi$. Third, it is well known that non-classical measurement error for the imputed fathers could lead to under or over-statement of the parameter of interest, $\text{plim } \hat{\pi}^{\ell i} > \pi$ or $\text{plim } \hat{\pi}^{\ell i} < \pi$.

As a final exercise, consider the effect of Type I errors on inference using exact ties. Consider a setting where N is the total number of records that one wishes to link and for M of these records, $r=1, 2, \dots, M$, there are J_r candidate matches that are tied. For instance, if the first record with ties involves 30 potential matches for a John Smith, age 40, then for $r=1$, $J_1=30$. A second record, however, may only have 4 ties, so $r=2$ and $J_2=4$. Assume that there is one correct link among the ties for record, r , indexed by $j=1, \log(y)$, and imputed (but incorrect links) $\log(y_j)$, $j=2, \dots, J_r$. From the researcher’s perspective, the correct link is unknown and the probability that any one of the ties is correct is $\frac{1}{J_r}$.

Assume that one of these records would be selected at random to use in the analysis. By the same logic as in equation (2), a regression estimate of the intergenerational income elasticity can be decomposed into a variance-weighted sum of elasticities for three groups of observations—correct, unique links, denoted $*$; a correct link from the ties, denoted $j=1$; incorrect links from the ties, denoted $j>1$, and a “selection term” (which we assume is zero). Therefore, the estimated elasticity will be

$$\hat{\pi}^\ell = \frac{s_{XX}^* \hat{\pi}^{*\ell}}{s_{XX}} + \frac{s_{XX}^{j=1}}{s_{XX}} \hat{\pi}^{j=1,\ell} + \frac{s_{XX}^{j>1,\ell}}{s_{XX}} \left(\sum_{j=2}^{J_r} \frac{s_{XX}^{j,\ell}}{s_{XX}^{j>1,\ell}} \hat{\pi}^{j,\ell} \right)$$
 and

$$\left(\sum_{j=2}^{J_r} \frac{s_{XX}^{j,\ell}}{s_{XX}^{j>1,\ell}} \hat{\pi}^{j,\ell} \right) = \sum_{j=2}^{J_r} \lambda_j \text{plim } \pi^{j,\ell}.$$
 Therefore, the estimate of the intergenerational income elasticity using random selection to break ties is,

$$\text{plim } \hat{\pi}^\ell = \pi \left[\theta^{*\ell} + \theta^{j=1,\ell} + \theta^{j>1,\ell} \left(\sum_{j=2}^{J_r} \lambda_j \psi_{j,\ell} \right) \right] \tag{5}$$

Note that this estimator is inconsistent if $\psi_{j,\ell} = \frac{\text{cov}(\log(x_1), \log(x_j))}{\sigma_{x_j}^2} < 1$.³² The degree of

inconsistency is, again, determined by how much information is in the incorrect ties. If the weights, θ and λ , simplify to the expected shares of observations in each group (as they would if variances were equal across all groups as we note above), the degree of inconsistency is also related to the share of all records with exact ties, $\frac{M}{N}$ (implicit in the weight), as well as the number of multiples for each record, J_r .

These conclusions are identical if the elasticity is estimated with a probability-weighted estimator where the weight is the probability that *any* exact multiple in a set of exact multiples is the true match, or $1/J_r$. The probability limit of the estimator will be the same, although the performance of these methods may diverge in smaller samples.³³ This result is intuitive because the same share of imputed observations would be present using probability weighting or random selection for exact ties. In summary, the presence of imputed links—either through random selection or probability weighting—will generally lead to inconsistency, with the degree of inconsistency increasing in the number of records with ties, the number of exact ties for a given record, as well as the relationship between imputed observation and the truth. After examining the role of Type II errors, we examine the quantitative importance of these errors in a case study.

B. How Type II Errors Affect Inferences

Social scientists are accustomed to working with small representative samples. As long as links are representative of the underlying population, higher Type II error rates should only reduce precision. Across linking methods and datasets, however, this paper finds evidence that samples of links are *not* representative. If Type II errors result in the selective representation of different groups *and* the relationship of interest is heterogeneous across

³²Note that $\frac{\text{cov}(\log(y_j), \log(x_1))}{\sigma_{x_j}^2} = \frac{\text{cov}(\pi \log(x_j) + \epsilon, \log(x_1))}{\sigma_{x_j}^2} = \frac{\text{cov}(\log(x_1), \log(x_j))}{\sigma_{x_j}^2} \pi = \text{plim}(\pi^j).$

³³Reducing the influence of observations with less information is why some statisticians recommend truncating lower probability links, where presumably the covariance between the income of the father for the imputed link and the true link is small (Scheuren and Winkler 1993, Lahiri and Larsen 2005). Although Lahiri and Larsen (2005) propose an exactly unbiased estimator of π , this result only holds when the estimated link probability is uncorrelated with father’s income and where an exact data generating process for links is estimated. But this result breaks down in many historical settings, because the distribution of matching variables (name, age, and birth place) are correlated with outcomes and, often, a parent’s socioeconomic status (see Appendix C and D).

these groups, Type II errors may also lead to inconsistent estimates of population parameters in linked samples.

Heterogeneity in intergenerational income elasticities is believed to exist for many reasons. For instance, researchers have concluded that intergenerational income elasticities are larger for blacks than whites (Duncan 1968, Margo 2016) and that patterns of mobility are substantially different for farmers compared to non-farmers (Hout and Guest 2013, Xie and Killewald 2013). If one group is over-represented in the linked data, this will bias inferences about the historical rate of the population’s intergenerational mobility.

To make this point concretely, assume that the two groups in equation (3) are high mobility, h , and low mobility, l (rather than correctly and imputed links). Denote the intergenerational income elasticities of these groups as π^h and π^l where ($\pi^h > \pi^l$), and the share of the variation attributable to each group is θ^h and θ^l , respectively. Finally, assume that there are no errors in linking. Therefore, following the logic of equation (2), the regression estimate of the population elasticity parameter for a given linking method, ℓ is,

$$\text{plim } \hat{\pi}^\ell = \theta^{\ell h} \text{plim } \hat{\pi}^{\ell h} + \theta^{\ell l} \text{plim } \hat{\pi}^{\ell l} + \theta^{\ell b} \text{plim } \hat{\pi}^{\ell b} \tag{6}$$

The inconsistency of the probability limit in equation (5) depends upon several factors. First, if $\pi^h = \pi^l$ and the means for both groups of fathers and sons are the same, the selection term is 0 and having a non-representative sample will not affect inference. Having a representative sample matters *only* to the extent that the relationship of interest varies across those groups or the group’s characteristics differ. Second, if $\pi^h > \pi^l$ (and the group means are the same), Type II errors that effectively decrease the share of variation attributable to one group will lead to an inconsistent estimate of the population intergenerational elasticity parameter. Suppose that a linking method introduces Type II errors, which effectively decreases the variation attributable to observations representing the low mobility group. (In the extreme, high rates of Type II errors could imply that none of the total variation is attributable to low mobility group.) These Type II errors would result in an elasticity estimate that puts lower weight on the low-mobility group, resulting in a lower estimated elasticity. Third, if $\pi^h > \pi^l$ but the group means are different, then the selection term will not be 0 and inferences will be affected in an ambiguous way. Both heterogeneity and selection, of course, may vary greatly across samples. The following case study examines the combined implications of non-representativeness (through heterogeneity and selection) using inverse propensity weights to adjust for differences in observed characteristics (DiNardo, Fortin, and Lemieux 1996, Heckman et al. 1998).

C. Results: Intergenerational Elasticity Estimates from the 1940 Census

Different linking methods could have large effects on intergenerational income elasticity estimates through their influence on both Type I and Type II error rates. Figure 7A reports estimates of the intergenerational elasticity using samples of sons linked using different methods. For the LIFE-M links, we estimate an income elasticity of 0.24 between fathers and sons. Consistent with lifecycle bias and transitory income fluctuations attenuating our

estimates, this estimate is lower than modern estimates.³⁴ These biases, however, should not affect our comparisons *across* different linking methods for the same set of records.

Several important patterns emerge. First, higher Type I errors in matching tend to be associated with smaller intergenerational elasticities. Consistent with attenuation described in equations (4) and (5), estimates using linking samples with higher Type I error rates tend to be smaller. Using NYSIIS and Soundex tends to increase Type I error rates and produce smaller estimates than using the reported name. Moreover, weighting ties results in Type I error rates ranging from 50 to 67 percent and yields intergenerational income elasticity estimates of 0.19 to 0.11. However, Type I error is not the only factor determining bias. It is notable that the more conservative Abramitzky, Mill, and Pérez (2018), the method with the lowest Type I error rates, yields an intergenerational income elasticity that is 20 percent smaller and statistically different than the true coefficient. *These findings show that* bias introduced by choice of linking algorithm may be *as large as transitory income bias or life-cycle bias*, which has been an important concern of the recent literature on intergenerational mobility (Solon 1999, Haider and Solon 2006, Mazumder 2018). Conversely, a method with a comparatively high Type I error rate such as Feigenbaum (2016) achieves an estimated intergenerational income elasticity that is statistically indistinguishable from the LIFE-M elasticity. These results may reflect the fact that sample composition or a more systematic correlations of the errors with certain characteristics impact the coefficient. Appendix Figures A1 through A10 consider alternate implementations of Feigenbaum (2016) and Abramitzky et al. (2018). In general, the implementations that place more weight on precision (minimizing Type I errors) achieve estimated elasticities that are closer to the estimate with the hand-linked data. However, the decrease in Type I error is accompanied by a decrease in match rates and an increase in standard errors.

