NBER WORKING PAPER SERIES

HOW WELL DO AUTOMATED METHODS PERFORM IN HISTORICAL SAMPLES? EVIDENCE FROM NEW GROUND TRUTH

Martha Bailey
Connor Cole
Morgan Henderson
Catherine Massey

Working Paper 24019
http://www.nber.org/papers/w24019

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2017

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 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 Erickson, James Feigenbaum, Joseph Ferrie, Katie Genadek, Tim Guinanne, Mary Hansen, Kris Inwood, Maggie Levenstein, Bhash Mazumder, Jorgen Modalsli, Adriana Lleras-Muney, Jared Murray, Joseph Price, Paul Rhode, Evan Roberts, Steve Ruggles, and Mel Stephens for their many helpful suggestions. We thank Sarah Anderson, Garrett Anstreicher, Ali Doxey, Meizi Li, Mike Ricks, and Hanna Zlotnick for their many contributions to the LIFE-M project. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth

Martha Bailey, Connor Cole, Morgan Henderson, and Catherine Massey

NBER Working Paper No. 24019
November 2017
JEL No. J62,N0

ABSTRACT

New large-scale data linking projects are revolutionizing empirical social science. Outside of selected samples and tightly restricted data enclaves, little is known about the quality of these "big data" or how the methods used to create them shape inferences. This paper evaluates the performance of commonly used automated record-linking algorithms in three high quality historical U.S. samples. Our findings show that (1) no method (including hand linking) consistently produces samples representative of the linkable population; (2) automated linking tends to produce very high rates of false matches, averaging around one third of links across datasets and methods; and (3) false links are systematically (though differently) related to baseline sample characteristics. A final exercise demonstrates the importance of these findings for inferences using linked data. For a common set of records, we show that algorithm assumptions can attenuate estimates of intergenerational income elasticities by almost 50 percent. Although differences in these findings across samples and methods caution against the generalizability of specific error rates, common patterns across multiple datasets offer broad lessons for improving current linking practice.

Martha Bailey
University of Michigan
Department of Economics
611 Tappan Street
207 Lorch Hall
Ann Arbor, MI 48109-1220
and NBER
baileymj@umich.edu

Connor Cole
University of Michigan
Department of Economics
611 Tappan Street
Ann Arbor, MI 48109
colecp@umich.edu

Morgan Henderson
University of Michigan
Department of Economics
611 Tappan St
Ann Arbor, MI 48103
morghend@umich.edu

Catherine Massey Institute
for Social Research
University of Michigan 426
Thompson Street
Ann Arbor, MI 48104
cgmassey@umich.edu

New large-scale linked data are revolutionizing empirical social science. A variety of current and proposed projects are linking national surveys, administrative data, and research samples to recently digitized historical records, such as the full-count 1880 (Ruggles 2006, Ruggles et al. 2015) and 1940 U.S. Censuses (the first to ask about education and wage income) and newly available administrative sources.[1] The resulting "big data" have the potential to break new ground on old questions and open entirely novel areas of inquiry.

Machine-linkage is critical to these projects, and new innovations in machine linkage are being introduced at record pace. Outside of protected administrative and proprietary data enclaves, however, little is known about how machine linkage influences data quality or social science inferences. The dearth of knowledge is especially acute for the U.S., where the lack of "ground truth" has limited the study of false (Type I errors) and missed matches (Type II errors).[2] Although some diagnostic exercises are suggestive, they rely on selected and often non-U.S. samples (Christen and Goiser 2007, Goeken et al. 2016) or rich administrative data unavailable to most researchers (Scheuren and Winkler 1993, Winkler 2006, Abowd 2017, Massey 2017). The resulting uncertainty about the quality of most linked samples limits both these samples' contribution to social science and the development of innovations in machine-linking methods.

This paper combines three different high quality U.S. samples in a novel evaluation of the impact of different linking assumptions on data quality and inference. Data availability leads us to focus on linking to historical U.S. census samples, but our conclusions apply broadly to contexts where linking variables are limited and measured with error.[3] Different ground-truth samples allow us to examine the robustness of our findings for various record types and contexts. The Longitudinal Intergenerational Family Electronic Micro-database's (LIFE-M) sample of birth certificates linked to the 1940 Census provides a perspective from the early 20th Century (Bailey et al. 2016), and the genealogically linked sample of Union Army veterans from the Early Indicators Project (Costa et al. 2017). The Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS) of the 1880 Census (Ruggles et al. 2015) cover the late 19th Century. In addition, we validate our conclusions by building a synthetic ground truth (for which we know true links), subjecting potential false links to *additional* independent human reviews, and using variables not used in the linking

[1] Many on-going initiatives link the 1940 Census to other datasets. The Census Bureau plans to link it 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 (PSID) and the Health and Retirement Survey (HRS) 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 et al. 2016). Supplementing these public infrastructure projects, entrepreneurial researchers have also combined large datasets. See, for example, Abramitzky et al. (2012), Abramitzky et al. (2013), Abramitzky et al. (2014), Boustan et al. (2012), Mill (2013), Hornbeck and Naidu (2014), Mill and Stein (2016), Aizer et al. (2016), Bleakley and Ferrie (2013, 2014, 2016), Nix and Qian (2015), Collins and Wanamaker (2016), and Eli et al. (2016). This paper discusses some of these approaches.

[2] "Ground truth" is defined as data obtained by direct measurement and is often used to refer to true data links.

[3] Unlike contemporary ground truth data, historical data are public and contain identifiable information necessary for linkage. Historical data allow us to be fully transparent about our samples and techniques.

process to characterize errors. We focus our analysis on four machine-linking algorithms, which represent the diversity of methods in current practice. In addition, we evaluate the impact of assumptions about phonetic name-cleaning (Soundex and NYSIIS), the treatment of common names, and procedures for dealing with multiple exact matches (such as probability weighting and random selection) on link quality.

Our findings highlight the important effects of linking algorithms on data quality. First, 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 method and record type. Second, automated linking algorithms result in high rates of incorrect matches ranging from 13 to 69 percent. Averaging across samples and methods, we find that 32 percent of links generated by automated methods are likely incorrect. Third, different algorithms induce different correlations between false links with baseline sample characteristics and the true link.

Although findings vary across datasets, common patterns reveal the importance of different linking assumptions. Phonetic name cleaning universally increases false link rates by an average of 23 percent and, in some cases, more than 60 percent. Trying to link more common names tends to worsen false link rates by 30 percent, when averaged across methods and datasets. Methods to break ties through random selection or simple probability weighting increase instances of false links in samples by an average of 40 percent and, in some cases, more than 75 percent.

A final case study considers the importance of linking algorithms on estimates of intergenerational mobility in the 1940s. After characterizing the theoretical implications of false links and missing links using a within-between decomposition framework, we show that—for the same sample of records—algorithm assumptions can attenuate intergenerational income elasticity estimates by almost 50 percent. Eliminating false matches reduces attenuation and renders intergenerational mobility elasticities—in very different samples—statistically indistinguishable.

These findings suggest that machine-linking algorithms *per se*—and the errors they introduce—can dramatically affect researchers' conclusions. Different linking assumptions induce different kinds of non-representativeness and measurement error, which may have difficult-to-predict consequences for inference. Although the specific error rates and attenuation factors found here are not broadly generalizable, we conclude by formulating recommendations for best practice that draw on common patterns across datasets and algorithms. Importantly, our findings caution against increasing linkage rates which tend to come at the cost of increasing linkage errors. Using low-cost ways to diagnose and purge false links together with weights to increase the sample representativeness holds great promise for improving large-scale record linkage.

I. AN OVERVIEW OF METHODOLOGY IN RECORD LINKAGE IN HISTORICAL DATA

Recent developments in computational speed, data availability, and probabilistic linking techniques have expanded the possibilities of record linkage. However, limitations in U.S. historical data often constrain what is possible, both relative to contemporary administrative sources and data for other countries. This section describes some limitations of historical U.S. records and illustrates how they can affect linking.

A. A Typical Historical Linking Problem

Several types of common errors in historical records limit researchers' ability to link them. Names may vary over time because Census enumerators may have misspelled respondent names, the household respondent may have reported incorrectly, or the individual may have changed their name (perhaps using a middle name or nickname in place of the given name). For example, Goeken et al. (2016) documents that in two enumerations of St. Louis in the 1880 Census, nearly 46 percent of first names are not exact matches, and 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). A similar problem relates to the common practice of rounding age to the nearest multiple of 5 (or "age heaping") (A'Hearn et al. 2009, Hacker 2013) and frequent mis-enumeration of birth place (Goeken et al. 2016), which make individuals hard to track over time.

Adding to these recording errors, the digitization of hand-written manuscripts adds another layer of uncertainty. Our comparison of two independently digitized versions of the 1940 Census by Ancestry.com and FamilySearch.org suggests that 25 percent of records have *different* transcriptions of last name in the two datasets. Consequently, linking individuals across historical datasets relies upon a handful of potentially mis-measured characteristics.

To illustrate common challenges associated with matching historical datasets, consider the exercise of linking birth certificates for boys to the 1940 U.S. Census. To minimize concerns about selection bias and non-representativeness of the linked sample, researchers typically use *time-invariant* characteristics to match records (Ruggles 2006).[4] When linking birth certificates to the 1940 Census, these characteristics include first

[4]The earliest linking 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 person within the same region (the block) in the next census (Malin 1935, Curti 1959, Bogue 1963, Thernstrom 1964, Guest 1987). This strategy missed individuals who relocated or changed names and, therefore, generated unrepresentative samples. The creation of state population indexes allowed researchers to draw a random sample of households with children at least 10 years old in a historical census. They then searched for the same household in the previous census using the birth state of the child to narrow the search (Schaefer 1985, Steckel 1988). This technique, therefore, could find individuals who moved between the census years, but it was similarly limited by geographic mobility between birth and first enumeration in the census. It also restricted the sample of linked households to those with children surviving to age 10.

name, last name, age, birth state, and sex—effectively four characteristics because sex contains little more information than first name.[5] One advantage of using birth certificates (relative to census) is that they measure age and birth state with much less error than the decennial censuses, because they also collect the *exact* date of birth and are recorded in a particular state. Issues with name transcription and digitization, however, will affect both birth certificates and Census, and errors in birth place and age will affect the 1940 Census records.

After limiting candidate matches to those with the same birth state and sex (as is common in the literature), the standard linking problem can be captured as two-dimensional scatter plots as shown in Figure 1. The x-axis captures the similarity between the name on the birth record and the names of its candidate links in the 1940 Census as measured by the Jaro-Winkler similarity score, a string-distance metric equal to 1 if the names are identical.[6] The y-axis captures the difference in the age in 1940 implied by the birth certificate (which contains exact date of birth) and the reported age in the 1940 Census. A perfect match occurs when the age difference is zero. In this two-dimensional space, link candidates fall into one of four categories:

(M1) A perfect (1,0), unique match in terms of name and age similarity (shown in Figure 1A).

(M2) A single, similar match is slightly different in terms of age, name, or both (shown in Figure 1B).

(M3) Many *perfect* (1,0) matches, leading to problems with match disambiguation (shown in Figure 1C).

(M4) Multiple similar matches are slightly different in terms of age, name, or both (shown in Figure 1D).

Historical linking algorithms generally consider M1 candidates as matches. However, methods differ in their treatment of candidates in the M2, M3, and M4 categories. A common approach for dealing with candidates in the M2 category is the use of Soundex or NYSIIS phonetic codes while allowing for differences in age. Phonetic string cleaning is designed to account for orthographic differences that could lead a true match to be missed, such as minor spelling differences, name Anglicization, and transcription errors. Soundex, for example, was developed in the early 20th century to help create Census links. It groups similar sounding names like “Smith,” “Smyth” and “Smythe” to the same code (in this example, S530). NYSIIS, the acronym for the New York State Identification and Intelligence System, was developed in 1970 as an improvement to the Soundex algorithm. NYSIIS transforms names like “Wilhem” and “William” to WALAN.[7] To account for

[5] Race is not available on birth certificates, but this has little effect on this paper’s results. Appendix A3 shows that using race would reduce error rates by at most 1 percentage point. Matching 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. The possible birthplaces in Norway number more than 500 municipalities for a population of under 2 million, whereas birthplace is limited to 48 states and foreign countries for 132 million residents in the 1940 U.S. Census.

[6] Jaro-Winkler similarity score adapts the Jaro (1989) string distance, the minimum number of single-character transpositions required to change one string into another, to up-weight differences earlier in the string. See Winkler (2006) for an overview.

[7] Phonetic name cleaning was included in a public probabilistic matching software, called PC Matchmaker, to link to Census samples in a DOS environment (Atack et al. 1992). PC Matchmaker coded names using Soundex or NYSIIS phonetic codes and allowed for user-specified blocking and weighting schemes. This software was then used to create a linked sample between the agricultural and population censuses between 1850 and 1880 (Atack 2004).

differences in age, researchers typically search within age ranges such as ±3 or ±5 years. Approaches to deal with ties in categories M3 or M4 include random selection among equally likely (tied) candidates (Nix and Qian 2015) or an equal probability weighting of tied candidates. The next sections discuss how the commonly used automated linking methods address challenges posed by M2, M3, and M4 cases in historical data.

B. Ferrie (1996)

Ferrie's (1996) path-breaking approach linked men in the 1850 U.S. Census to men who were 10 years and older in the 1860 U.S. Census. Ferrie (1996) began by selecting a sample of uncommon names from the 1850 Census, which reduces problems with record ties (for example, the uncommon name sample excludes names like "John Smith" or records in M3 and M4 above).[8] To identify likely links among those that were similar and to correct for minor orthographic differences (category M2 above), Ferrie's method transformed last names using NYSIIS codes and also truncated the untransformed first name after the fourth letter. Ferrie then linked his sample to the 1860 Census, and eliminated candidate links that were not born in the same state and not living with the same family. Ferrie kept all candidate links within a ±5 year difference in age and, if more than two links remained, Ferrie chose the link with the smallest age difference.[9] This process produced a linked sample of 4,938 men between—9 percent of the male population in 1850, 19 percent of the population of men with uncommon names. Ferrie's approach has varied in his more recent work, and he has used a smaller age ranges, different phonetic name cleaning (such as SPEDIS in SAS), or required matches of birth places. More than 20 years later, Ferrie's original approach has become the foundation for much of the historical linking literature, which is why we consider it in this paper.

C. Abramitzky, Boustan, and Erikson (2012 and 2014)

Abramitzky et al.'s "Iterative Method" automates 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. As summarized in a detailed web appendix (Abramitzky et al. 2012a), Abramitzky et al. (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 the NYSIIS, and look for exact, unique matches. If this method does not identify an exact, unique link, it searches for an exact match within a ±1 year band as a second step and, if they do not find a unique link, they repeat this process up to a ±2 year band around the reported birth year. The Iterative Method does not link a record if there is no exact name match in NYSIIS, if

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

the best link is tied (has multiple exact candidate matches, as in the M3 case above), or if no candidate falls within the maximum allowable age band. Abramitzky et al. (2012b) ultimately linked 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 et al. (2014) uses the Iterative Method to link men ages 18-35 with unique age-name combinations from the 1900 U.S. Census to the 1910 and 1920 censuses, achieving a match rate of 16 percent for native-born and 12 percent for foreign-born men. The authors generously provided the most recent version of their code for our analysis, which has been used for record linkage in a number of high profile papers.[10] Appendix A includes the results for Abramitzky et al.'s (2014) additional robustness check of limiting the sample to individuals unique in name within a five-year age band in both censuses.

D. Feigenbaum (2016) and IPUMS (2015)

A common feature of Ferrie (1996) and Abramitzky et al. (2014) is that they search along the line segment at phonetically cleaned name similarity = 1 in Figure 1 for a *unique* match in NYSIIS-cleaned name with a minimum age difference. New methods in probabilistic linking extend these methods by identifying the empirical importance of name and age differences using human decisions as a basis. 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 when simultaneously considering *both* the age and name differences.

IPUMS models these trade-offs using a Support Vector Machine (SVM) to create IPUMS-LRS of the 1880 U.S. Census. IPUMS-LRS consists of linkages between the 1850-1930 one-percent samples and the 1880 Census (Ruggles et al. 2015) using clerically reviewed data to train a SVM that, following a few tuning parameter choices, classifies all potential links as true or false (Goeken et al. 2011). Illustrating the conservative nature of their approach, they produce final match rates of 12.2 percent for native-born whites, 3 percent for foreign-born whites, and 6.4 percent for African Americans for the 1870-1880 links.[11]

---

[10] Abramitzky et al. (2012) use the Iterative Method to show that the returns to migration were relatively low, and Abramitzky et al. (2013) use the Iterative Method show that immigrants from Norway to the U.S. were negatively selected based upon parents' wealth. To study migrant assimilation for 16 sending countries, Abramitzky et al. (2014) make use of the Iterative Method to link men between the ages of 18 and 35 across the 1900, 1910, and 1920 U.S. Censuses, producing a sample of 20,225 immigrant and 1,650 native-born men for a match rate of 12 percent and 16 percent, respectively. In another application of the Iterative Method, Boustan et al. (2012) produce a sample of men ages 30 to 40 years old linked from the 1920 to the 1930 U.S. Censuses to study migration responses to natural disasters, for a match rate of 24 percent across the 1920 and 1930 Censuses. Hornbeck and Naidu (2014) use the same linked sample restricted to 1920 and 1930 links to study black migration out of the Mississippi Delta after the 1927 Flood. Aizer et al. (2016) use an updated Ferrie method to link male children of applicants to the mother's pension programs between 1911 and 1935 to the Social Security Death Master File to study the effect of cash transfers on later-life outcomes. They find that mother's pension programs increased longevity of recipient children by one year relative to children of mothers who were denied benefits.

[11] Researchers 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).

In a similar spirit, Feigenbaum (2016) employs a probit regression to quantify the empirical importance of name and age differences in addition to other record features. Using clerically reviewed data, he estimates the effect of different record characteristics on the probability that the records are a match. These record characteristics include name Jaro-Winkler similarity scores, differences in age, indicators for Soundex matches of first and last name, indicators for matches in letters or names, and whether truncated parts of the first and last name match. 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 than the second-best candidate's match probability (if a second-best candidate exists). Feigenbaum (2016) then uses this method to link a large number of men in the 1915 Iowa Census to the 1940 Census, achieving a match rate of 57 percent.[12]

E. Breaking Ties: Probability Weighting and Random Selection

Another challenge in historical data linkage is the treatment of multiple identical matches on name and age (e.g., cases like M3 in Figure 1C and M4 in Figure 1D). Nix and Qian (2015) propose *randomly* selecting among identical candidate matches to choose one from these exact ties. This results in a draw of *one* of the candidate matches with probability $1/J_r$, where J_r is the number of ties for a primary record, r. Statisticians have, instead, proposed weighting by the conditional probability that the match is correct (Scheuren and Winkler 1993, Lahiri and Larsen 2005). In Census linking with limited information, multiple *exact* ties arise often. Simple probability weighting uses information from these identical ties and weights each by $1/J_r$.[13] The results for simple probability weighting and random selection are the same in expectation.[14] For brevity, we

[12] We focus on Feigenbaum (2016) in this analysis because it is for U.S. data, transparent, and easy to replicate. Many other researchers have also incorporated probabilistic and machine learning. Mill (2013) and Mill and Stein (2016) use a method that employs string comparators and scoring of matches and maximum likelihood estimation. Similar to the IPUMS-LRS, Antonie et al. (2014) and Wisselgren et al. (2014) use a text-string comparator and estimate probability scores using truth data and machine learning techniques. Antonie et al. (2014) describe the linkage system they developed to link across historical Canadian census data and their application of this system to linkage of men from the 1871 Canadian Census to the 1881 Canadian Census. Their linkage rates range from 17.5 percent (Quebec) to 25.5 percent (New Brunswick). Wisselgren et al. (2014) use an approach similar to Antonie et al. (2014) and the MPC to link the 1890 and 1900 Swedish Censuses. Depending on their treatment of names, they achieve match rates ranging from 18 percent to 70 percent for men and 24 percent to 66 percent for women (Wisselgren et al. 2014, pp. 148). They use parish records as "truth" data to show 3 percent Type I error rates in their links. Preliminary work by Eriksson (2016) shows the error rate of the Swedish linked data increases by as much as 24 percent if linked using county of birth rather than parish of birth. Another approach is the use of Ancestry.com's probabilistic search algorithm to link records. Bailey et al. (2011) link records of lynching to the 1900 to 1930 U.S. Censuses to determine which community characteristics were associated with lynching. Collins and Wanamaker (2015) enter information on name, age, and place of birth for black men ages 0 to 40 resident in Southern states in the 1910 Census into Ancestry to search for individuals who uniquely match these criteria. They match 19.4 percent of individuals to a unique person in 1930. Eli et al. (2016) use probabilistic matching to link military records from the War and Treasury Departments to study the effect of the Civil War on migration decisions for those living on the border of Union and Confederate states.

[13] This 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.

[14] To see this, 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, …, 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

summarize these results under Nix and Qian (2015) in this paper's presentation.

In summary, existing linking methods involve multiple, interrelated procedures that may improve match rates and accuracy and also address problems with tie breaking. Which set of assumptions should researchers use in different contexts? What are the implications of using different phonetic name cleaning algorithms or weighting exact ties? This paper seeks to answer these questions by presenting a systematic comparison of methods in different records and periods.

II. METRICS OF AUTOMATED METHOD PERFORMANCE

This paper evaluates the performance of different linking methods for linking U.S. historical data: Ferrie (1996); the Iterative Method (Abramitzky et al. 2014); and regression-based prediction (Feigenbaum 2016). 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 as described.[15] We also present the results for an adaptation of Ferrie (1996) that relaxes the common name restriction (i.e., includes names with more than 10 exact matches). 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 our ground truth samples (called "estimated coef."; see Appendix A for details and coefficient estimates). To implement Abramitzky et al. (2014), we use the code provided by the authors (see Appendix A). Finally, we implement tie breaking by random selection (Nix and Qian 2015) for links generated by Ferrie's method without the uncommon name restriction. We additionally present results using different phonetic name cleaning algorithms for each of these algorithms.

We examine the performance of each algorithm in several high-quality historical samples, including the LIFE-M sample of birth certificates linked to the 1940 Census (Bailey et al. 2016), a genealogically linked sample of Union Army veterans from the Early Indicators Project (Costa et al. 2017), and the IPUMS-LRS of the 1880 Census (Ruggles et al. 2015). Because these ground truth data may also contain errors, we validate our conclusions by building a synthetic ground truth (which we create with full knowledge of true links), subjecting potential false links to additional independent human reviews, and using variables not used in the linking process to characterize errors. We use four main criteria to measure performance:

(1) Match rate: We calculate the match rate as the share of records that were successfully matched

---

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.

[15] Unlike 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.

from the original sample. Even for perfect matching, this rate is expected to be less than 100 percent due to emigration and mortality. Notwithstanding, comparisons across methods are still valid, as emigration and mortality affect each sample and, therefore, all of our methods equally.

(2) Representativeness: We compare the characteristics of the linked samples to the characteristics of the unlinked sample using a heteroskedasticity-robust Wald test of the overall significance of the covariates in a multivariate, linear probability model with Huber-White standard errors (Huber 1967, White 1980). The exact variables in the regression vary across datasets. Under the null hypothesis that the covariates are jointly unrelated to successful linkage, the Wald statistic should be statistically insignificant.[16] This approach provides a straightforward, single summary measure, and its regression coefficients describe the extent and subgroups that are under-represented.

Although these first two measures are the most common statistics reported for linked samples, they 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 (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, rendering the data useless for making population inferences. We, therefore, use two more criteria to assess link performance (Abowd and Vilhuber 2005, Kim and Chambers 2012):

(3) False link rate (also called the Type I error rate)[17]: We compare links for each automated method to a measure of the truth. In the synthetic data, we know the true link and take disagreements as errors. We also take the hand-linked Early Indicators data as the truth. For the LIFE-M and IPUMS data, we include an independent assessment of link quality. For the LIFE-M data, we stage a "police line-up." In the line-up, two independent reviewers see the LIFE-M link, the automated method link, and a number of close candidate links, which means that links from the ground truth and the automated method have an equal shot at being chosen. This process also allows reviewers to identify errors in LIFE-M. For the IPUMS-LRS, we use concordance of parents' birth places in the two linked Censuses—information that IPUMS did not use to link the data—as an independent

[16] We 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.
[17] Computer scientists focus on 1-T1 error rate, which is called "precision."

measure of quality.[18] The false positive rate for each method is defined as the share of links that disagree with the synthetic ground truth and Early Indicators' links, the links that reviewers reject in the police line-up (LIFE-M), or the links with discordant parents' birth places (IPUMS-LRS). This analysis may understate the true Type I error rate if the ground truths are incorrect or the ground truth and the automated method agree on an incorrect link. We also look for evidence of systematic measurement error by examining the correlation of false links with baseline characteristics.

(4) False negative rate (also called the Type II error rate)[19]: This metric captures the fraction of true links that are not found, or 1 – Match Rate*(1 – Type I Error Rate).

Because a central focus of a growing literature is linking to the newly available 1940 Census, we begin our analysis with the LIFE-M sample to the 1940 Census.

III. METHOD PERFORMANCE IN THE LIFE-M DATA

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 linked LIFE-M sample consists of 45,442 boys born from 1881 to 1940 and 25,352 born in North Carolina and 19,090 born in Ohio.

The LIFE-M sample of boys was then linked to the 1940 full-count U.S. Census using a semi-automated process, making use of both computer programming and human input. After cleaning and standardizing the data, we use bi-gram matching on name and age similarity within birth state to generate a set of candidate links (Wasi 2014).[20] Each candidate match is 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

[18] We acknowledge that birthplace discordance is a noisy measure of Type I errors, but Bailey et al. (2017) shows that birthplace disagreement is correlated with being an incorrect link. Requiring disagreement in both parents' birthplaces may understate the true rate of Type I errors, because 18.7 percent of IPUMS links have at least one disagreement in parent's birthplace. On the other hand, parental birth place discordance may overstate errors if enumerators filled out this variable incorrectly (e.g., indicated state of residence instead of birth), if there are errors in imputed family relationships within a household, if an adult misremembers the birthplace of their parents, or if there were changes in household composition that would lead to a different father's birthplace being listed on a child's record (Goeken et al. 2016). Although parental birthplace discordance is noisy measure of Type I errors, limiting the sample to non-discordant links reduces the Type I error rate. For this reason, IPUMS-LRS also used disagreement in parent birthplaces as part of its matching algorithm when matching the 1880 to the 1900, 1910, 1920 and 1930 census samples.