To examine the role of non-representativeness, we use inverse propensity-score weights to reweight the linked sample to have the characteristics of the LIFE-M birth certificate sample (Bailey, Cole, and Massey 2019).³⁵ Figure 7B shows that the reweighted intergenerational income elasticities tend to be slightly smaller in magnitude than the unweighted Figure 7A estimates. This result may stem from the modest over-representation of larger, less-mobile families in the linked sample. The attenuation of the coefficient for the more conservative Abramitzky, Mill, and Pérez (2018) is cut in half, however, by using weights. For this case study, however, the effects of non-representativeness (as measured by the changes induced by reweighting) on observed characteristics appear modest in comparison to the role of

³⁴For instance, Chetty et al. (2014) estimates 0.33, which is itself smaller than estimates for the same period using survey data (Mazumder 2015). Life-cycle bias may attenuate the estimated intergenerational elasticity regardless of matching method (Mazumder 2005, Haider and Solon 2006, Black and Devereux 2011, Mazumder 2015). In addition, wage income observed in the 1940 Census is an imperfect measure of permanent income, and we expect the single year observation of income for both generations can generate downward bias in estimated elasticities due to the importance of transitory income (Solon 1992, Zimmerman 1992, Mazumder 2005). On the other hand, the absence of farm and self-employed income in 1940 may lead this analysis to overstate mobility by excluding father-son pairs of farmers—an occupation that tends to be highly persistent across generations (Hout and Guest 2013, Xie and Killewald 2013). However, lifecycle bias and transitory income fluctuations should have similar effects for all methods.

³⁵To construct these weights, we first run a probit model of link status (for each method) on covariates, \mathbf{X} , which include an indicator variable for presence of middle name, length of first, middle, and last name, polynomials in day of birth, polynomials in age, an index for first name commonness, an index for last name commonness, number of siblings, an indicator variable for presence of siblings, and the length of own name as well as father's and mother's names. We then use the estimated propensity of being linked, $P(L_j = 1 | \mathbf{X}_j)$, for each method and reweight observations by $(1 - P(L_j = 1 | \mathbf{X}_j)) / P(L_j = 1 | \mathbf{X}_j) * q / (1 - q)$, where q is the share of records that are linked. Distributions of inverse propensity score weights are plotted in Appendix E.

errors in linking. Of course, one might also choose to re-weight the sample to resemble the 1940 Census. Appendix E shows that these results are nearly identical to results presented here.

While estimates using machine-linked samples appear attenuated relative to LIFE-M, the attenuation is not always as severe as one might expect with random error. For instance, Ferrie (1996) with name results in a 20 percent Type I error rate but the intergenerational elasticity estimate obtained from these links is one percentage point different from the LIFE-M estimate. If the selection term in equation (3) were zero, and the signal to noise ratio in equation (4) were zero, one would expect to estimate 0.19 ($=0.80 \times 0.24$). Therefore, one might think that fathers' incomes for imputed links positively covaries with the truth or that the Ferrie (1996) is positively selected on immobility. For tie-breaking methods, however, the attenuation appears more consistent with random error. For instance, Ferrie (1996) with common names and ties and Soundex shows a 69 percent Type I error rate and the intergenerational elasticity estimate is 0.11 in Figure 7A.

Figure 7C and Figure 7D *directly* examine the effects of incorrect matches by plotting $\hat{\pi}^*$, or the estimated elasticity for the “true” links (plotted as o with 95-percent confidence intervals) and $\hat{\pi}^i$, or the estimated elasticity for the “false” links (plotted as x) from separate regressions. Without the incorrect links, the estimates of the intergenerational income elasticity are very similar across groups at around 0.23 without weights (Figure 7C) and 0.23 with inverse propensity-score weights (Figure 7D). The comparability of unweighted estimates is especially striking given how different in size and representativeness the samples are. For instance, the number of true links varies from around 482 for Ferrie 1996 (Soundex) to 1,600 when using exact ties Ferrie 1996 (Soundex), but the unweighted intergenerational income elasticities are estimated to be 0.22 and 0.23, respectively. Consistent with Type I errors introducing attenuation, $\hat{\pi}^i$ tends to be smaller than $\hat{\pi}^*$ across methods. And, consistent with the observations about the magnitudes above, the unweighted estimated intergenerational elasticities for the imputed links for Ferrie 1996 (Name) are 0.15 and only 0.05 for Abramitzky, Boustan, and Eriksson (2014) (Soundex) and 0.05 for Ferrie 1996 with common names and ties (Soundex)—a statistical zero in the latter two cases. On the other hand, the correlation of incorrect links for Feigenbaum (2016) is very high, which shows how the regression-based classification system selects links with a very high correlation to the true link in this setting—even when incorrect. In short, the inclusion of imputed links appears to have large effects on OLS estimates of the intergenerational income elasticities, biasing them toward zero in most cases. After purging incorrect links, reweighting the linked sample to resemble the set of birth certificates has a minimal effect on the estimates. In short, smaller attenuation for some methods reflects the fact that the sample of correct links is selected on having *lower* intergenerational mobility (i.e., a higher elasticity), pulling the point estimate up, while the measurement error tends to pull the estimates down (see Figures 7C and 7D).

Given the expected attenuation due to transitory income and life-cycle biases, how do our estimates of intergenerational mobility relate to the literature? If we assume the independence of these two biases, the LIFE-M estimates imply that the true income

elasticity could be around 0.36 for the early 20th century (over 56 percent or $1/(0.64)$ larger than observed). This is similar in magnitude to Chetty et al. (2014) but smaller than in Mazumder (2018). However, our estimate is based on fathers and sons with wage income in Ohio and North Carolina. To the degree that the patterns of mobility present for this sample are not representative of the country as a whole, these findings may not generalize. For a similar period, Feigenbaum (2018), who links the 1915 Iowa Census to the 1940 Census, finds a correlation in log income earnings of between 0.20 and 0.26. This estimate is very similar to the LIFE-M estimate, although one might expect Iowa's intergenerational elasticity to differ due to Iowa's commitment to and investment in education (Goldin and Katz 2011).

Recent work by Ward (2019), however, recommends caution in interpreting early 20th century estimates of intergenerational mobility. His findings challenge the notion that the US was more mobile in the early 20th century. Correcting for measurement error—due both to linking error and occupational coding—using multiple observations of fathers' occupations suggests that intergenerational mobility may have *increased* (rather than decreased) over the 20th century. A better understanding of the role of measurement error in historical linking is, therefore, crucial for understanding whether the US held more opportunities for its citizens in the early 20th century—especially as researchers seek to create more linked historical data to understand the variation in mobility and also understand its causes.

VII. Lessons for Historical Record Linking

New large-scale linked data hold the potential to shift the knowledge frontier, increasing the urgency for developing reliable linking methods. Using different U.S. samples, this paper documents how linking algorithms and resulting errors could have large effects on scientific conclusions and policy inferences. Not only are linked samples not representative, but existing algorithms yield high rates of false matches. Moreover, the incidence of false matches are systematically related to baseline sample characteristics, suggesting that linking-induced measurement error may introduce complicated forms of bias into analyses. Our case study shows that linking algorithms may severely attenuate estimates of intergenerational income elasticities.

The variability in our estimates across datasets implies that it is difficult to diagnose how much linking assumptions matter for different research questions using different records. Nevertheless, our results suggest that reducing false matches and choosing methods that generate false matches more highly correlated with the truth are *crucial* for improving inferences with linked data—even when reducing Type I errors increases Type II errors.

An easy remedy when linking richer data is to use more information—especially continuous variables or those with many values (e.g., Social Security Numbers or exact dates of birth). In addition, higher quality information (e.g., administrative records rather than individual reports) will result in lower error rates than we document. For contexts with limited linking variables which are measured with error, systematic clerical review (e.g., LIFE-M) and genealogical methods (e.g., Early Indicators) generally attain lower error rates than machine algorithms. Because these methods are cost prohibitive for most projects, we draw on our

findings to recommend several easy-to-implement and lower-cost changes to current practice.

First, we recommend careful examination of a sample of links resulting from automated algorithms. Applying close scrutiny to a sample of links allows researchers to diagnose and potentially remedy systematic problems with machine-linking algorithms arising for specific records or in a particular historical context. In fact, many of the links coded as incorrect in clerical review are easy to identify as such. These cases can be used to improve machine-linking algorithms.

Second, caution is advised in linking phonetically cleaned names in deterministic algorithms or in commonly occurring name-age combinations. Phonetic cleaning tends to remove meaningful variation in names that allows algorithms to make better links. Eliminating commonly occurring name-age combinations, like Ferrie's (1996) approach of only linking uncommon names or Abramitzky, Boustan, and Eriksson (2014)'s robustness check using unique name-age combinations in a five-year window, substantially reduces the incidence of false matches. Together with reweighting, these restrictions achieve results in our case study that are statistically indistinguishable from hand-linked data. In contrast, weighting name-age ties equally by the inverse of their empirical frequency incorporates information from a large number of false links and results in substantial attenuation. In addition, researchers may incorporate more information in the linking process to break ties and distinguish true links from close alternatives. One such example in historical data is middle name or middle initial.

Third, using even a small sample of clerically reviewed data to train a machine-learning algorithm (or applying the results of another researcher's model based on similar training data) can improve the quality of linked samples. Notably, even when these machine-methods make *incorrect* links, the correlation of these links with the truth appears to be much higher than for other algorithms in our setting. These errors, therefore, have less impact on inference. An additional feature of some machine-linking algorithms is that they allow researchers to choose the importance of Type I and Type II errors, balancing the trade-off to fit a particular application. Although Feigenbaum (2016) and Abramitzky, Mill, and Pérez (2018) choose a specific penalty for Type I and Type II errors, different parameter choices can drive Type I error rates lower while linking much of the sample correctly.

A fourth strategy for reducing Type I errors is to *combine multiple methods* and use the intersection of the links across sets—a form of ensemble machine learning in the spirit of “bagging” or “boosting.” By construction, requiring links to be classified as such by more than one algorithm should tend to decrease match rates. But, to the extent that different methods make errors for different reasons, taking the set of common links helps avoid idiosyncratic reasons for errors. We illustrate the value of this approach in Figure 8 for our example of intergenerational elasticity, where we plot the Type I and Type II errors associated with the 131,071 possible combinations ($2^{17}-1$) of the 17 algorithms in this paper for each dataset. Overall, combining methods drives down Type I error rates and increases Type II error rates. For example, when using LIFE-M data, combining two methods like Ferrie (1996) and Feigenbaum (2016) drives the Type I error rate to 10 percent—a

substantial improvement over error rates for either method individually. Combining 12 methods achieves error rates as low as 6 percent, which is almost as precise as hand-linking.

Using combinations of methods may also improve inference. As shown in Figure 9A, across all combinations of methods, unweighted intergenerational elasticity estimates range from 0.11 to 0.24 (circle markers) and inverse-propensity-score reweighted estimates range from 0.13 to 0.24 (square markers). Based on an unweighted linear regression, a 10 percentage point increase in the Type I error rate tends to decrease the elasticity by 0.028, whereas this number is 0.015 in the weighted regression. Interestingly, in both weighted and unweighted cases, the mean over all combinations yields the value to the elasticity obtained in the hand-linked LIFE-M sample. As in our intergenerational elasticity example using single methods, Figure 9B shows that eliminating the incorrect links yields an average intergenerational elasticity nearly identical to the hand-linked sample (0.22) whereas the average intergenerational elasticity estimates for the incorrect links are less than half that value (0.096). These findings hold even when considering only the most prominent matching algorithms.

Finally, after limiting the role of linking errors, we recommend using multiple record features to assess and improve sample representativeness. Survey methods for constructing weights and allocating values are easy to implement and have well-documented properties. Making greater use of common record features such as name length or other socio-demographic information also allows researchers to use survey research methods or, as is more common in economics (and used in this paper), construct inverse-propensity weights to reduce sample selection and improve representativeness in observed characteristics (see Bailey, Cole, and Massey (2019) for a simple how-to guide). Ideally, this reweighting also improves balance in terms of unobserved characteristics, but there is no way to test this claim. A close examination of what is referred to as the common support assumption also informs researchers about where more time-intensive genealogical or clerical review methods may increase the representation of hard-to-link groups.

Many discussions of inference with linked data implicitly or explicitly assume that the match rate is just as important to inference as match quality. Our findings suggest that the quality of inferences with linked data may be improved by putting less emphasis on increasing sample sizes (which in our analysis tend to be associated with higher rates of false matches) and more emphasis on increasing the share of *correct* links. That is, social scientists wishing to conduct inference on linked data might increase the weight they place on decreasing Type I error rates over increasing sample sizes (decreasing Type II error rates). In the parlance of machine learning, this would involve weighting precision more heavily. Indeed, modern surveys such as the *Panel Survey of Income Dynamics* and the *National Longitudinal Survey* demonstrate that much can be learned from high-quality small samples with summary statistics and weights to describe and adjust for non-representativeness. Ultimately, increasing sample sizes for difficult-to-link subgroups (such as individuals with common names) will not likely be achieved *without* more data or higher quality record features to disambiguate similar records. More research to uncover data to describe the groups underrepresented in linked samples will serve both to broaden

knowledge about them and improve the ability of modern machine learning methods to link them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This project was generously supported by the National Science Foundation (SMA 1539228), the National Institute on Aging (R21 AG05691201), the University of Michigan Population Studies Center Small Grants (R24 HD041028), the Michigan Center for the Demography of Aging (MiCDA, P30 AG012846-21), the University of Michigan Associate Professor Fund, and the Michigan Institute on Research and Teaching in Economics (MITRE). We gratefully acknowledge the use of the Population Studies Center's services and facilities at the University of Michigan (R24 HD041028). During work on this project, Cole was supported by the NICHD (T32 HD0007339) as a UM Population Studies Center Trainee. We are grateful to Ran Abramitzky, Eytan Adar, George Alter, Jeremy Atack, Hoyt Bleakley, Leah Boustan, John Bound, Charlie Brown, Matias Cattaneo, William Collins, Dora Costa, Shari Eli, Katherine Eriksson, James Feigenbaum, Joseph Ferrie, Katie Genadek, Tim Guinane, Mary Hansen, Kris Inwood, Maggie Levenstein, Bhash Mazumder, Jorgen Modalsli, Adriana Lleras-Muney, Jared Murray, Joseph Price, Paul Rhode, Evan Roberts, Steve Ruggles, Mel Stephens and five anonymous referees for their many helpful suggestions. We thank Sarah Anderson, Garrett Anstreicher, Ali Doxey, Meizi Li, Shariq Mohammed, Paul Mohnen, Mike Ricks, and Hanna Zlotnick for their many contributions to the LIFE-M project.

VIII. References

- A'Hearn Brian, Baten Jörg, and Crayen Dorothee. 2009. "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital." *The Journal of Economic History* 69 (3):783–808. doi: 10.1017/S0022050709001120.
- Abowd John M. 2017. "Large-scale Data Linkage from Multiple Sources: Methodology and Research Challenges." NBER Summer Institute Methods Lecture.
- Abowd John M., and Vilhuber Lars. 2005. "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers." *Journal of Business and Economic Statistics* 23 (2):133–165.
- Abramitzky Ran, Boustan Leah, and Eriksson Katherine. 2013. "Have the Poor Always been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration." *Journal of Development Economics* 102:2–14. [PubMed: 26609192]
- Abramitzky Ran, Boustan Leah, and Eriksson Katherine. 2012a. "Web Appendix: Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration."
- Abramitzky Ran, Boustan Leah, and Eriksson Katherine. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122 (3):467–506.
- Abramitzky Ran, Boustan Leah Platt, and Eriksson Katherine. 2012b. "Europe's Tired, Poor, and Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102 (5):1832–1856.
- Abramitzky Ran, Mill Roy, and Pérez Santiago. 2018. "Linking Individuals Across Historical Sources: a Fully Automated Approach." National Bureau of Economic Research Working Paper Series No. 24324. doi: 10.3386/w24324.
- Aizer Anna, Eli Shari, Ferrie Joseph, and Lleras-Muney Adriana. 2016. "The Long Term Impact of Cash Transfers to Poor Families." *American Economic Review* 106 (4):935–971.
- Antonie Luiza, Inwood Kris, Lizotte Daniel J., and Ross J. Andrew. 2014. "Tracking People Over Time in 19th Century Canada for Longitudinal Analysis." *Machine Learning* 95 (1):129–146.
- Atack Jeremy. 2004. "A Nineteenth-Century Resource for Agricultural History Research in the Twenty-First Century." *Agricultural History* 78 (4):389–412.
- Atack Jeremy, Bateman Fred, and Gregson Mary Eschelbach. 1992. "Matchmaker, Matchmaker, Make Me a Match." *Historical Methods* 25 (2):53–65.

- Bailey Amy Kate, Tolnay Stewart E., Beck EM, and Laird Jennifer D.. 2011. "Targeting Lynch Victims: Social Marginality or Status Transgressions?" *American Sociological Review* 76 (3):412–436. [PubMed: 25067845]
- Bailey Martha, Cole Connor, and Massey Catherine G.. 2019. "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850-1930 IPUMS Linked Representative Historical Samples." University of Michigan Working Paper.
- Bailey Martha J. 2018. "Creating LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database." University of Michigan Working Paper.
- Bailey Martha J., and Cole Connor. 2019. "Autolink.ado." accessed 2019-06-13. 10.3886/E110164V1.
- Black Sandra E., and Devereux Paul J.. 2011. "Recent Developments in Intergenerational Mobility." In *Handbook of Labor Economics*, edited by David Card and Orley Ashenfelter, 1487–1541. Amsterdam: Elsevier.
- Bleakley Hoyt , and Ferrie Joseph. 2016. "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations." *Quarterly Journal of Economics* 131 (3):1455–1495.
- Bleakley Hoyt , and Ferrie Joseph P.. 2013. "Up from Poverty? The 1832 Cherokee Land Lottery and the Long-run Distribution of Wealth." NBER Working Paper 19175.
- Bleakley Hoyt, and Ferrie Joseph. 2017. "Land Opening on the Georgia Frontier and the Coase Theorem in the Short- and Long- Run." http://www-personal.umich.edu/~hoytb/Bleakley_Ferrie_Farmsize.pdf
- Bogue A 1963. *From Prairie to Corn Belt: Farming on the Illinois and Iowa Prairies in the Nineteenth Century*. Chicago: University of Chicago Press.
- Boustan Leah Platt , Kahn Matthew E., and Rhode Paul W.. 2012. "Moving to Higher Ground: Migration Response to Natural Disasters in the Early Twentieth Century." *American Economic Review: Papers and Proceedings* 102 (3):238–244.
- Boustan Leah Platt, and Collins William. 2014. "The Origins and Persistence of Black-White Differences in Women’s Labor Force Participation from the Civil War to the Present." In *Human Capital and History: The American Record*, edited by Boustan Leah, Frydman Carola and Margo Robert A.. Chicago, IL: University of Chicago Press.
- Buckles Karey S., and Hungerman Daniel M.. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics* 95 (3):711–724.
- Chetty Raj, Hendren Nathaniel, Kline Patrick, Saez Emmanuel, and Turner Nicholas. 2014. "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility." *American Economic Review* 104 (5):141–47. doi: 10.1257/aer.104.5.141.
- Christen Peter, and Goiser Karl. 2007. *Quality and Complexity Measures for Data Linkage and Deduplication*. Vol. 43.
- Collins William J., and Wanamaker Marianne H.. 2014. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics* 6 (1):220–252.
- Collins William J., and Wanamaker Marianne H.. 2015. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants." *Journal of Economic History* 75 (4):947–992.
- Collins William J., and Wanamaker Marianne H.. 2016. "Up from Slavery? African American Intergenerational Economic Mobility Since 1880." NBER Working Paper 23395.
- Costa Dora L., Heather DeSommer Eric Hanss, Roudiez Christopher, Wilson Sven E., and Yetter Noelle. 2017. "Union Army Veterans, All Grown Up." *Historical Methods* 50 (2):79–95. [PubMed: 28690347]
- Curti Merle. 1959. *The Making of an American Community: A Case Study of Democracy in a Frontier County*. Stanford: Stanford University Press.
- Dempster AP, Laird NM, and Rubin DB. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1):1–38.
- DiNardo John, Fortin Nicole M., and Lemieux Thomas. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64 (5):1001–1044.