[19] Computer science focus on a similar statistic, "recall." This is defined as the number of true links found by the algorithm divided by number of linkable observations. It is different from 1-T2 error rate in that the false links are excluded from the denominator. In historical contexts, it is difficult to know how many individuals have a true links in the data, because mortality rates, emigration, and Census under-enumeration are unknown.

[20] A bi-gram algorithm compares two strings by selecting all combinations of two consecutive characters within each string.

the truth. In cases where the two trainers disagree, the records are *re*-reviewed by three new trainers to resolve these discrepancies.[21] This automated system maintains quality by randomly assigning these discrepancies cases among the 15 to 30 trainers who are employed at any time, so it is difficult for trainers to coordinate with peers. In addition, random audits provide feedback to trainers about the accuracy of their decisions, and weekly meetings help trainers achieve consistent and accurate matches.

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 543 of the 19,090 Ohio boys, 241 of which had been linked by LIFE-M. The BYU team had no knowledge of LIFE-M's links. Its links, however, 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 "Tree." (FamilySearch.org Tree links are created by genealogists and users of FamilySearch.org, which are also independent of the LIFE-M process.) For 1,043 birth certificates linked to the 1940 Census by LIFE-M and FamilySearch.org users, the LIFE-M links agreed with FamilySearch.org users 96.7 percent of the time. Taking genealogical linking as the gold standard implies that LIFE-M's false link rate is between 3.3 and 6.6 percent. However, not all cases where LIFE-M differs are necessarily incorrect. 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 as described in section II.

A. Match Rates

Table 1 and Figure 2A describe the match rates for the LIFE-M clerical review and for each automated algorithm. Match rates are computed as the share of the baseline sample of 45,442 boys who were successfully matched to the 1940 complete count Census. LIFE-M matched 43 percent of the baseline sample, whereas Ferrie's (1996) method matches between 19 and 31 percent of baseline sample, depending on the phonetic name cleaning. His match rates fall with phonetic name cleaning, because this cleaning creates more common name strings, leading the algorithm to discard more links. As expected, relaxing the uncommon name sample restriction (labeled "Ferrie 1996 + common names") results in higher match rates, ranging from 39 to 43 percent. Abramitzky et al.'s Iterative Method is very similar to Ferrie with common names and yields comparable match rates, ranging from 37 to 39 percent. Feigenbaum's (2016) regression-based machine learning method matches approximately 47 percent of the baseline sample when using Iowa coefficients and

[21] "Data trainers" participate in a rigorous orientation process when 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.

43 percent when we estimate the coefficients using a random sample of the LIFE-M links. As intended, the Nix and Qian (2015) method of randomly choosing a match for records with ties substantially increases the match rate to between 67 to 82 percent, highlighting how important tie breaking is for historical linking.

Three findings stand out. First, match rates across methods using the birth certificates are higher than in published studies. For instance, the Ferrie (1996) method matches between 19 and 31 percent of our sample versus his published figure of 9 percent of all men between 1850 and 1860. Similarly, the Abramitzky et al. (2014) algorithm links 37 to 39 percent of the LIFE-M sample, whereas it links only 29 percent in the 2012b paper and 16 percent of native-born men in their 2014 paper. This result likely reflects the fact that birth certificates have better information than data used in many U.S. Census-linking contexts: (1) they contain a complete and correct *full* name, often including middle names omitted in the Census; (2) they record the exact date of birth rather than age in years;[22] and (3) state of birth reflects where the birth certificate was recorded and so should have less error than Census reports.[23] In addition, the LIFE-M boys are on average 24 years old in the 1940 Census, so mortality and outmigration are likely lower than in other contexts.

Second, phonetic name cleaning may increase or decrease match rates, depending upon how it affects ties. Although phonetic cleaning may correct for orthographic and transcription errors (tending to increase match rates), it may also remove meaningful spelling variation from names and increase ties (tending to decrease match rates). This interaction is important for the Ferrie (1996) method. This method's uncommon name restriction (i.e., allowing fewer than 10 potential matches in the 1940 Census regardless of age) results in the match rate falling from 31 to 26 (NYSIIS), to 19 percent (Soundex). Because the Nix and Qian (2015) method does not require unique matches, phonetic name cleaning unambiguously increases match rates.

B. Representativeness of the Linked Samples

We next compare the representativeness of the linked records. 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 the individual's exact date of birth[24] and the number of siblings in the family. In addition, we include as covariates 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. Finally, we include the share of family records with a misspelled mother's or

[22] Massey (2017) shows that decreasing the noise in age results in higher match rates and lower Type I error rates.
[23]Goeken et al. (2016) compare the first and second enumeration of St. Louis in the 1880 Census, conducted five months apart. They find that own birthplace disagreed for 8.4 percent of records, that father's birthplace disagreed for 17.9 percent of records, and mother's birthplace disagreed for 18.7 percent of records.
[24] Exact day of birth (1-366, due to leap year) is as close to a continuous measure as we can get in historical records, and season of birth is strongly correlated with socio-economic in modern data (Buckles and Hungerman 2013).

father's name, which we expect to be negatively correlated with years of schooling and income (Aizer et al. 2016). Column 1 of Table 2 reports the heteroskedasticity-robust Wald-statistic (p-value beneath) by method for the LIFE-M data. Under the null that the links are representative of the underlying population, the Wald-statistic would be small and the p-value large.

Consistent with findings in other papers (Abramitzky et al. 2012, Abramitzky et al. 2014, Collins and Wanamaker 2015), p-values indicate that we can reject representativeness (i.e., we reject that baseline covariates are unrelated to whether a record is linked) for all linking methods. Regression estimates in Appendix C show that the LIFE-M ground truth is less likely to link boys with higher incidence of misspelled father's last name, but more likely to link boys with a longer mother's name. All methods are more likely to link children with longer names, indicating that linked records may come from more affluent families. At the same time, most methods tend to link children with more siblings, indicating that their families may have been less affluent. These findings in part reflect the fact that the linkable sample—those surviving and not emigrating before the Census—is not randomly selected. Still, differences in regression coefficients across samples likely imply that the linking algorithm itself has consequences for representativeness of the data as well. This finding suggests caution when generalizing findings from the linked data to the population or comparing results across samples linked through different matching algorithms.

C. False Links (Type I Errors)

Figure 2A and Table 1also describes the share of links that are incorrect using the "police line-up" method described in section II. False links are presented in two ways. First, the *share of the entire sample* determined to be wrong for each method is displayed in red in Figure 2A. For slightly less than 1 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, these reversals are rare. Second, the column on the far right in Figure 2A and in column 5 in Table 1 presents that *share of links* that are wrong (false positive rate). This number is computed by dividing the share of the total sample that is incorrect by the match rate. Because the LIFE-M match rate is 43 percent, this implies a Type I error rate of 1.0 percent (approximately 0.005/0.43). The implications of measurement error are closely 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 much higher, ranging from 22 percent for Ferrie (1996) to 69 percent for Nix and Qian-Soundex (2015). In his 1996 paper, Ferrie

used NYSIIS, which is associated with a 26 percent false link rate in the LIFE-M sample.[25] Although 20 years old, Ferrie's (1996) method of selecting uncommon names achieves the lowest Type I error rate at 22 percent. Including more common names worsens the share of false links. Holding the rest of Ferrie's (1996) algorithm constant, using common names increases the false match rate by 8 to 13 percentage points, resulting in Type I error rates ranging from 30 to 44 percent. This pattern explains the higher error rates, ranging from 26 to 43 percent rates of all links for Abramitzky et al. (2014) that includes more common names. Underscoring this point is the fact that their robustness check, which eliminates more common names, reduces error rates to levels similar to Ferrie (1996) (see Appendix A3). Feigenbaum's (2016) supervised, regression-based machine learning model produces a Type I error rate of 35 percent when using the Iowa coefficients and decreases to 30 percent when estimated using LIFE-M data. In the case of Nix and Qian (2015), the share of links that are wrong ranges from 54 to 69 percent. As noted, a high share of false positives for Nix and Qian (2015) is not surprising, because this method randomly selects one link from sets of *identical* ties on name and age. For instance, if 10 identical matches are found, nine of these will be incorrect. Although this method ensures a high match rate, the random tie breaking is unlikely to identify the *correct* match in the majority of cases.

Importantly, the incidence of Type I error universally increases with the use of phonetic name cleaning in our sample. Across methods, using NYSIIS rather than uncleaned names increases Type I error rates by an average of 22 percent and as much as 31 percent. Using Soundex rather than uncleaned names increases Type I error rates by an average of 48 percent and as much as 65 percent. This increase occurs because, in addition to orthographic and transcription errors, phonetic codes remove *meaningful* spelling variation from names. For example, both Soundex and NYSIIS would code "Meyer" and "Moore" as the same name, whereas reviewers tend to treat these as different names. By interacting with the requirements of many automated methods that names (or a phonetically cleaned names) match *exactly*, phonetic name cleaning causes some automated methods to classify records as links that humans would reject.

Table 3 shows that false links appear to be systematically related to baseline sample characteristics. Our test regresses a binary dependent variable equal to one for false links and zero for correct links on the characteristics used in the representativeness analysis: the individual's exact date of birth, the number of siblings, the number of characters in the infant, mother and father's names; and misspelled names (a common measure of illiteracy). Column 1 of Table 3 reports the Wald-statistic by method for the LIFE-M data. For all methods, including LIFE-M's clerical review, p-values of the Wald test are lower than 0.05. These results

[25] This error rate 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 a 19-23 percent false positive rate.

suggest that false links are not random with respect to sample characteristics may complicate inference by introducing systematic measurement error into analyses. See Appendix D for the full set of regression results.

D. Missed Links (Type II Errors)

Our last analysis compares the incidence of Type I errors to Type II errors which we plot in Figure 2B and present in column 9 of Table 1. The horizontal axis shows that Type II error rates tend to be high, ranging from a low of 58 percent for LIFE-M to 88 percent for Ferrie (1996)-Soundex (with the uncommon name restriction). Of course, the *level* of these errors 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.[26] Because mortality, emigration, and under-enumeration affect all methods equally, these factors do not influence our comparisons across methods.

To facilitate an overall comparison of methods, Figure 2B plots Type I error rates on the vertical axis against Type II error rates on the horizontal axis. A linking method that linked all individuals correctly would locate at (0.15, 0), missing only the individuals who are unlinkable. As linking methods become less conservative (and, therefore, have lower Type II error rates), we generally expect the share of incorrectly classified links (Type I errors) to fall. This expectation predicts a negative relationship between Type II and Type I errors. The plot, however, exhibits a positive relationship with a slope coefficient of 1. As methods have increased match rates, they have also tended to worsen Type I errors *and increase* Type II errors.

In summary, the LIFE-M data suggest that a large share of links used for inference in historical settings appear erroneous. Methods aimed to increase match rates such as the use of more common names, phonetic name cleaning, and methods for tie-breaking increase the share of false matches and, counter-intuitively, increase the share of missed links. Adding to these problems, false links appear to be systematically related to baseline sample characteristics, which suggest they might introduce difficult-to-correct measurement error into analyses. Next, we examine the robustness of these findings in other high quality linked datasets.

IV. AUTOMATED METHODS PERFORMANCE IN ALTERNATIVE DATASETS

Variability in record quality and the skill and care with which microfilmed data are transcribed should lead researchers to expect different results in different datasets. Consequently, the LIFE-M results may not generalize to other datasets or periods. Moreover, high Type I error rates in the LIFE-M data could result from

[26] Based on life tables from 1939 to 1941, we calculate that 8.27 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.

reviewers (trained by the LIFE-M project) making decisions more likely to favor the LIFE-M links.

To characterize the robustness of our findings in different contexts, we use three alternative high-quality samples. First, we *simulate* a synthetic ground truth dataset, so that we know the true link objectively. Because our objective truth is not influenced by human reviewers at all, this validation exercise helps interpret the exercise with the LIFE-M data. Second, we use the Oldest Old sample from the Early Indicators Project that was linked by genealogists and is known to be highly accurate. Third, we use 1850, 1860, 1870, and 1900 IPUMS-LRS. This section describes each of these samples and then presents the results.

A. Descriptions of the Synthetic Ground Truth, Early Indicators, and IPUMS-LRS Samples

We construct the synthetic ground truth in in two steps. First, we take all of our Ohio and North Carolina born boys, randomly drop 10 percent to reflect mortality and emigration and drop another 5 percent to reflect under-enumeration (see footnote 26). Using the LIFE-M names 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 (Hacker 2010, Hacker 2013, Goeken et al. 2016).[27] The result 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 retain regional naming conventions and have similar demographic and economic characteristics and, consequently, provide good candidate links. We choose the size of our random sample of boys from neighboring states so that our total set of candidate matches (the noisy truth − 15 percent for emigration and mortality + incorrect boys) has the same number of observations as in the LIFE-M linking exercise described in section III: 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 South Carolina for North Carolina. When linking to this 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.