- Duncan Otis Dudley. 1968. "Patterns of Occupational Mobility among Negro Men." *Demography* 5 (1):11–22.
- Eli Shari, Salisbury Laura, and Shertzer Allison. 2018. "Ideology and Migration after the American Civil War." *Journal of Economic History* 78 (3). doi: 10.1017/S0022050718000384.
- Eriksson Björn. 2016. "The Missing Links: Data Quality and Bias to Estimates of Social Mobility." www.fas.nus.edu.sg/cfpr/RC28/089.pdf. Accessed September 15, 2016.
- Feigenbaum James J. 2016. "A Machine Learning Approach to Census Record Linking." <http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf?m=1423080976>. Accessed March 28, 2016.
- Feigenbaum James J. 2018. "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940." *Economic Journal* 128 (612):F446–F481.
- Fellegi Ivan P., and Sunter Alan B.. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328):1183–1210. doi: 10.1080/01621459.1969.10501049.
- Ferrie Joseph P. 1996. "A New Sample of Males Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules." *Historical Methods* 29 (4):141–156.
- Goeken Ronald, Lynch Tom, Yu Na Lee Jacob Wellington, and Magnuson Diana. 2017.
- Goldin Claudia, and Katz Lawrence F., eds. 2011. *Mass secondary schooling and the state: the role of state compulsion in the high school movement*. Edited by Costa Dora and Lamoreaux Naomi R., *Understanding Long Run Economic Growth*. Chicago: University of Chicago Press.
- Greene William H. 2008. *Econometric Analysis*, 6th Edition. New York: Pearson.
- Guest AM 1987. "Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century." *Historical Methods* 20 (2):63–77.
- Hacker J. David. 2010. "Decennial Life Tables for the White Population of the United States, 1790-1900." *Historical Methods* 43 (3):45–79. [PubMed: 20563225]
- Hacker J. David. 2013. "New Estimates of Census Coverage in the United States, 1850-1930." *Social Science History* 37 (1):71–101.
- Haider Steven J., and Solon Gary. 2006. "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review* 96 (4):1308–1320.
- Heckman James J., Ichimura Hidehiko, Smith Jeff, and Todd Petra. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5):1017–1098.
- Hornbeck Richard, and Naidu Suresh. 2014. "When the Levee Breaks: Black Migration and Economic Development in the American South." *American Economic Review* 104 (3):963–990.
- Horowitz Joel L., and Manski Charles F.. 1995. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica* 63 (2):281–302. doi: 10.2307/2951627.
- Hout Michael, and Guest Avery M.. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States since 1850: Comment." *American Economic Review* 103 (5):2021–40. doi: 10.1257/aer.103.5.2021.
- Huber PJ . 1967. "The behavior of maximum likelihood estimates under nonstandard conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1:221–233.
- Jaro Matthew A. 1989. "Advances in Record Linking Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84 (406):414–420.
- Kim Gunky, and Chambers Raymond. 2012. "Regression Analysis under Probabilistic Multi-Linkage." *Statistica Neerlandica* 66 (1):64–79.
- Lahiri P, and Larsen Michael D.. 2005. "Regression Analysis with Linked Data." *Journal of the American Statistical Association* 100 (469):222–230.
- Malin J 1935. "The Turnover of Farm Population in Kansas." *Kansas Historical* 20:339–372.
- Margo Robert A. 2016. "Obama, Katrina, and the Persistence of Racial Inequality." *Journal of Economic History* 76 (2):301–341.
- Massey Catherine G. 2017. "Playing with matches: An assessment of accuracy in linked historical data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*:1–15. doi: 10.1080/01615440.2017.1288598.

- Mazumder Bhashkar. 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *Review of Economics and Statistics* 87 (2):235–255. doi: 10.1162/0034653053970249.
- Mazumder Bhashkar. 2015. "Estimating the Intergenerational Elasticity and Rank Association in the U.S.: Overcoming the Current Limitations of Tax Data." Federal Reserve Bank of Chicago Working Paper.
- Mazumder Bhashkar. 2018. "Intergenerational Mobility in the United States: What We Have Learned from the PSID." *Annals of the American Academy of Political and Social Science* 680 (1):213–234.
- Michelson M and Knoblock CA 2006. "Learning Blocking Schemes for Record Linkage." *Proceedings of the 21st National Conference on Artificial Intelligence AAAI-06*.
- Mill Roy. 2013. "Record Linkage across Historical Datasets." Stanford University Dissertation. <https://searchworks.stanford.edu/view/10232417>.
- Mill Roy, and Stein Luke C.. 2016. "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America." Last Modified October 24, 2016. <http://www.public.asu.edu/~lstein2/research/mill-stein-skincolor.pdf>.
- Modalsli Jorgen. 2017. "Intergenerational Mobility in Norway, 1865-2011." *The Scandanavian Journal of Economics* 119 (1):34–71. doi: 10.1111/sjoe.12196.
- Mullainathan Sendhil, and Spiess Jann. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2):87–106.
- National Office of Vital Statistics. 1948. *State and Regional Life Tables, 1939-1941*.
- Nix Emily, and Qian Nancy. 2015. "The Fluidity of Race: 'Passing' in the United States, 1880-1940." <http://www.nber.org/papers/w20828>.
- Ruggles Steven. 2006. "Linked Historical Censuses: A New Approach." *History and Computing* 14:213–224.
- Ruggles Steven. 2011. "Intergenerational Coredisence and Family Transitions in the United States, 1850-1880." *Journal of Marriage and the Family* 73 (1):138–148. [PubMed: 22039309]
- Ruggles Steven, Fitch Catherine A., and Roberts Evan. 2018. "Historical Census Record Linkage." *Annual Review of Sociology* 44.
- Ruggles Steven, Genadek Katie, Grover Josiah, and Sobek Matthew. 2015. *Integrated Public Use Microdata Series (Version 6.0) [Machine-Readable database]*. edited by University of Minnesota. Minneapolis: University of Minnesota.
- Saperstein Aliya, and Gullickson Aaron. 2013. "A Mulatto Escape Hatch? Examining Evidence of U.S. Racial and Social Mobility in the Jim Crow Era." *Demography* 50 (5):1921–1942. [PubMed: 23606347]
- Scheuren Fritz, and Winkler William E.. 1993. "Regression analysis of data files that are computer matched." *Survey methodology* 19 (1):39–58.
- Solon Gary. 1992. "Intergenerational Income Mobility in the United States." *American Economic Review* 82 (3):393–408.
- Solon Gary. 1999. "Intergenerational Mobility in the Labor Market." In *Handbook of Labor Economics*, edited by Ashenfelter Orley and Card David, 1761–1800. Amsterdam: Elsevier.
- Steckel R 1988. "Census Matching and Migration: A Research Strategy." *Historical Methods* 21:52–60.
- Stephens Melvin Jr., and Unayama Takashi. 2017.
- Thernstrom S 1964. *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Cambridge: Harvard University Press.
- Ward Zachary. 2019. "Intergenerational Mobility in American History: Accounting for Race and Measurement Error."
- West Kirsten K., and Robinson J. Gregory. 1999. "What do we know about the Undercount of Children?" U.S. Census Bureau Population Division Working Paper 39.
- White H . 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica* 48:817–830.

- Winkler William E. 2006. "Overview of Record Linkage and Current Research Directions." Research Report Series, Statistics 2006 (2).
- Wisselgren Maria J., Edvinsson Soren, Berggren Mats, and Larsson Maria. 2014. "Testing Methods of Record Linkage on Swedish Censuses." *Historical Methods* 47 (3):138–151.
- Xie Yue, and Killewald Alexandra. 2013. "Intergenerational Occupational Mobility in Britain and the U.S. since 1850: Comment." *American Economic Review* 103 (5):2003–2020.
- Zimmerman David J. 1992. "Regression Toward Mediocrity in Economic Stature." *American Economic Review* 82 (3):409–429.

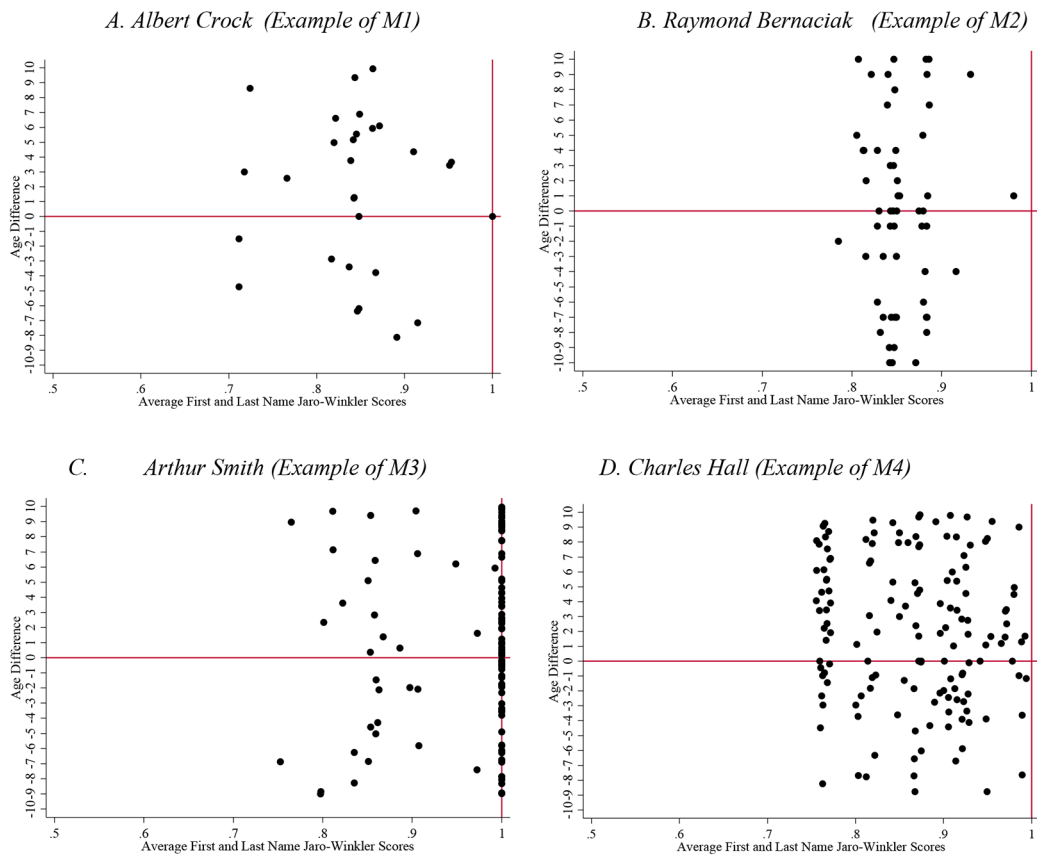


Figure 1.
Examples of Common Linking Problems in Historical Samples

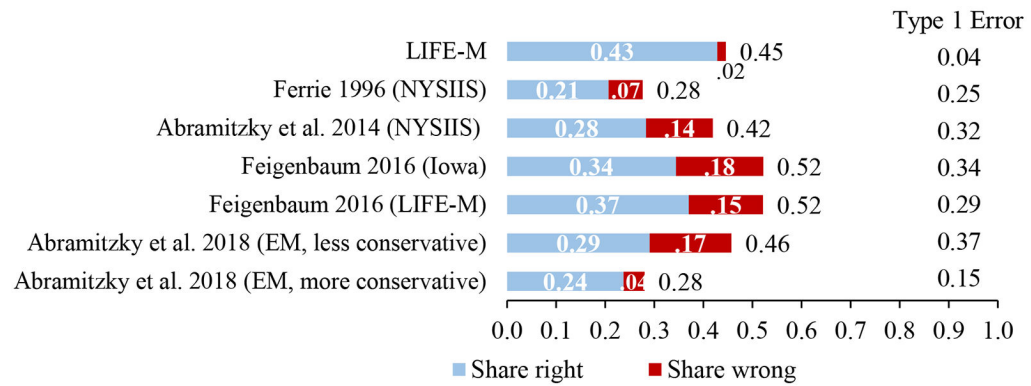


Figure 2. Match Rates and False Links for LIFE-M Hand-Linked Data and Selected Automated Linking Methods

Notes: The bars show the performance of different algorithms linking LIFE-M boys to the 1940 Census. See text for details and Table 1 for numerical estimates. EM stands for Abramitzky, Mill, and Pérez’s (2018) “Expectation maximization” method. More and less conservative refer to the two different, manually set thresholds recommended in the paper. Error rates for LIFE-M hand-links and other methods are computed based the police-review process described in the text.

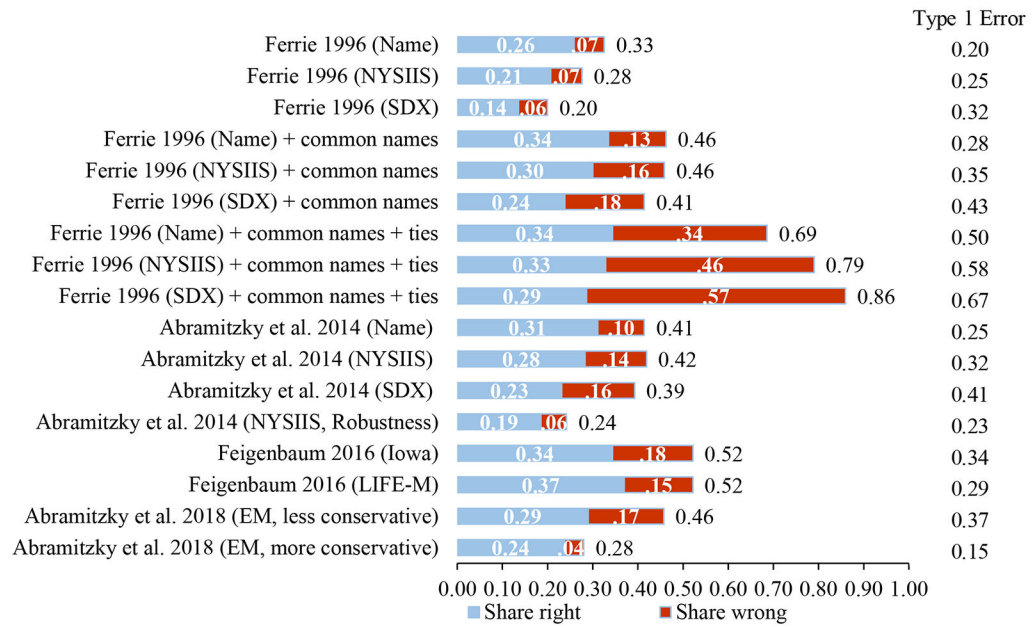


Figure 3. Match Rates and False Links for Common Variations on Automated Linking Methods
 Notes: See Figure 2 notes and text for details and Table 4 for numerical estimates.

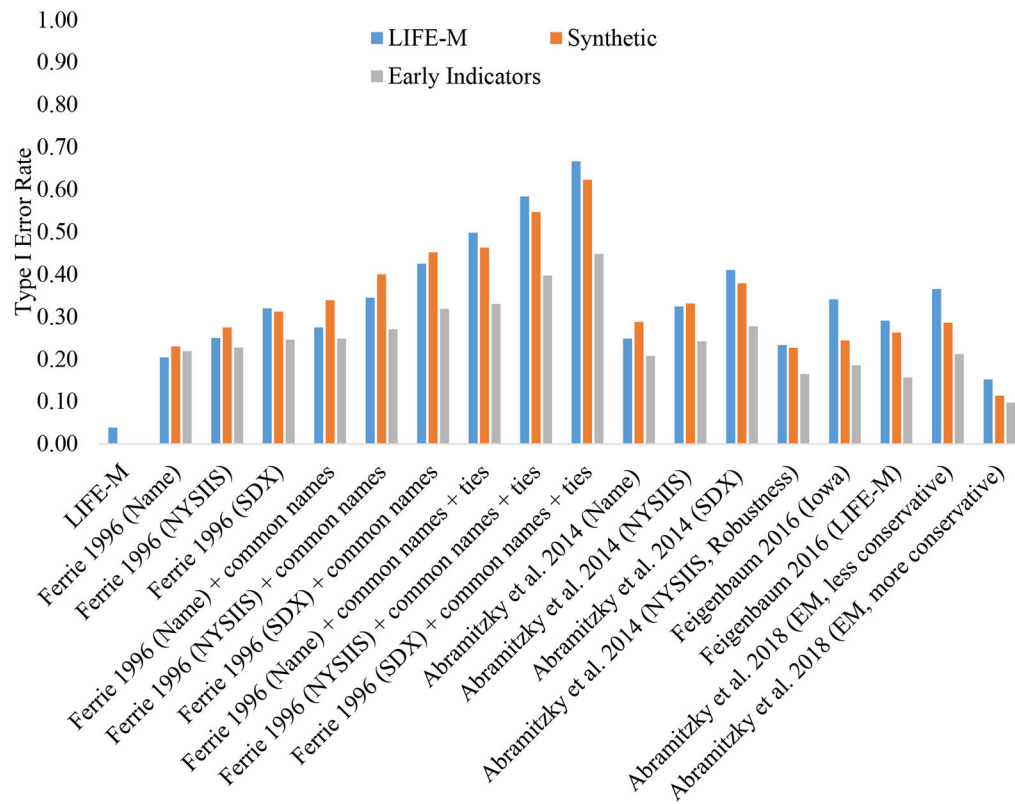


Figure 4. Share of Incorrect Links (Type I Error Rate) by Method and Dataset

Notes: See Figure 2 notes and Table 1 for numerical estimates.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