[27] To 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 name 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, the second character of the first name or last name is randomly changed. In 5 percent of cases, the 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 each, we replace the first name with an initial.

The advantage using this synthetic dataset is that it characterizes the performance of each matching algorithm relative to an *objective* truth similar to the LIFE-M sample. One disadvantage of this approach is that the error structure in names and ages is unknown, so our decisions about how to simulate error may be incomplete. In robustness checks (omitted for brevity), we find that the patterns of results (although not necessarily the exact levels) are very similar under different assumptions.

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 rich set of supplementary information. These veterans tended to report very 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 an extremely high match rate of 90 percent among men confirmed to live beyond the 1900 Census. This makes the match rates appear higher than in other samples, which do not make adjustments for mortality between census years. The Early Indicators project scores matches on a scale of 1 to 4 to indicate their confidence in a match. We use 1,875 matches coded as the highest quality (1 and 2) as the ground truth sample. Importantly, we do not use all possible records for which matches were attempted, so tests of representativeness should be interpreted as relative to the set of high quality links, which is not likely representative.

Our third sample is the IPUMS-LRS, which links the 1850, 1860, 1870 and 1900 Census samples to the 1880 full count Census. We restrict attention to men who were age 15 or younger and living with a mother and father in the 1850, 1860, 1870 Censuses or ages 15 to 40 in the 1900 Census for a sample of 349,712 men linked to the 1880 Census. Because IPUMS-LRS is itself a machine-linked sample, we use discordance in parents' birthplaces across years as an independent check of link quality (described in section II.).[28]

B. Results from Alternative Ground Truth Data

Table 1 presents the match rates for each method in each ground truth. For the synthetic data, the best 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

[28] When linking the 1-percent 1850, 1.2-percent 1860 sample, and 1.2-percent 1870 samples to the full-count 1880 Census, we use only men aged 15 and younger and living at home with both their parents in 1850-1870. For the 1.2-percent 1900 sample, we use only men who would have plausibly been age 0 to 15 in 1880 and include a five-year age window. When matching people in 1900 to 1880, we use men aged 15 to 40 in 1900. We exclude the 1910-1930 data, because there is no identifier on these samples to link the samples to the IPUMS-LRS. We restrict attention to people living at home because parent birthplaces are only available for men living at home in the 1850-1870 samples, which we require for our birth place discordance validation. We also consider results for our evaluation when applying IPUMS-provided person weights for each sample, including those specifically made for the IPUMS-LRS, and we find that they do not meaningfully impact the results we offer here.

LIFE-M match rates, with the exception of Feigenbaum (2016). Notably, Feigenbaum's (2016) method performs substantially better in the synthetic data than in the LIFE-M data with a match rate of 55 and 60 percent with the Iowa and estimated coefficients, respectively. The match rates for Early Indicators' veterans linked to the 1900 complete count Census are generally higher than in the LIFE-M sample. This pattern is due at least in part to the fact that we are linking a sample that are known to be linkable (i.e., did not die before 1900). In contrast, the match rates are much lower for all methods in the IPUMS-LRS. This result could reflect higher mortality, greater under-enumeration in the 1880 Census, or higher incidence of errors in enumeration or transcription.

Comparisons of match rates across methods in different ground truths also yield patterns similar to LIFE-M, although the levels differ. Match rates for Ferrie (1996) are in general lower than other methods and decrease with use of phonetic name cleaning. Although differences in record quality, the types of names and spelling variations, or other differences in the socio-demographic composition of the samples (i.e., differences in country of origin of parents, socio-economic characteristics of individuals represented, etc.) could affect conclusions about the role of phonetic name cleaning, the use of NYSIIS and Soundex increases match rates as does random selection among ties (Nix and Qian 2015). Finally, as in the LIFE-M data, restricting to a sample of uncommon names tends to lower match rates (Ferrie 1996).

Linked samples also appear unrepresentative across different ground truths. Columns 2 through 4 of Table 2 presents a heteroskedasticity-robust Wald-statistic from regressions of whether a record was linked on baseline covariates (described in the table notes; see full regression results in Appendix C). For the synthetic data, this exercise allows us to test the hypothesis that the non-representativeness in the LIFE-M sample 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.

Column 3 and 4 present these analyses for the Early Indicators and IPUMS-LRS data. As indicated in the table notes, data availability requires that we use a slightly different set of covariates across different ground truths.[29] As in the synthetic data, we reject that the linked samples are representative at the 10-percent

[29] For the synthetic dataset, we use the same covariates as in the LIFE-M data when considering representativeness. For the Early Indicators data, we use age, speaks English, owns a farm, currently married, foreign born, day of birth by year, literacy, length of first and last names, and foreign born status of parents. For the IPUMS-LRS data, we use age category variables, the size of the local place,

significance level for all methods except for Abramitzky et al. (2014) used with NYSIIS and with Soundex. 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. For the IPUMS-LRS, we also reject representativeness: linked individuals tend to have fewer siblings, and are less likely to be native born, have native-born parents, be married, and live in urban areas.

In terms of false links, automated methods produce relatively high Type I error rates in each these three additional datasets. Despite differences in the match rates, period, and type of records, Figure 3 and Table 1 show similar *patterns* (though different levels) in Type I error rates as found in the LIFE-M data. Averaging across all methods except the ground truths, roughly one third of links are incorrect, a similar finding to that in the LIFE-M data. Furthermore, using phonetic name cleaning and trying to link common names raises the share of false matches in all samples, just as in the LIFE-M data by 14 percent with NYSIIS and 32 percent for Soundex. Feigenbaum's (2016) method with estimated coefficients performs better outside the LIFE-M sample, achieving the low average of 18 percent errors versus the Ferrie (Name) method averaging 24 percent errors in the alternative ground-truth data. In particular, the Feigenbaum (2016) method with estimated coefficients does especially well in the synthetic data, achieving a Type I error rate of 16 percent. Part of this result may stem from the quality of the training data used in the synthetic case, where the training data was the true links that were known to us from having created the data. This evidence suggests that this regression-based method is well-suited to address the specific age and name perturbations we introduced in the synthetic data, but its performance varies considerably in the other ground truth samples.

A key factor in differences in Type I error rates across datasets is the number of tied links. Notably, the higher Type I error rate for the LIFE-M than synthetic data for the Nix and Qian method indicates that the number of candidate ties for a given record is higher in the LIFE-M data. This finding provides insight about the difference between our synthetic error generating process and the true error process: in practice, very common names are *less* likely to be misenumerated and transcribed. Because our synthetic process creates errors as frequently for common and uncommon names, the synthetic data error process *lowers* the number of exact ties by more than happens in the LIFE-M data. Since ties are less common as in Early Indicators and IPUMS-LRS, random selection or probability weighting have smaller effects on outcomes. Finally, as with the LIFE-M data, Table 3 shows that erroneous links tend to be strongly related to baseline characteristics (full set

birth region, relationship to household head, occupation categories, whether or not the respondent lives at home with parents, whether or not the respondent's parents were foreign born, region of residence, marital status, farm status, number of siblings, and whether or not the respondent lives in the same state of birth. We compare representativeness of the matched data by appending on the relevant final year Censuses, as these were the data that were used to compute the IPUMS-LRS weights.

of results in Appendix D).

Lastly, in our evaluation of errors in the IPUMS-LRS matches, we find that 4 percent of IPUMS-LRS links are probable Type I errors using disagreement in both parents' birthplaces as a metric. This rate is a higher number than reported in IPUMS internal assessments of 0.2 and 2 percent (Goeken et al. 2011) but similar to clerically reviewed data like the LIFE-M sample. This rate is substantially lower than the rate we observe with other automated matching algorithms, which is one reason we use this as a "ground truth" sample.

Summarizing these results, Figure 4 plots the relationship between Type I and Type II error rates for each method and dataset. The plot 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, methods have also tended to increase Type I errors (and decrease precision). For the synthetic and Early Indicators data, Type I and Type II errors exhibit a positive relationship, with slope parameters of 0.43 and 0.51, respectively. These results reinforce the finding that recent methods have often increased both Type 1 errors *and* Type II errors. In contrast, the IPUMS-LRS data show a negative relationship: for an increase in the Type II error rate of 0.1, the false positive rate falls by 0.07. There is no obvious reason why these results should differ, but they underscore the point that generalizing from one sample to others may prove misleading.

The main finding is the overall similarity of our results across datasets. Our analyses in four high quality linked datasets robustly show that an average of 32 percent of links using common automated methods are likely incorrect. Phonetic name cleaning universally increases false link rates by an average of 23 percent (14 and 32 percent with NYSIIS and Soundex, respectively), but sometimes by more than 60 percent. Linking more common names also worsens false link rates by an average of 30 percent. Methods to break ties through random selection or simple probability weighting increase the instance of false links by an average of 40 percent, but in some cases more than 75 percent. Considering all automated methods in this paper, using Ferrie's (1996) method without *common names* and *without name cleaning* achieves the lowest and most robust average rate of false matches of 22 percent. Feigenbaum's (2016) method also achieves an average rate of false positives around 22 percent, but this rate is less robust across datasets (varies from 13 to 35 percent of records). Finally, erroneous links are systematically related to baseline sample characteristics which suggests they may introduce systematic measurement error into analyses.

V. 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' income plays in determining their children's wage earnings. Intergenerational mobility is measured as 1- π, 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 PSID, 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 in links for sons may have vastly 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 $s_{xy}^{\ell *} + s_{xy}^{\ell i} = \sum_g s_{xy}^{\ell g} = \sum_g \sum_k \left(\log(x_{kg}) - \overline{\log(x_{kg})}\right)\left(\log(y_{kg}) - \overline{\log(y_{kg})}\right)$, $s_{xy}^{\ell b} = \sum_g N_g \left(\overline{\log(x_{kg})} - \overline{\overline{\log(x_{kg})}}\right)\left(\overline{\log(y_{kg})} - \overline{\overline{\log(y_{kg})}}\right)$ where group means are defined with a single bar and overall means are defined by two bars, such that,

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

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 to be 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 could happen in practice if errors (e.g., enumeration error or transcription) randomly assign records to these groups. Initially, we assume this term is zero to simplify exposition but later relax this assumption. Furthermore, note 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 $\text{cov}(\varepsilon, \log(x^{\ell i}))=0$, then

$$\text{plim}\,\hat{\pi}^{\ell i} = \frac{\text{cov}\left(\log(y^{*}),\log(x^{\ell i})\right)}{\text{var}(\log(x^{\ell i}))} = \frac{\text{cov}\left(\pi\log(x^{*})+\varepsilon,\log(x^{\ell i})\right)}{\text{var}(\log(x^{\ell i}))} = \pi\frac{\text{cov}\left(\log(x^{*}),\log(x^{\ell i})\right)}{\text{var}(\log(x^{\ell i}))} \quad (4)$$

If the imputed father's income is the same as the true father's income, $\log(x^{*}) = \log\left(x^{\ell i}\right)$, then $\text{plim}\,\hat{\pi}^{\ell *} = \text{plim}\,\hat{\pi}^{i}$. However, if $\frac{\text{cov}\left(\log(x^{*}),\log(x^{\ell i})\right)}{\text{var}(\log(x^{\ell i}))} \neq 1$, 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 $\frac{\text{cov}\left(\log(x^{*}),\log(x^{\ell i})\right)}{\text{var}(\log(x^{\ell i}))} = 0$. Then, the $\text{plim}\,\hat{\pi}^{\ell i} = 0$ and the estimator is inconsistent in proportion to the share of imputed links, $\text{plim}\hat{\pi}^{\ell} = \theta^{\ell *}\pi$. Second, consider the case where imputed father's income equals the true father's income plus noise, or $\log\left(x^{\ell i}\right) = \log(x^{*}) + u$. Under the assumptions of the classical errors in variables model ($\text{plim}(u\varepsilon) = 0$, $\text{plim}\left(\text{ulog}\,(x^{*})\right) = 0$, and $\text{plim}\left(\text{ulog}\,(y)\right) = 0$), then $\text{plim}\,\hat{\pi}^{\ell i} = \theta^{\ell i}\left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\right)\pi$. Moreover, $\text{plim}\,\hat{\pi}^{\ell} = (1-\theta^{\ell *})\,\pi + \theta^{\ell *}\left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\right)\pi$. 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 error on inference using Nix and Qian's (2015) method

of random selection among identical ties. Consider a setting where N is the total number of records that one wishes to link and for $M \leq N$ of these records, $r=1,2, \ldots, M$, there are J_r candidate matches that are exact ties. 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 <u>one</u> correct link among the ties for record, r, indexed by $j=1$, $\log(x_1)$, and imputed (but incorrect links) $\log(x_j)$, $j=2,\ldots,J_r$. From the researcher's perspective, the correct link is unknown and the probability that any one of the ties is correct is $\frac{1}{J_r}$. Under the Nix and Qian (2015) method, 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 $\hat{\pi}^{\ell} = \frac{s^{*}_{xx}}{s_{xx}}\hat{\pi}^{*\ell} + \frac{s^{j=1}_{xx}}{s_{xx}}\hat{\pi}^{j=1,\ell} + \frac{s^{j>1,\ell}_{xx}}{s_{xx}}\left(\sum_{j=2}^{J_r}\frac{s^{j,\ell}_{xx}}{s^{j>1,\ell}_{xx}}\hat{\pi}^{j\ell}\right)$ and $\operatorname{plim}\left(\sum_{j=2}^{J_r}\frac{s^{j\ell}_{xx}}{s^{j>1,\ell}_{xx}}\hat{\pi}^{j\ell}\right) = \sum_{j=2}^{J_r}\lambda_j \operatorname{plim}\pi^{j\ell}$. Therefore, the estimate of the intergenerational income elasticity using random selection to break ties is,

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

Note that this estimator is inconsistent if $\psi_{j\ell} = \frac{\operatorname{cov}(\log(x_1),\log(x_j))}{\sigma^2_{x_j}} < 1$.[30] 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.[31] 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

[30] Note that $\operatorname{plim}(\pi^j) = \frac{\operatorname{cov}(\log(y_1),\log(x_j))}{\sigma^2_{x_j}} = \frac{\operatorname{cov}(\pi\log(x_1)+\varepsilon,\log(x_j))}{\sigma^2_{x_j}} = \frac{\operatorname{cov}(\log(x_1),\log(x_j))}{\sigma^2_{x_j}}\pi$.

[31] Reducing 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 estimator 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).

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, 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 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, Collins and Wanamaker 2016, 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 a 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 (where $\pi^{\mathrm{h}} \leq \pi^{\mathrm{l}}$), 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,

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

The inconsistency of the probability limit in equation (5) depends upon several factors. First, if $\pi^{\mathrm{h}}=\pi^{l}$ 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 $\pi^{\mathrm{h}} \neq \pi^{l}$ (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 $\pi^{\mathrm{h}}=\pi^{l}$ but the group means are different, then the selection term will not be 0 and inferences will be affected in an

ambiguous way. Both effect 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 et al. 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 5A reports estimates of the intergenerational elasticity of income using samples of sons linked using different methods. For the LIFE-M links, we estimate an income elasticity of 0.22 between fathers and sons. Consistent with lifecycle bias and transitory income fluctuations attenuating our estimates, this estimate is lower than modern estimates.[32] 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, Nix and Qian's (2015) method of random selection among ties results in Type I error rates ranging from 54 to 69 percent and yields intergenerational income elasticity estimates of 0.19 to 0.12.

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 population (Bailey et al. 2017).[33] Figure 5B shows that the reweighted intergenerational income elasticities tend to be slightly smaller in magnitude than the unweighted Figure 5A estimates. This result may stem from the modest over-representation of larger, less-mobile families in the linked sample. The effects of non-representativeness (as measured by the changes induced by reweighting) on observed characteristics, however, appears modest in comparison to the role of

[32] For 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 famers—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.

[33] To construct these weights, we first run a probit model of link status (for each method) on covariates, $\boldsymbol{X}$, 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, $P_i(L_i = 1|\boldsymbol{X_i})$, for each method and reweight observations by $(1 - P_i(L_i = 1|\boldsymbol{X_i}))/P_i(L_i = 1|\boldsymbol{X_i}) * q/(1 - q)$, where q is the share of records that are linked. Distributions of inverse propensity score weights are plotted in Appendix E.

errors in linking.

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 22 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 0.16 (=0.78*0.21). 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, Nix and Qian (2012) with Soundex shows a 69 percent Type I error rate and the intergenerational elasticity estimate is 0.12 in Figure 5A.

Figure 5C and Figure 5D examine the effects of incorrect matches *directly* 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 ×) 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 5C) and 0.21 with inverse propensity-score weights (Figure 5D). 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 (Soundex) to 1,600 for Nix and Qian (Soundex), but the unweighted intergenerational income elasticities are estimated to be 0.22 and 0.21, respectively. Consistent with Type I errors introducing attenuation, $\hat{\pi}^{i}$ tend 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 (Name) is 0.11 and only 0.04 for Abramitzky et al. (Soundex) and 0.05 for Nix and Qian (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 show how the regression-based classification system selects links with a very high correlation to the true link—even when false. 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.[34]

[34] Reweighting our sample to represent individuals in the 1940 Census, as we do in Appendix E, produces estimates that are closer to the LIFE-M elasticity across all methods—*even* when incorrect links are included. Interestingly, this reweighting procedure sometimes assigns lower p-score weights to matches that are incorrect, thus reducing the influence of these observations for some methods (Abramitzky et al. 2004, NYSIIS). For other methods (Ferrie 1996-Name + uncommon names; Nix and Qian Name), incorrect links receive more weight substantially attenuating the estimates. Of course, the appropriate way to re-weight a sample depends upon the research question of interest.

VI. LESSONS FOR HISTORICAL RECORD LINKING

New large-scale longitudinal and intergenerational data for the U.S. have the potential to revolutionize empirical social science. The need for reliable linking methods is especially high in the U.S., where the linking variables in public data tend to be more limited than in modern administrative data and relative to records in other countries. Using different U.S. ground truth samples, this paper documents how linking errors could have large effects on scientific conclusions and policy inferences. Not only are linked samples non-representative, existing algorithms yield very high rates of false matches. Our case study of intergenerational income elasticities shows that some combinations of linking assumptions attenuate estimates by more than 50 percent. Of course, linking based on more variables—especially those with many values (e.g., social security numbers or exact dates of birth)—or higher quality information (e.g., administrative records rather than individual reports) will attain lower error rates than we document. Although we cannot diagnose how much linking assumptions matter in other contexts, our results broadly 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.

For limited linking variables which are measured with error, clerical review (e.g., LIFE-M) and genealogical methods (e.g., Early Indicators) attain much lower error rates than machine algorithms. Because these labor-intensive methods are cost prohibitive for most projects, our findings 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 links allows researchers to diagnose and potentially remedy algorithm errors arising for specific records or in a particular historical context.

Second, we recommend against using phonetic cleaning and using commonly occurring observations. Limiting attention to distinctive observations, like Ferrie's (1996) approach of only linking uncommon names or Abramitzky et al.'s (2014) robustness check using unique name-age combinations in a five year window (see Appendix A), reduces sample sizes but substantially reduces false matches. In contrast, the common practice of weighting ties equally by the inverse of their empirical frequency incorporates a large number of known false links and may result in substantial attenuation.

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. Feigenbaum's (2016) method or Ruggles et al.'s (2015) SVM are examples of this. 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 and these errors, therefore, have less impact on inferences.

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 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 leverages the common strengths of different methods. For example, when using LIFE-M data, combining methods like Ferrie (1996) and Feigenbaum (2016) drives the match rate to 25 percent and the Type I error rate to 12 percent—much lower than the 22 to 30 percent error rate for either method individually. Using this same combination in other ground truths achieves a similarly low error rate. In addition, this technique provides a low-cost way to identify false links and remedy algorithm issues contributing to bad links.

After purging samples of incorrect links, 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 improve representativeness.[35] 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.

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. In the parlance of machine learning, "precision" should be weighted more heavily than "recall." Modern surveys such as the Panel Survey of Income Dynamics and National Longitudinal Surveys 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 solved without more data 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.

[35] See Bailey et al. (2017) for a simple implementation description.

VII. REFERENCES

A'Hearn, B., J. Baten and D. Crayen (2009). "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital." The Journal of Economic History 69(3): 783-808.

Abowd, J. M. (2017). "Large-Scale Data Linkage from Multiple Sources: Methodology and Research Challenges." NBER Summer Institute Methods Lecture.

Abowd, J. M. and L. Vilhuber (2005). "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers." Journal of Business and Economic Statistics 23(2): 133-165.

Abramitzky, R., L. Platt Boustan and K. Eriksson (2012). "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, R., L. Platt Boustan and K. Eriksson (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.

Abramitzky, R., L. Platt Boustan and K. Eriksson (2014). "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." Journal of Political Economy 122(3): 467-506.

Aizer, A., S. Eli, J. Ferrie and A. Lleras-Muney (2016). "The Long Term Impact of Cash Transfers to Poor Families." American Economic Review 106(4): 935-971.

Antonie, L., K. Inwood, D. J. Lizotte and J. A. Ross (2014). "Tracking People over Time in 19th Century Canada for Longitudinal Analysis." Machine Learning 95(1): 129-146.

Atack, J. (2004). "A Nineteenth-Century Resource for Agricultural History Research in the Twenty-First Century." Agricultural History 78(4): 389-412.

Atack, J., F. Bateman and M. E. Gregson (1992). "Matchmaker, Matchmaker, Make Me a Match." Historical Methods 25(2): 53-65.

Bailey, A. K., S. E. Tolnay, E. M. Beck and J. D. Laird (2011). "Targeting Lynch Victims: Social Marginality or Status Transgressions?" American Sociological Review 76(3): 412-436.

Bailey, M., C. Cole and C. G. Massey (2017). "Representativeness and False Links in the 1850-1930 Ipums Linked Representative Historical Samples." Retrieved October 1, 2017. Available at http://www-personal.umich.edu/~baileymj/Bailey_Cole_Massey.pdf.

Bailey, M. J., S. Anderson, A. Karimova and C. G. Massey (2016). "Creating Life-M: The Longitudinal, Intergenerational Family Electronic Micro-Database."

Black, S. E. and P. J. Devereux (2011). Recent Developments in Intergenerational Mobility. In Handbook of Labor Economics, edited by David, C. and A. Orley. Amsterdam: Elsevier. Volume 4B: 1487-1541.

Bleakley, H. and J. Ferrie (2014). "Land Opening on the Georgia Frontier and the Coase Theorem in the Short- and Long- Run." Retrieved Available at http://www-personal.umich.edu/~hoytb/Bleakley_Ferrie_Farmsize.pdf.

Bleakley, H. and J. Ferrie (2016). "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital across Generations." Quarterly Journal of Economics 131(3): 1455-1495.

Bleakley, H. and J. P. Ferrie (2013). "Up from Poverty? The 1832 Cherokee Land Lottery and the Long-Run Distribution of Wealth." NBER Working Paper 19175.

Bogue, A. G. (1963). From Prairie to Corn Belt: Farming on the Illinois and Iowa Prairies in the Nineteenth Century. Chicago: University of Chicago Press.

Boustan, L. P. and W. Collins (2014). The Origins and Persistence of Black-White Differences in Women's Labor Force Participation from the Civil War to the Present. In In Human Capital and History: The American Record, edited by Boustan, L., C. Frydman and R. A. Margo. Chicago, IL: University of Chicago Press.

Boustan, L. P., M. E. Kahn and P. W. Rhode (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.

Buckles, K. S. and D. M. Hungerman (2013). "Season of Birth and Later Outcomes: Old Questions, New Answers." Review of Economics and Statistics 95(3): 711-724.

Chetty, R., N. Hendren, P. Kline, E. Saez and N. Turner (2014). "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility." American Economic Review 104(5): 141-147.

Christen, P. and K. Goiser (2007). Quality and Complexity Measures for Data Linkage and Deduplication.

Collins, W. J. and M. H. Wanamaker (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, W. J. and M. H. Wanamaker (2016). "Up from Slavery? African American Intergenerational Economic Mobility since 1880." Retrieved November 8, 2016. Available at http://www.nber.org/papers/w23395.pdf.

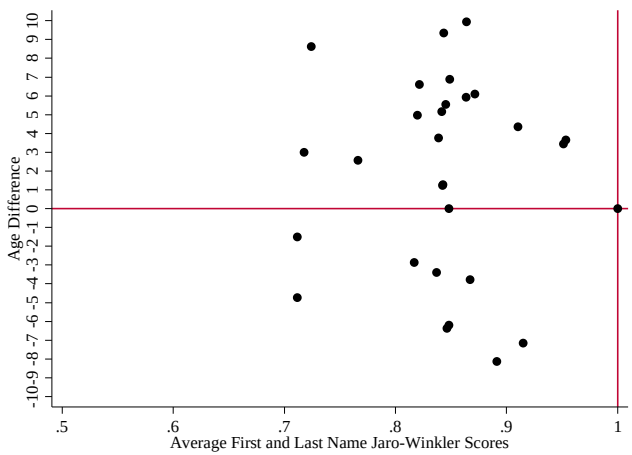
Costa, D. L., H. DeSomer, E. Hanss, C. Roudiez, S. E. Wilson and N. Yetter (2017). "Union Army Veterans, All Grown Up." Historical Methods 50(2): 79-95.