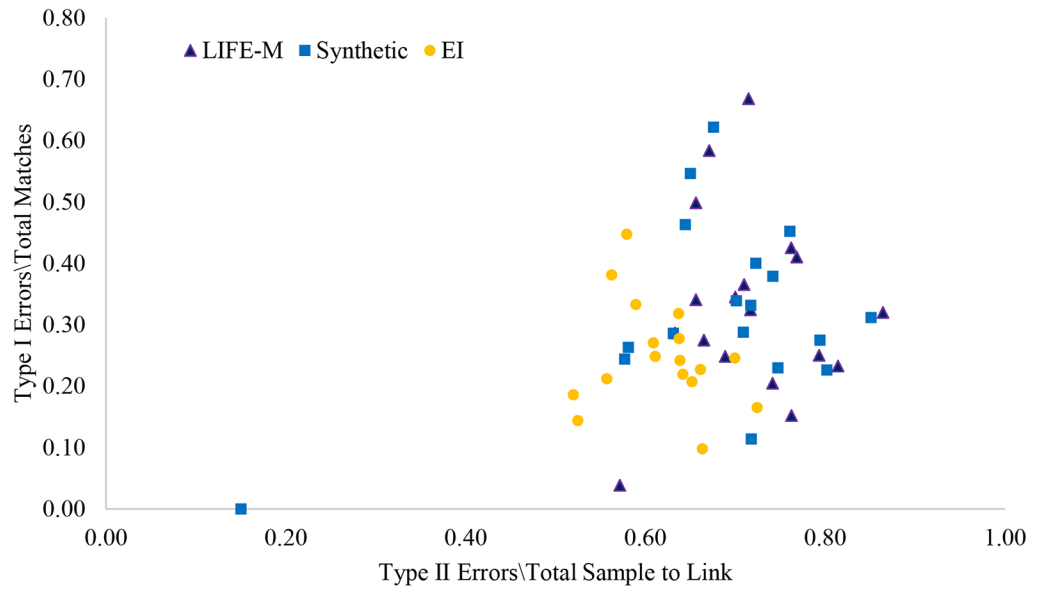


Figure 5. Type I vs. Type II Error Rates by Method and Dataset
Notes: Points plot Type I and Type II error rates using different algorithms and data in Table 2.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

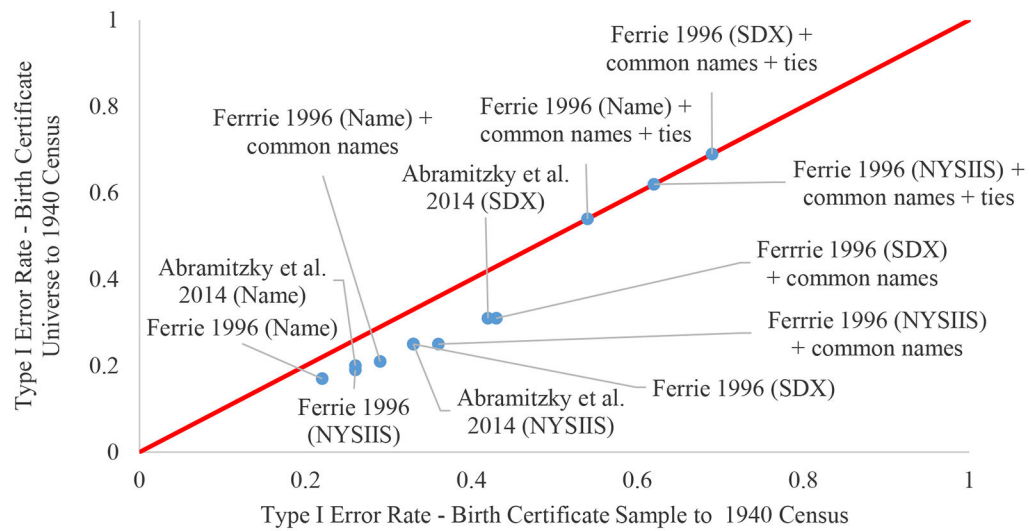
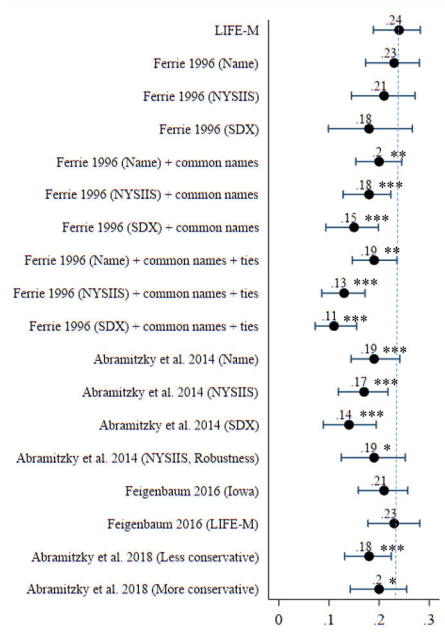


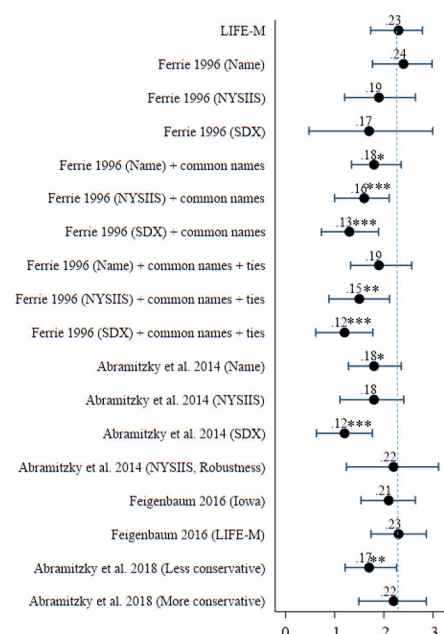
Figure 6. A Comparison of Method Performance in Sample-to-Population and Population-to-Population Linking

Notes: The y-axis plots the Type I error rate implied by linking birth certificates for *all* boys in the same cohorts as the Ohio and North Carolina LIFE-M sample to the 1940 Census using different automated methods. The x-axis plots the Type I error rate implied by linking the LIFE-M sample of boys in the Ohio and North Carolina birth certificates to the 1940 Census using different automated methods.

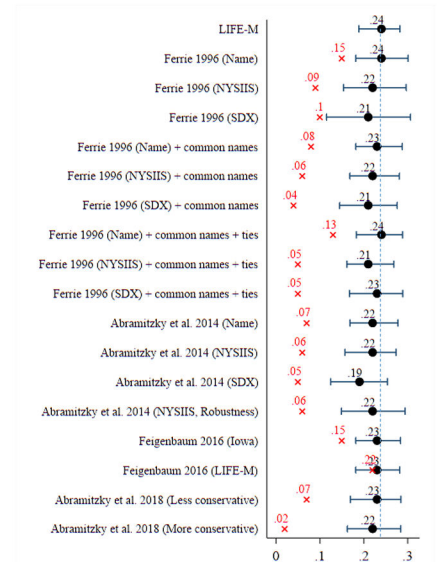
A. Unweighted Linked Samples



B. Inverse Propensity-Score Weighted Linked Samples



C. Separate Regressions for Imputed and Correct Links, Unweighted



D. Separate Regressions for Imputed and Correct Links, Weighted

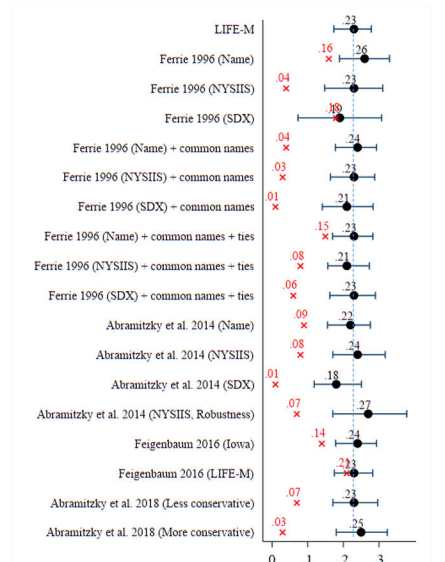


Figure 7. Intergenerational Income Elasticity Estimates

Notes: Differences in estimates reflect the incidence of Type I and Type II errors. The sample sizes of father-son pairs are lower than when matching sons only, because not all linked sons had income from wages and fathers who were also linked who also had income from wages. Sample sizes are 1,834 for LIFE-M, 1,313 for Ferrie 1996 (Name), 1,064 for Ferrie 1996 (NYSIIS), 708 for Ferrie 1996 (Soundex), 1,751 for Ferrie 1996 (Name) + common names, 1,702 for Ferrie 1996 (NYSIIS) + common names, 1,466 for Ferrie 1996 (SDX) + common names, 2,354 for Ferrie 1996 (Name) + common names + ties, 2,648 for Ferrie 1996 (NYSIIS) + common names + ties, 2,875 for Ferrie 1996 (SDX) + common

names + ties, 1,610 for Abramitzky et al. 2014 (Name), 1,600 for Abramitzky et al. 2014 (NYSIIS, Robustness), 1,412 for Abramitzky et al. 2014 (SDX), 999 for Abramitzky et al. 2014 (NYSIIS, Robustness) 1,955 for Feigenbaum 2016 (Iowa), 1,855 for Feigenbaum 2016 (LIFE-M), 1774 Abramitzky et al. 2018 (EM, less conservative), 1206 Abramitzky et al. 2018 (EM, more conservative). Weighted estimates reweight the linked sample of LIFE-M birth certificates so that they match a representative sample of birth certificates using the following characteristics: day of year measured from one to 365, polynomials in age, first and last name commonness indexes and the interaction of the two, a dummy variable for presence of siblings, polynomials in the number of siblings, polynomials in the length of child, mother, and father name, and state fixed effects. * indicates that the estimate is statistically different from the LIFE-M estimate at the 10-percent, ** at the 5-percent, and *** at the 1-percent levels.