Curti, M. (1959). The Making of an American Community: A Case Study of Democracy in a Frontier County. Stanford: Stanford University Press.
DiNardo, J., N. M. Fortin and T. Lemieux (1996). "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." Econometrica 64(5): 1001-1044.
Duncan, O. D. (1968). "Patterns of Occupational Mobility among Negro Men." Demography 5(1): 11-22.
Eli, S., L. Salisbury and A. Shertzer (2016). Migration in Response to Civil Conflict: Evidence from the Border of the American Civil War. Available at http://www.nber.org/papers/w22591.
Eriksson, B. (2016). "The Missing Links: Data Quality and Bias to Estimates of Social Mobility." Retrieved September 15, 2016. Available at www.fas.nus.edu.sg/cfpr/RC28/089.pdf.
Feigenbaum, J. J. (2016). "A Machine Learning Approach to Census Record Linking." Retrieved March 28, 2016. Available at http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf?m=1423080976.
Ferrie, J. 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, R., T. Lynch, Y. N. Lee, J. Wellington and D. Magnuson (2016). "Evaluating the Accuracy of Linked U. S. Census Data: A Household Approach." Retrieved Available at request of the authors.
Greene, W. H. (2008). Econometric Analysis, 6th Edition. New York: Pearson.
Guest, A. M. (1987). "Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century." Historical Methods 20(2): 63-77.
Hacker, J. D. (2010). "Decennial Life Tables for the White Population of the United States, 1790-1900." Historical Methods 43(3): 45-79.
Hacker, J. D. (2013). "New Estimates of Census Coverage in the United States, 1850-1930." Social Science History 37(1): 71-101.
Haider, S. J. and G. Solon (2006). "Life-Cycle Variation in the Association between Current and Lifetime Earnings." American Economic Review 96(4): 1308-1320.
Heckman, J. J., H. Ichimura, J. Smith and P. Todd (1998). "Characterizing Selection Bias Using Experimental Data." Econometrica 66(5): 1017-1098.
Hornbeck, R. and S. Naidu (2014). "When the Levee Breaks: Black Migration and Economic Development in the American South." American Economic Review 104(3): 963-990.
Horowitz, J. L. and C. F. Manski (1995). "Identification and Robustness with Contaminated and Corrupted Data." Econometrica 63(2): 281-302.
Hout, M. and A. M. Guest (2013). "Intergenerational Occupational Mobility in Great Britain and the United States since 1850: Comment." American Economic Review 103(5): 2021-2040.
Huber, P. J. (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, M. 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, G. and R. Chambers (2012). "Regression Analysis under Probabilistic Multi-Linkage." Statistica Neerlandica 66(1): 64-79.
Lahiri, P. and M. D. Larsen (2005). "Regression Analysis with Linked Data." Journal of the American Statistical Association 100(469): 222-230.
Malin, J. (1935). "The Turnover of Farm Pouplation in Kansas." Kansas Historical 20: 339-372.
Margo, R. A. (2016). "Obama, Katrina, and the Persistence of Racial Inequality." Journal of Economic History 76(2): 301-341.
Massey, C. G. (2017). "Playing with Matches: An Assessment of Accuracy in Linked Historical Data." Historical Methods: A Journal of Quantitative and Interdisciplinary History: 1-15.
Mazumder, B. (2015). "Estimating the Intergenerational Elasticity and Rank Association in the U.S.: Overcoming the Current Limitations of Tax Data." Retrieved November 10, 2015. Available at https://www.chicagofed.org/publications/working-papers/2015/wp2015-04.
Mazumder, B. (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.
Mill, R. (2013). Record Linkage across Historical Datasets. Stanford University.
Mill, R. and L. C. Stein (2016). "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America." Retrieved March 1, 2016. Available at http://www.public.asu.edu/~lstein2/research/mill-stein-skincolor.pdf.
Modalsli, J. (2017). "Intergenerational Mobility in Norway, 1865-2011." The Scandanavian Journal of Economics 119(1): 34-71.
National Office of Vital Statistics (1948). State and Regional Life Tables, 1939-1941.
Nix, E. and N. Qian (2015). "The Fluidity of Race: 'Passing' in the United States, 1880-1940." Retrieved January 30,

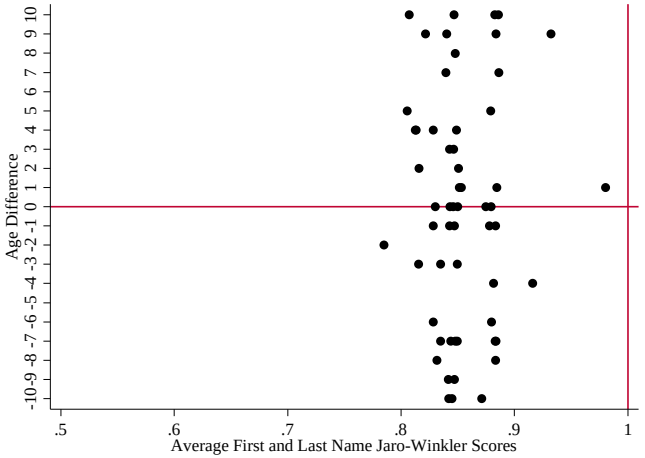
2015. Available at http://www.nber.org/papers/w20828.
Ruggles, S. (2011). "Intergenerational Coredisence and Family Transitions in the United States, 1850-1880." Journal of Marriage and the Family 73(1): 138-148.
Ruggles, S. (2006). "Linking Historocal Censuses: A New Approach." History and Computing 14(1-2): 213-224.
Ruggles, S., K. Genadek, J. Grover and M. Sobek (2015). Integrated Public Use Microdata Series (Version 6.0) [Machine-Readable Database]. Minneapolis: University of Minnesota.
Saperstein, A. and A. Gullickson (2013). "A Mulatto Escape Hatch? Examining Evidence of U.S. Racial and Social Mobility in the Jim Crow Era." Demography 50(5): 1921-1942.
Schaefer, D. (1985). "A Statistical Profile of Frontier and New South Migration: 1850-1860." Agricultural History 59: 563-567.
Scheuren, F. and W. E. Winkler (1993). "Regression Analysis of Data Files That Are Computer Matched." Survey methodology 19(1): 39-58.
Solon, G. (1992). "Intergenerational Income Mobility in the United States." American Economic Review 82(3): 393-408.
Solon, G. (1999). Intergenerational Mobility in the Labor Market. In Handbook of Labor Economics, edited by Ashenfelter, O. and D. Card. Amsterdam: Elsevier. 1761-1800.
Steckel, R. (1988). "Census Matching and Migration: A Research Strategy." Historical Methods 21(2): 52-60.
Stephens, M., Jr. and T. Unayama (2017). "Estimating the Impacts of Program Benefits: Using Instrumental Variables with Underreported and Imputed Data." Retrieved February 1, 2017. Available at http://www-personal.umich.edu/~mstep/IV_Underreporting_20170131.pdf.
Thernstrom, S. (1964). Poverty and Progress: Social Mobility in a Nineteenth Century City. Cambridge: Harvard University Press.
Wasi, N. (2014). Reclink3. Available at http://EconPapers.repec.org/RePEc:boc:bocode:s456876.
West, K. K. and J. G. Robinson (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, W. E. (2006). "Overview of Record Linkage and Current Research Directions." Research Report Series, Statistics 2006(2).
Wisselgren, M. J., S. Edvinsson, M. Berggren and M. Larsson (2014). "Testing Methods of Record Linkage on Swedish Censuses." Historical Methods 47(3): 138-151.
Xie, Y. and A. Killewald (2013). "Intergenerational Occupational Mobility in Britain and the U.S. Since 1850: Comment." American Economic Review 103(5): 2003-2020.
Zimmerman, D. 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

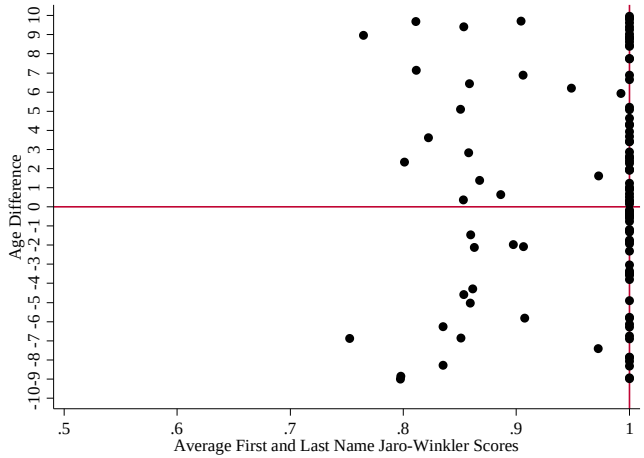
A. Albert Crock (Example of M1)



B. Raymond Bernaciak (Example of M2)



C. Author Smith (Example of M3)



D. Charles Hall (Example of M4)

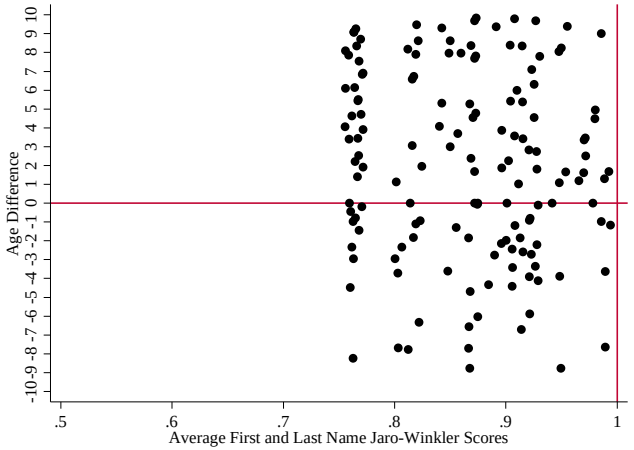
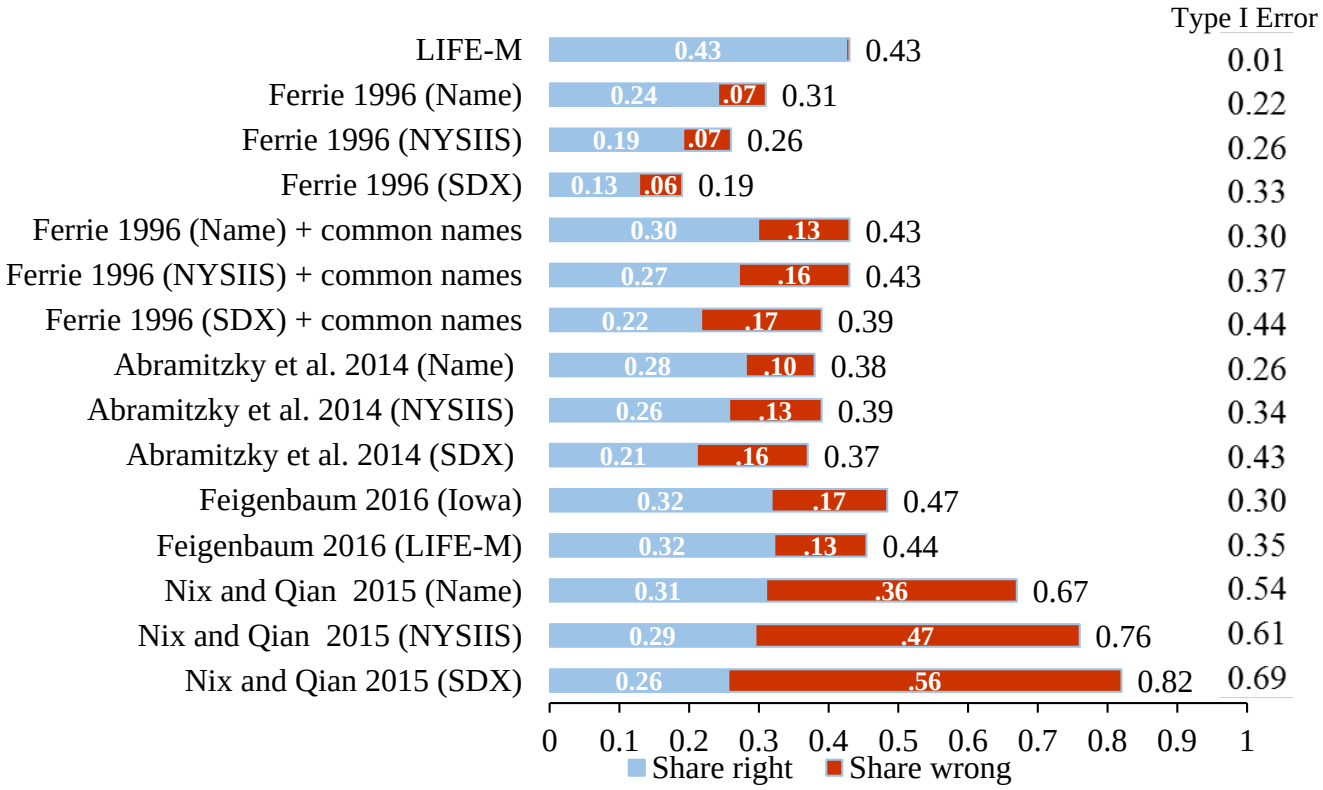
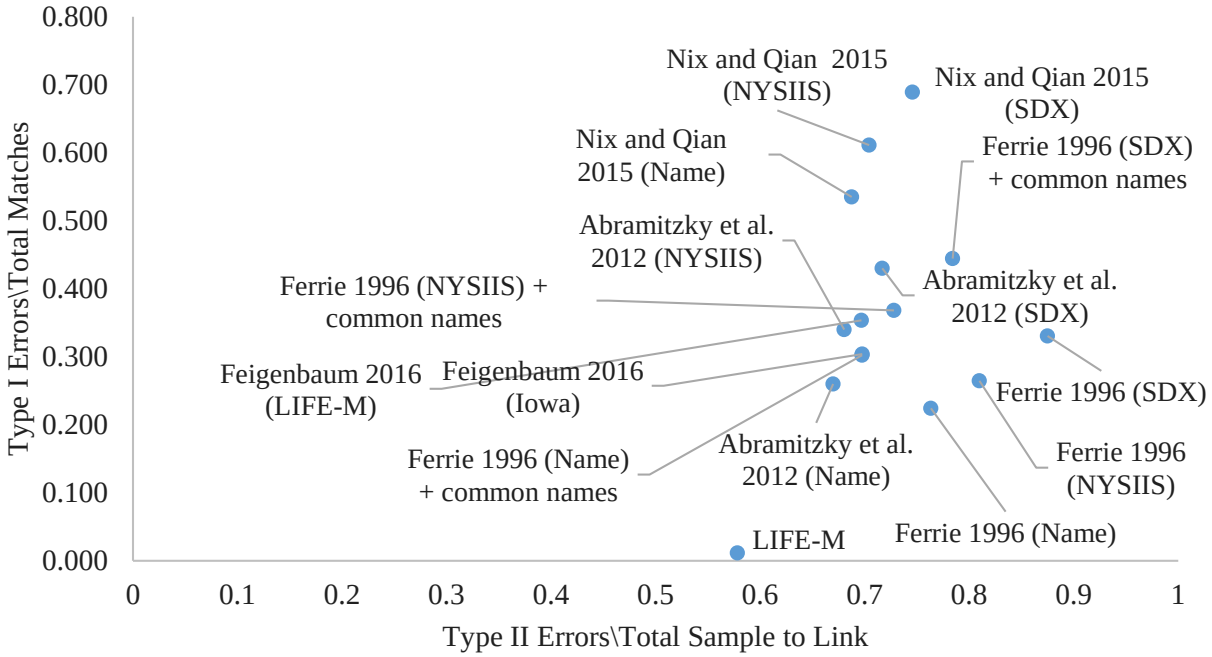