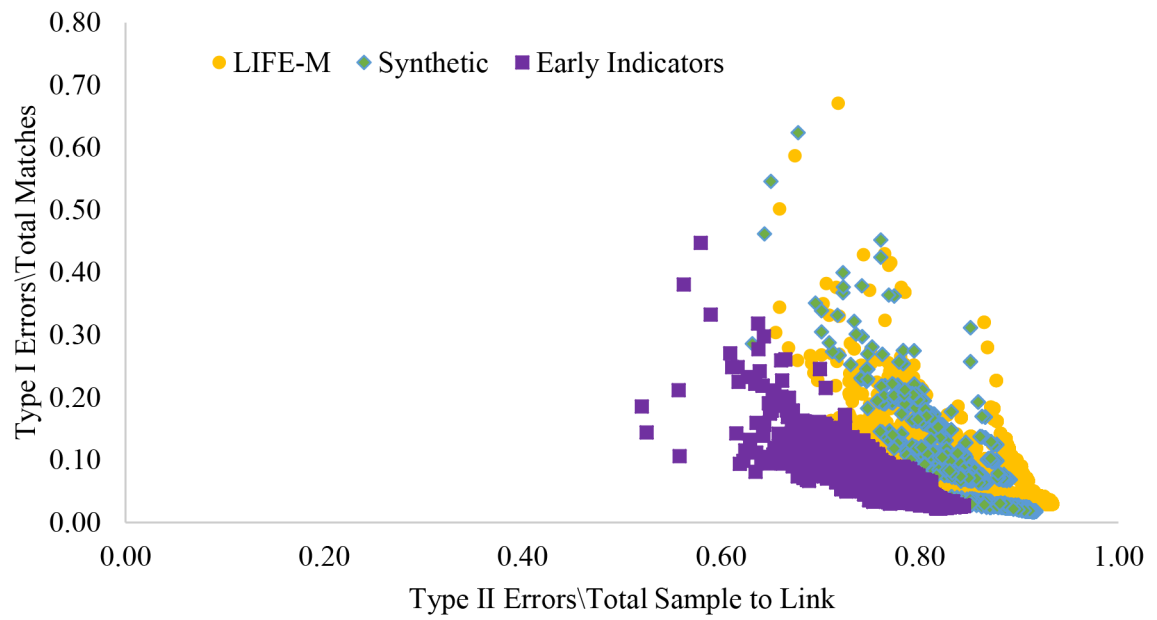
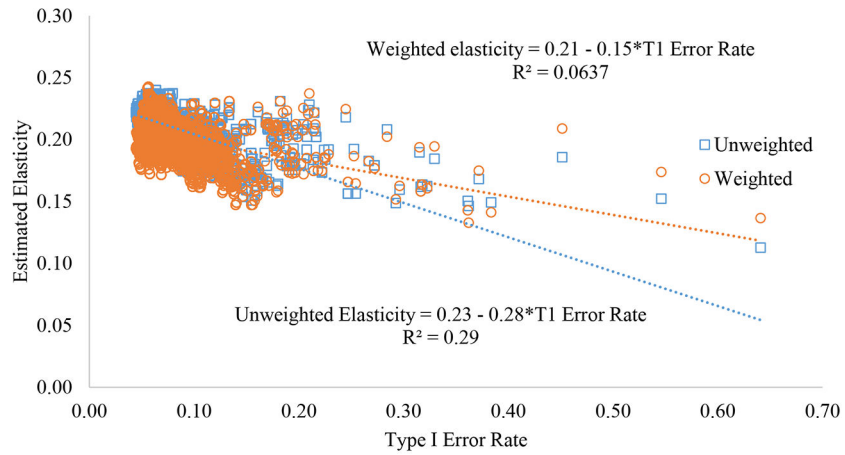


Figure 8. Type I vs. Type II Error Rates for Different Combinations of Methods and Dataset
Notes: Each point represents the Type I and Type II error rate for 131,071 different combinations of the 17 methods considered in this paper by dataset.

A. Unweighted and Weighted Intergenerational Income Elasticity Estimates



B. Intergenerational Income Elasticity Estimates using Correct versus Incorrect Links

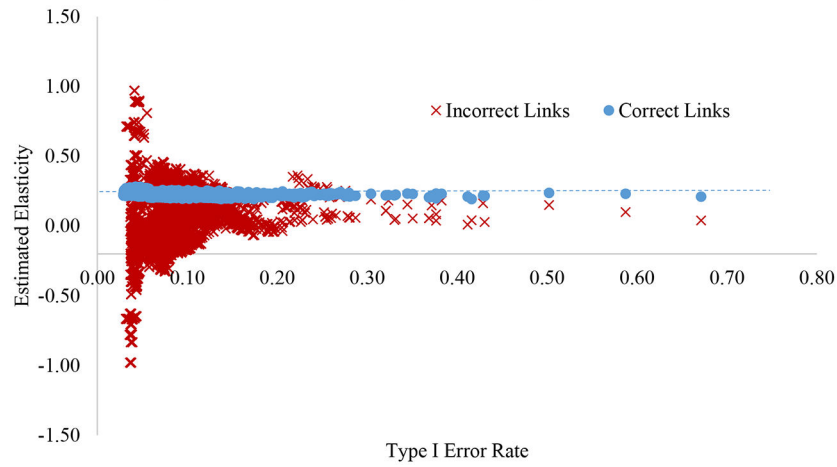


Figure 9. Intergenerational Income Elasticity Estimates across Method Combinations
 Notes: Each point represents intergenerational elasticity estimate plotted against the Type I error rate for one of the 131,071 different combinations of the 17 methods considered in this paper. Panel A pools all links and panel B plots estimates separately for correct and incorrect links. See also Figure 6 notes.

Table 1.
Summary of Performance of Prominent Linking Methods, by Algorithm and Dataset

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Hand-links or Synthetic Data	0.45	0.85	1.00	0.04	0.00	0.00	0.57	0.15	0.00
Ferrie 1996	0.28	0.28	0.44	0.25	0.27	0.23	0.79	0.79	0.66
Abramitzky et al. 2014	0.42	0.42	0.48	0.32	0.33	0.24	0.72	0.72	0.64
Feigenbaum 2016 (Iowa coefficients)	0.52	0.56	0.59	0.34	0.24	0.19	0.66	0.58	0.52
Feigenbaum 2016 (Estimated coefficients)	0.52	0.57	0.57	0.29	0.26	0.14	0.63	0.58	0.52
Abramitzky et al. 2018 (EM, less conservative)	0.46	0.52	0.56	0.37	0.29	0.21	0.71	0.63	0.56
Abramitzky et al. 2018 (EM, more conservative)	0.28	0.32	0.37	0.15	0.11	0.10	0.76	0.72	0.66

Notes: EI stands for the “Early Indicators” data. Each estimate in the table is for a match rate, Type I error rate, or Type II error rate as described in text. These estimates are depicted in graphical form in Figures 2, 3 and 4.

Table 2.

Representativeness of Links Created by Prominent Linking Methods, by Algorithm and Dataset

	LIFE-M	Synthetic Data	Early Indicators
Ferrie 1996 (NYSIIS)	445.9	277.5	38.2
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (NYSIIS)	457.0	387.2	12.7
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.24</i>
Feigenbaum 2016 (Iowa coef.)	195.7	34.9	50.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Estimated coef.)	334.9	62.2	44.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2018 (EM, less conservative)	788.3	485.0	46.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2018 (EM, more conservative)	1350.0	673.0	51.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Observations	42,869	42,869	1,785

Notes: Each estimate is a heteroskedasticity-robust Wald-test from a separate regression of a binary dependent variable (=1 for linked record) for samples described in the text. Relevant p-values are reported in italics. The covariates included in the LIFE-M sample and synthetic data are age, number of siblings, length of names of individuals and parents, fraction of siblings with misspelled parents' names, and an observation coming from Ohio. The covariates included in the Early Indicators data are age, currently married, foreign born, day of birth by year, literacy, length of first and last names, and foreign born status of parents. These sample sizes are slightly smaller due to missing values. See appendices for full regression results.

Table 3.

Representativeness of False Links Created from Prominent Linking Methods, by Algorithm and Dataset

	LIFE-M	Synthetic Data	Early Indicators
Ferrie 1996 (NYSIIS)	242.9	35.1	26.2
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (NYSIIS)	500.9	64.3	39.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Iowa coef.)	1806.0	448.0	38.9
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Estimated coef.)	1559.0	802.0	19.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.03</i>
Abramitzky et al. 2018 (EM, less conservative)	559.3	112.9	43.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2018 (EM, more conservative)	139.8	17.4	18.3
<i>p-value</i>	<i>0.00</i>	<i>0.03</i>	<i>0.05</i>

Notes: Each estimate is a heteroskedasticity-robust Wald-test from a separate regression of a binary dependent variable (=1 for falsely linked record) for samples described in the text. Relevant p-values are reported in italics. The covariates included in the LIFE-M sample and synthetic data are age, number of siblings, length of names of individuals and parents, fraction of siblings with misspelled parents' names, and an observation coming from Ohio. The covariates included in the Early Indicators data are age, currently married, foreign born, day of birth by year, literacy, length of first and last names, and foreign born status of parents. See appendices for full regression results.

Table 4.