Figure 2. Performance of Automated Linking Methods using the LIFE-M Data

A. Match Rates and False Links

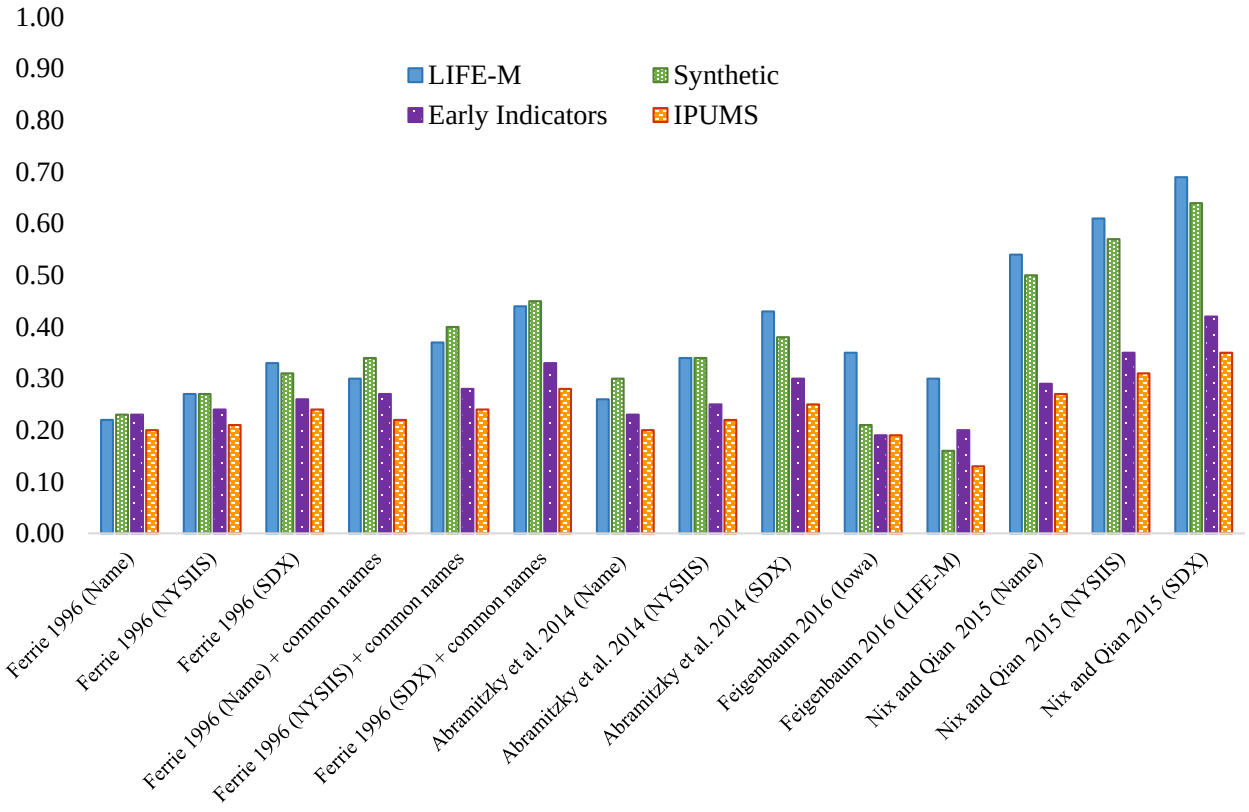


B. Type I versus Type II Errors



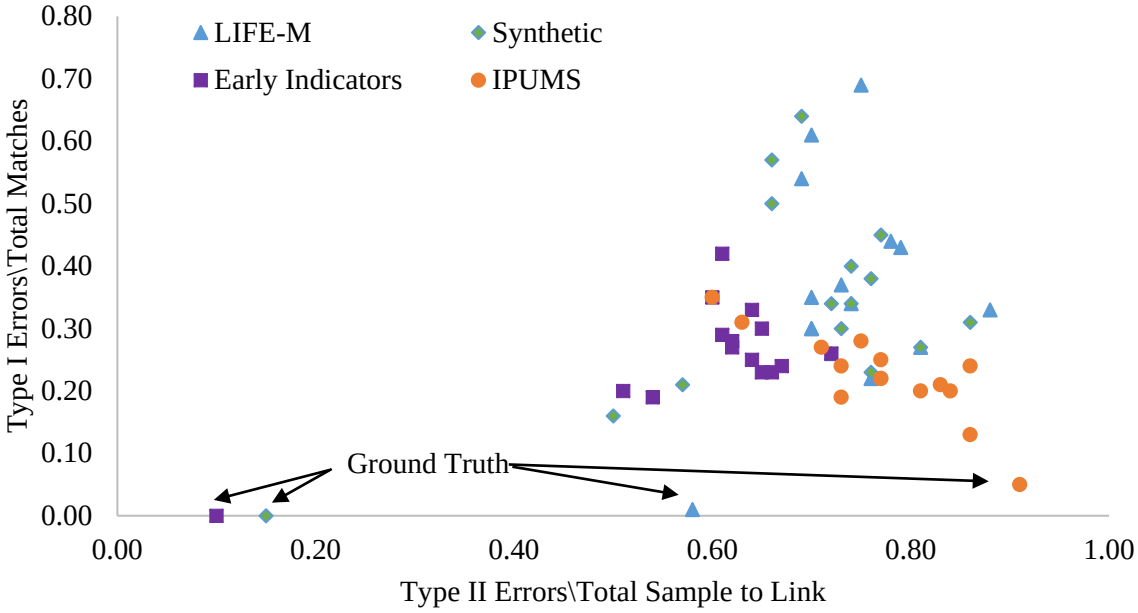
Notes: Panels use the LIFE-M sample of Ohio boys linked to the 1940 Census using different automated methods. See Table 1 for numerical estimates.

Figure 3. Share of Matches Incorrect (Type I Error Rate) by Method and Ground Truth



Notes: Type I error is the share of links that are incorrect. See text for details and descriptions of samples. See Table 1 for numerical estimates for match rates, Type I errors, and Type II errors.

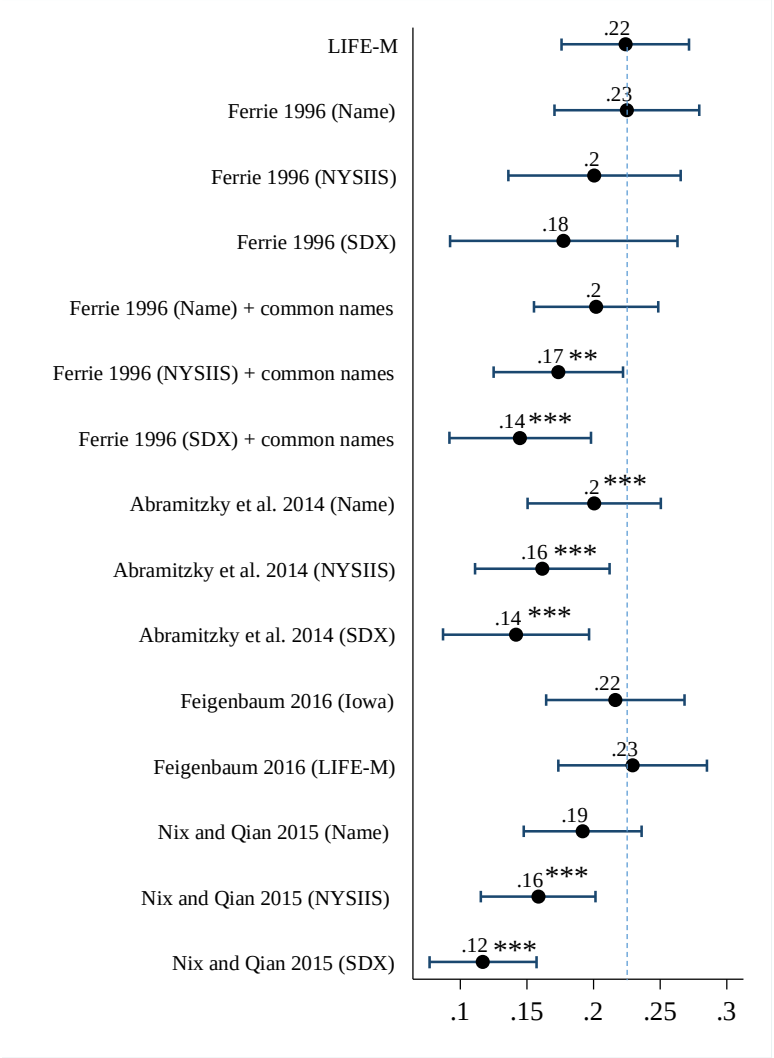
Figure 4. Type I vs. Type II Error Rates by Method and Data



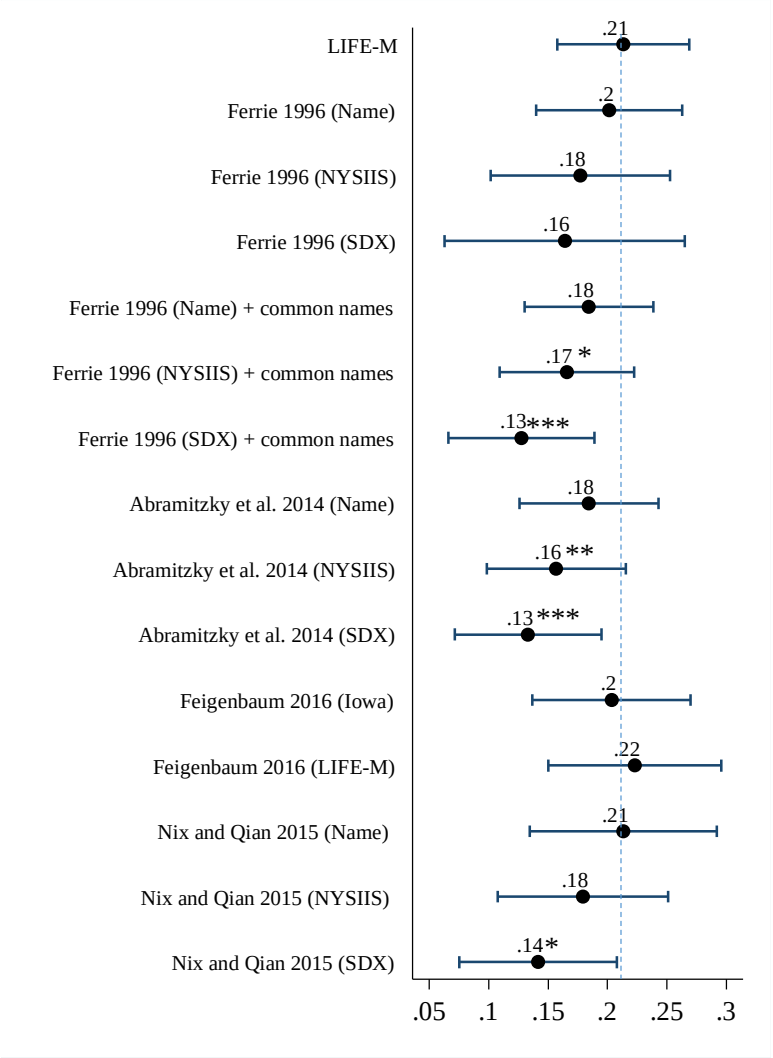
Notes: See Figure 3 notes.