Summary of Algorithm Performance When Varying Assumptions

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Ferrie 1996 (Name)	0.33	0.33	0.46	0.20	0.23	0.22	0.74	0.75	0.64
Ferrie 1996 (NYSIIS)	0.28	0.28	0.44	0.25	0.27	0.23	0.79	0.79	0.66
Ferrie 1996 (SDX)	0.20	0.22	0.40	0.32	0.31	0.25	0.86	0.85	0.70
Ferrie 1996 (Name) + common names	0.46	0.45	0.52	0.28	0.34	0.25	0.66	0.70	0.61
Ferrie 1996 (NYSIIS) + common names	0.46	0.46	0.54	0.35	0.40	0.27	0.70	0.72	0.61
Ferrie 1996 (SDX) + common names	0.41	0.44	0.53	0.43	0.45	0.32	0.76	0.76	0.64
Ferrie 1996 (Name) + common names + ties	0.69	0.66	0.62	0.50	0.46	0.33	0.66	0.64	0.59
Ferrie 1996 (NYSIIS) + common names + ties	0.79	0.77	0.71	0.58	0.55	0.40	0.67	0.65	0.57
Ferrie 1996 (SDX) + common names + ties	0.86	0.86	0.76	0.67	0.62	0.45	0.71	0.68	0.58
Abramitzky et al. 2014 (Name)	0.41	0.41	0.44	0.25	0.29	0.21	0.69	0.71	0.65
Abramitzky et al. 2014 (NYSIIS)	0.42	0.42	0.48	0.32	0.33	0.24	0.72	0.72	0.64
Abramitzky et al. 2014 (SDX)	0.39	0.42	0.50	0.41	0.38	0.28	0.77	0.74	0.64
Abramitzky et al. 2014 (NYSIIS, Robustness)	0.24	0.26	0.33	0.23	0.23	0.17	0.81	0.80	0.72
Feigenbaum 2016 (Iowa coef.)	0.52	0.56	0.59	0.34	0.24	0.19	0.66	0.58	0.52
Feigenbaum 2016 (LIFE-M coef.)	0.52	0.57	0.57	0.29	0.26	0.16	0.63	0.58	0.52
Abramitzky et al. 2018 (EM, less conservative)	0.46	0.52	0.56	0.37	0.29	0.21	0.71	0.63	0.56
Abramitzky et al. 2018 (EM, more conservative)	0.28	0.32	0.37	0.15	0.11	0.10	0.76	0.72	0.66

Notes: See Table 1 notes and text for details.

Table 5.

Representativeness of Links When Varying Algorithm Assumptions

	LIFE-M	Synthetic Data	Early Indicators
Ferrie 1996 (Name)	688.3	390.7	47.5
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (NYSIIS)	445.9	277.5	38.2
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (SDX)	130.6	71.1	57.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (Name) + common names	412.3	378.1	35.5
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (NYSIIS) + common names	402.6	447	16.0
<i>p-value</i>	<i>0.00</i>	<i>0.10</i>	<i>0.10</i>
Ferrie 1996 (SDX) + common names	208.8	310.6	25.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (Name) + common names + exact ties	178.8	452.1	75.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (NYSIIS) + common names + exact ties	148.2	363.9	69.9
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (SDX) + common names + exact ties	104.7	174.7	43.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (Name)	454.6	271.9	32.3
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (NYSIIS)	457.0	387.2	12.7
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.24</i>
Abramitzky et al. 2014 (SDX)	255.7	257.8	17.1
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.07</i>
Abramitzky et al. 2014 (NYSIIS, Robustness)	568.6	397.7	31.2
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Iowa coef.)	195.7	34.9	50.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Estimated coef.)	334.9	62.2	44.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2018 (EM, less conservative)	788.3	485.0	46.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2018 (EM, more conservative)	1350.0	673.0	51.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Observations	42,869	42,869	1,785

Notes: See Table 2 notes and text for details.

Table 6.

Representativeness of False Links When Varying Algorithm Assumptions

	LIFE-M	Synthetic Data	Early Indicators
Ferrie 1996 (Name)	468.3	79.3	45.8
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (NYSIIS)	242.9	35.1	26.2
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (SDX)	81.5	0.9	17.5
<i>p-value</i>	<i>0.00</i>	<i>0.99</i>	<i>0.06</i>
Ferrie 1996 (Name) + common names	772.1	115.3	48.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (NYSIIS) + common names	429.0	64.4	39.2
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (SDX) + common names	157.7	17.4	32.4
<i>p-value</i>	<i>0.00</i>	<i>0.03</i>	<i>0.00</i>
Ferrie 1996 (Name) + common names + exact ties	1859.0	466.2	60.1
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (NYSIIS) + common names + exact ties	1163.0	249.6	92.3
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Ferrie 1996 (SDX) + common names + exact ties	457.8	55.4	61.7
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (Name)	744.4	100.2	54.3
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (NYSIIS)	500.9	64.3	39.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (SDX)	223.2	41.7	28.6
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2014 (NYSIIS, Robustness)	239	24.3	32.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Iowa coef.)	1806.0	448	38.9
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Feigenbaum 2016 (Estimated coef.)	1559.0	802	19.4
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.03</i>
Abramitzky et al. 2018 (EM, less conservative)	559.3	112.9	43.0
<i>p-value</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>
Abramitzky et al. 2018 (EM, more conservative)	139.8	17.4	18.3
<i>p-value</i>	<i>0.00</i>	<i>0.03</i>	<i>0.05</i>

Notes: See Table 3 notes and text for details.

Table 7.

How Middle Initials Could Reduce Errors in Linking in LIFE-M Data

	(1) Table 4 Match Rate	(2) Table 4 Type Error Rate	(3) Share Matches with Middle Initials for Both Records	(4) Share of (3) with Discordant Middle Initials	(5) Type I Error Rate in (4)	(6) Revised Match Rate	(7) Revised Type I Error Rate
Ferrie 1996 (Name)	0.33	0.20	0.28	0.26	0.90	0.30	0.15
Ferrie 1996 (NYSIIS)	0.28	0.25	0.26	0.27	0.91	0.26	0.20
Ferrie 1996 (SDX)	0.20	0.32	0.24	0.30	0.93	0.19	0.27
Ferrie 1996 (Name) + common names	0.46	0.28	0.29	0.35	0.94	0.42	0.20
Ferrie 1996 (NYSIIS) + common names	0.46	0.35	0.27	0.37	0.95	0.41	0.28
Ferrie 1996 (SDX) + common names	0.41	0.43	0.26	0.41	0.96	0.37	0.36
Ferrie 1996 (Name) + common names + exact ties	0.69	0.50	0.30	0.44	0.97	0.60	0.43
Ferrie 1996 (NYSIIS) + common names + exact ties	0.79	0.58	0.29	0.50	0.98	0.68	0.52
Ferrie 1996 (SDX) + common names + exact ties	0.86	0.67	0.28	0.57	0.98	0.72	0.61
Abramitzky et al. 2014 (Name)	0.41	0.25	0.30	0.31	0.94	0.38	0.18
Abramitzky et al. 2014 (NYSIIS)	0.42	0.32	0.27	0.33	0.94	0.38	0.26
Abramitzky et al. 2014 (SDX)	0.39	0.41	0.27	0.39	0.96	0.35	0.35
Abramitzky et al. 2014 (NYSIIS, Robustness)	0.24	0.23	0.26	0.26	0.89	0.23	0.18
Feigenbaum 2016 (Iowa)	0.52	0.34	0.31	0.23	0.93	0.48	0.30
Feigenbaum 2016 (LIFEM)	0.52	0.29	0.33	0.23	0.93	0.48	0.24
Abramitzky et al. 2018 (EM, less conservative)	0.46	0.37	0.29	0.34	0.95	0.41	0.30
Abramitzky et al. 2018 (EM, more conservative)	0.28	0.15	0.29	0.20	0.87	0.26	0.11

Notes: This table uses the LIFE-M data to evaluate changes in algorithm match rates and Type I error rates with the addition of middle initials. Column 7 computes match rates after dropping links with discordant middle initials. Column 8 computes revised Type I error rates by dropping links with discordant middle initials. See text for details.

Table 8.

How Using Race Could Reduce Errors in Linking in LIFE-M Data

	(1) Table 4 Match Rate	(2) Table 4 Type I Error Rate	(3) Share Matches - 1940 Race Variables Different than LIFE-M	(4) Revised match Rate	(5) Revised Type I Error Rate
Ferrie 1996 (Name)	0.33	0.20	0.00	0.33	0.20
Ferrie 1996 (NYSIIS)	0.28	0.25	0.01	0.28	0.25
Ferrie 1996 (SDX)	0.20	0.32	0.01	0.20	0.31
Ferrie 1996 (Name) + common names	0.46	0.28	0.01	0.46	0.27
Ferrie 1996 (NYSIIS) + common names	0.46	0.35	0.01	0.46	0.34
Ferrie 1996 (SDX) + common names	0.41	0.43	0.02	0.41	0.42
Ferrie 1996 (Name) + common names + exact ties	0.69	0.50	0.01	0.69	0.49
Ferrie 1996 (NYSIIS) + common names + exact ties	0.79	0.58	0.03	0.79	0.58
Ferrie 1996 (SDX) + common names + exact ties	0.86	0.67	0.06	0.86	0.66
Abramitzky et al. 2014 (Name)	0.41	0.25	0.01	0.41	0.25
Abramitzky et al. 2014 (NYSIIS)	0.42	0.32	0.01	0.42	0.32
Abramitzky et al. 2014 (SDX)	0.39	0.41	0.02	0.39	0.40
Abramitzky et al. 2014 (NYSIIS, Robustness)	0.24	0.23	0.01	0.24	0.23
Feigenbaum 2016 (Iowa)	0.52	0.34	0.01	0.52	0.34
Feigenbaum 2016 (LIFEM)	0.52	0.29	0.00	0.52	0.29
Abramitzky et al. 2018 (EM, less conservative)	0.46	0.37	0.02	0.46	0.36
Abramitzky et al. 2018 (EM, more conservative)	0.28	0.15	0.00	0.28	0.15

Notes: This table uses the LIFE-M to evaluate changes in linking rates with the addition of race. Column 4 computes match rates after dropping links with discordant race. Column 5 computes revised Type I error rates by dropping links with discordant race. See text for details.