Figure 5. Intergenerational Income Elasticity Estimates

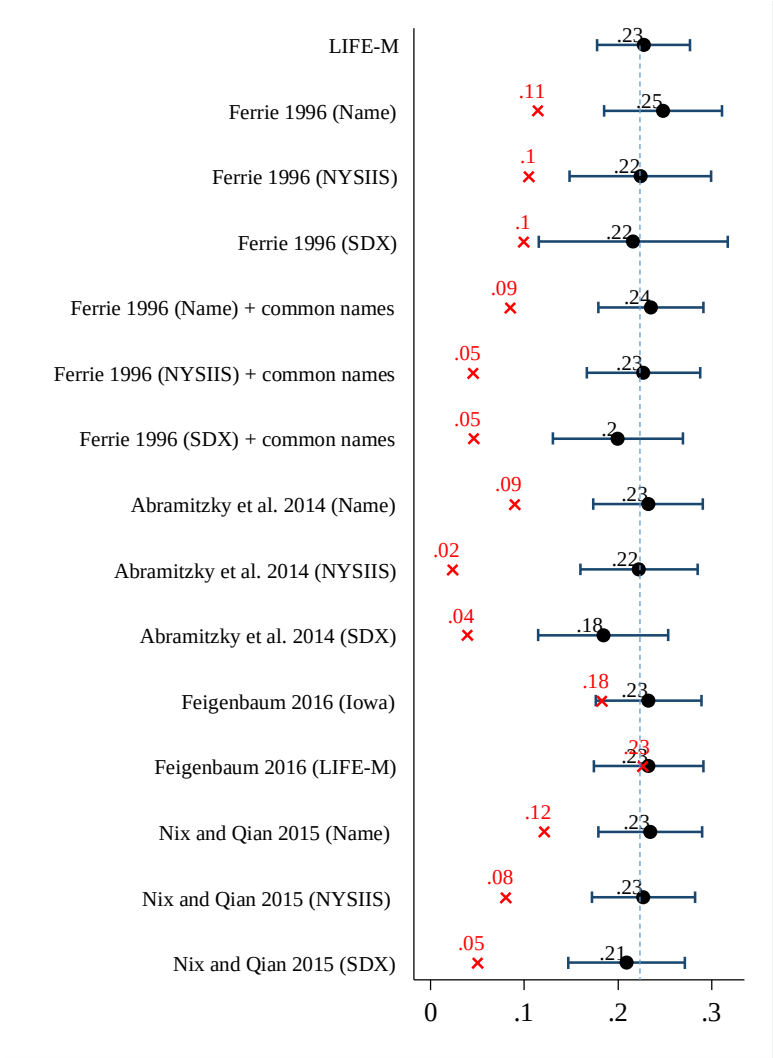
A. Unweighted Linked Samples



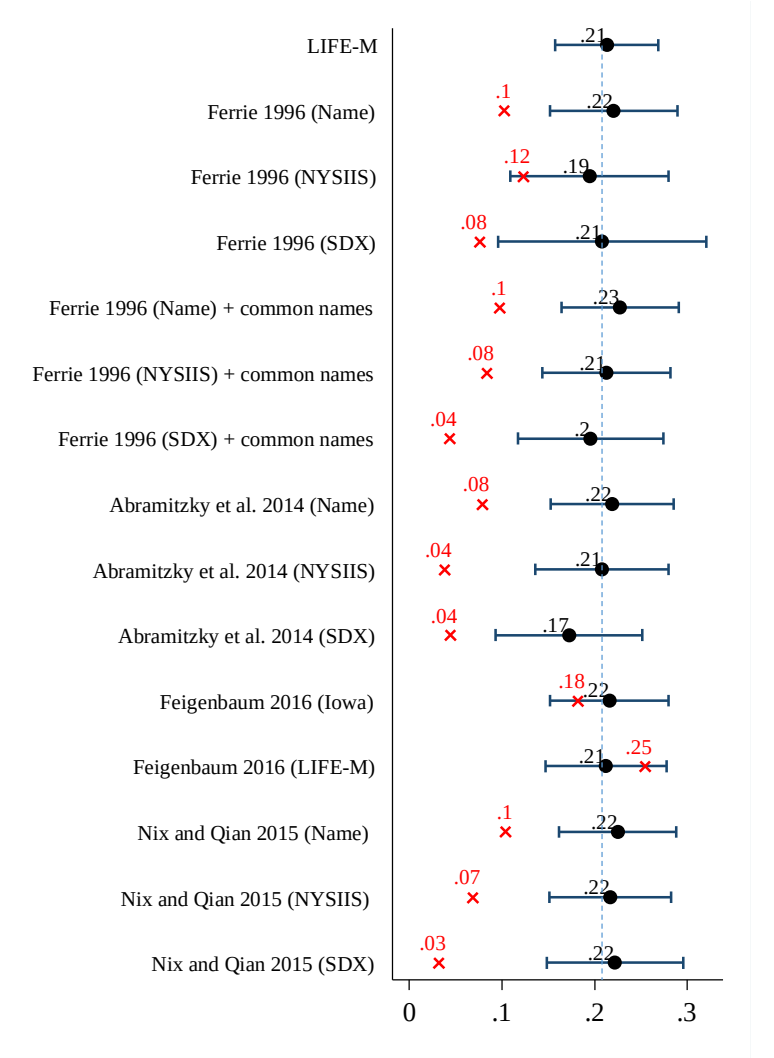
B. Inverse Propensity-Score Weighted Linked Samples



C. Separate Regressions for Imputed and Correct Links, Unweighted



D. Separate Regressions for Imputed and Correct Links, Weighted



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 1,702 for LIFE-M, 1,282 for Ferrie 1996 (Name), 1,034 for Ferrie 1996 (NYSIIS), 684 for Ferrie 1996 (Soundex), 1,699 for Ferrie 1996 (Name) + common names, 1,644 for Ferrie 1996 (NYSIIS) + common names, 1,406 for Ferrie 1996 (SDX) + common names, 1,549 for Abramitzky et al. 2014 (Name), 1,549 for Abramitzky et al. 2014 (NYSIIS), 1,364 for Abramitzky et al. 2014 (SDX), 2,371 for Nix and Qian (Name), 2,683 for Nix and Qian (NYSIIS), 2,884 for Nix and Qian (SDX), 1,814 for Feigenbaum 2016 (Iowa), and 1,715 for Feigenbaum 2016 (LIFE-M). * 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.

Table 1. Summary of Match Rates and Error Rates, by Method and Ground Truth

	A. Match Rates				B. Type I Error Rate (False Links)				C. Type II Error Rate (Missed links)			
	LIFE-M	Synthetic	EI	IPUMS	LIFE-M	Synthetic	EI	IPUMS	LIFE-M	Synthetic	EI	IPUMS
Ground Truth	0.43	0.85	0.90	0.09	0.01	0.00	0.00	0.05	0.58	0.15	0.10	0.91
Ferrie 1996 (Name)	0.31	0.31	0.45	0.20	0.22	0.23	0.23	0.20	0.76	0.76	0.65	0.84
Ferrie 1996 (NYSIIS)	0.26	0.26	0.43	0.21	0.27	0.27	0.24	0.21	0.81	0.81	0.67	0.83
Ferrie 1996 (SDX)	0.19	0.20	0.38	0.19	0.33	0.31	0.26	0.24	0.88	0.86	0.72	0.86
Ferrie 1996 (Name) + common names	0.43	0.43	0.52	0.29	0.30	0.34	0.27	0.22	0.70	0.72	0.62	0.77
Ferrie 1996 (NYSIIS) + common names	0.43	0.44	0.53	0.35	0.37	0.40	0.28	0.24	0.73	0.74	0.62	0.73
Ferrie 1996 (SDX) + common names	0.39	0.41	0.53	0.35	0.44	0.45	0.33	0.28	0.78	0.77	0.64	0.75
Abramitzky et al. 2014 (Name)	0.38	0.38	0.44	0.23	0.26	0.30	0.23	0.20	0.72	0.73	0.66	0.81
Abramitzky et al. 2014 (NYSIIS)	0.39	0.40	0.48	0.29	0.34	0.34	0.25	0.22	0.74	0.74	0.64	0.77
Abramitzky et al. 2014 (SDX)	0.37	0.39	0.50	0.31	0.43	0.38	0.30	0.25	0.79	0.76	0.65	0.77
Feigenbaum 2016 (Iowa coef.)	0.47	0.55	0.57	0.33	0.35	0.21	0.19	0.19	0.70	0.57	0.54	0.73
Feigenbaum 2016 (estimated coef.)	0.44	0.60	0.61	0.16	0.30	0.16	0.20	0.13	0.70	0.50	0.51	0.86
Nix and Qian 2015 (Name)	0.67	0.68	0.55	0.40	0.54	0.50	0.29	0.27	0.69	0.66	0.61	0.71
Nix and Qian 2015 (NYSIIS)	0.76	0.79	0.62	0.54	0.61	0.57	0.35	0.31	0.70	0.66	0.60	0.63
Nix and Qian 2015 (SDX)	0.82	0.87	0.67	0.61	0.69	0.64	0.42	0.35	0.75	0.69	0.61	0.60

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. Wald Test of Representativeness of Linked Sample, by Dataset and Linking Method

	LIFE-M	Synthetic Data	Early Indicators	IPUMS - LRS
Ground Truth	1,029	10.1	N/A	16,669
p-value	*0.00*	*0.26*		*0.00*
Ferrie 1996 (Name)	838.5	565.2	42.3	25,325
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (NYSIIS)	544.6	417.8	38.4	23,105
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (SDX)	146	124.0	56.2	17,255
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (Name) + common names	605.3	541.5	36.5	27,470
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (NYSIIS) + common names	594.4	636.1	16.4	26,634
p-value	*0.00*	*0.00*	*0.09*	*0.00*
Ferrie 1996 (SDX) + common names	345.2	488.9	24.8	22,174
p-value	*0.00*	*0.00*	*0.01*	*0.00*
Abramitzky et al. 2014 (Name)	727.9	426.8	36.5	24,509
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Abramitzky et al. 2014 (NYSIIS)	640.2	566.3	15.7	24,909
p-value	*0.00*	*0.00*	*0.11*	*0.00*
Abramitzky et al. 2014 (SDX)	412.5	451.2	16.1	22,876
p-value	*0.00*	*0.00*	*0.10*	*0.00*
Feigenbaum 2016 (Iowa coef.)	471.9	129.5	53.4	22,690
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Feigenbaum 2016 (estimated coef.)	782.7	253.5	27.3	39,242
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Nix and Qian 2015 (Name)	739.9	500.2	62.8	39,915
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Nix and Qian 2015 (NYSIIS)	825.2	345.0	54.2	46,453
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Nix and Qian 2015 (SDX)	672.1	94.2	45.5	48,038
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Observations	45,417	45,417	1,774	32,959,605

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 North Carolina. The covariates included in the Early Indicators data are age, speaks English, owns a farm, currently married, foreign born, day of birth by year, literacy, length of first and last names, and foreign born status of parents. The covariates included in the IPUMS data are the number of siblings, age, living with parents, foreign-born status of parents, indicators for geographic location, farm status, is currently married, indicators for occupation, race, and foreign-born status. These sample sizes are slightly smaller due to missing values. See appendices for full regression results.

Table 3. Wald Test of Whether Erroneous Links are Unrelated to Baseline Characteristics

	LIFE-M	Synthetic Data	Early Indicators	IPUMS - LRS
Ground Truth	212.5	N/A	N/A	743.1
p-value	*0.00*			*0.00*
Ferrie 1996 (Name)	414.4	83.5	39.1	1,957
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (NYSIIS)	257.7	44.0	22.1	2,708
p-value	*0.00*	*0.00*	*0.01*	*0.00*
Ferrie 1996 (SDX)	114.2	9.9	13.7	3,101
p-value	*0.00*	*0.27*	*0.19*	*0.00*
Ferrie 1996 (Name) + common names	877.2	145.1	43.5	3,625
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (NYSIIS) + common names	609.1	82.7	34.1	5,431
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Ferrie 1996 (SDX) + common names	301.2	32.7	27.2	7,815
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Abramitzky et al. 2014 (Name)	829.5	174.3	50.8	2,639
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Abramitzky et al. 2014 (NYSIIS)	673.7	107.5	37.6	4,246
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Abramitzky et al. 2014 (SDX)	375.3	44.6	27.3	6,088
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Feigenbaum 2016 (Iowa coef.)	1698.0	492.9	25.9	2,330
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Feigenbaum 2016 (estimated coef.)	1788.0	439.2	19.9	989.8
p-value	*0.00*	*0.00*	*0.01*	*0.00*
Nix and Qian 2015 (Name)	3130.0	682.9	40.7	7,796
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Nix and Qian 2015 (NYSIIS)	2084.0	362.4	56.7	14,472
p-value	*0.00*	*0.00*	*0.00*	*0.00*
Nix and Qian 2015 (SDX)	1078.0	91.0	57.2	22,514
p-value	*0.00*	*0.00*	*0.00*	*0.00*

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. P-values are reported in italics. The covariates included in each regression are indicated in Table 2 notes. See appendices for full regression results.

[click here for online appendices]