UCLA
UCLA Previously Published Works

Title

Simple strategies for improving inference with linked data: a case study of the 1850–1930 IPUMS linked representative historical samples

Permalink

https://escholarship.org/uc/item/5d6260zw

Journal

Historical Methods A Journal of Quantitative and Interdisciplinary History, 53(2)

ISSN

0161-5440

Authors

Bailey, Martha
Cole, Connor
Massey, Catherine

Publication Date

2020-04-02

DOI

10.1080/01615440.2019.1630343

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Hist Methods. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Hist Methods. 2020 ; 53(2): 80–93. doi:10.1080/01615440.2019.1630343.

Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850–1930 IPUMS Linked Representative Historical Samples

Martha Bailey[1,2], **Connor Cole**[1], **Catherine Massey**[1]

[1]University of Michigan

[2]National Bureau of Economic Research

Abstract

New large-scale linked data are revolutionizing quantitative history and demography. This paper proposes two complementary strategies for improving inference with linked historical data: the use of validation variables to identify higher quality links and a simple, regression-based weighting procedure to increase the representativeness of custom research samples. We demonstrate the potential value of these strategies using the 1850–1930 Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS)—a high quality, publicly available linked historical dataset. We show that, while incorrect linking rates appear low in the IPUMS-LRS, researchers can reduce error rates further using validation variables. We also show how researchers can reweight linked samples to balance observed characteristics in the linked sample with those in a reference population using a simple regression-based procedure.

Until recently, the dearth of longitudinal or intergenerational U.S. data for the late 19th and 20th centuries limited the study of important social, economic, demographic, and health questions. Much of the existing work on these questions has instead used cross-sectional or aggregated data—data that answer some questions but that often leave the mechanisms for both observed effects and policy generalizability unclear.[1]

Large-scale linked data are allowing researchers to break new ground on older questions and open entirely novel areas of inquiry.[2] New work, however, suggests that the prevalence of false links and missed matches in historical U.S. linked data may limit the contributions of this research. Bailey, Cole, Henderson, and Massey (2019) show that commonly used

[1]See, for instance, early-life public health initiatives (Alsan & Goldin, 2015; Cutler & Miller, 2005), exposures to environmental pollutants (Clay, Lewis, & Severnini, 2016) and animal diseases (Rhode & Olmstead, 2015), and access to medicines (Bleakley, 2007). Other examples include the long-run effects of exposure to human capital initiatives through Rosenwald schools (Mazumder & Aaronson, 2011).

[2]On-going 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, Genadek, Grover, & Sobek, 2015) and 1940 U.S. Censuses (the first U.S. census to ask about education and wage income) and newly available administrative sources. The Census Bureau plans to link the 1940 Census to current administrative and census data (Census Longitudinal Infrastructure Project, CLIP) and the Minnesota Population Center plans to link it to other historical censuses. The Panel Survey of Income Dynamics (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, Anderson, Karimova, & Massey, 2016). Supplementing these public infrastructure projects, entrepreneurial researchers have also combined large datasets. See, for example, Abramitzky, Platt Boustan, and Eriksson (2012, 2013, 2014), Boustan, Kahn, and Rhode (2012), Hornbeck and Naidu (2014); Mill (2013); Mill and Stein (2016), Aizer, Eli, Ferrie, and Lleras-Muney (2016), Bleakley and Ferrie (2014; 2016; 2013), Nix and Qian (2015), Collins and Wanamaker (2016), and Eli, Salisbury, and Shertzer (2016).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

methods consistently produce non-representative samples and high rates of false matches (or Type I errors), ranging from 15 to 37 percent, and higher rates of missed matches (or Type II errors), ranging from 63 to 79 percent, depending on the linking algorithm used. In addition, false matches do not occur at random; they are systematically predicted by baseline characteristics, suggesting that machine linking algorithms may introduce complicated forms of bias into analyses. To this point, Bailey et al.'s (2019) case study of linking birth certificates to the 1940 Census shows that—for the same set of records—prominent linking algorithms attenuate intergenerational income elasticity estimates by up to 20 percent. In that setting, Bailey et al. (2019) show that false links generate a critical part of this bias, and eliminating Type I errors from matches produces estimates that are indistinguishable from estimates of elasticities in data linked by hand.

This paper proposes two practical and complementary methods that aim to address these concerns and improve inference with linked data, regardless of the linking method used to create the data. First, we suggest using "validation variables"—variables that include information on the likelihood that a link is correct and information that was not used in the original linking process. Validation variables can help identify subsets of lower quality links for greater scrutiny. Second, we recommend creating *custom* weights for linked samples to improve their representativeness. These weights mitigate the biases that arise from low linking rates (high Type II errors) as well as the biases introduced by restricting samples with validation variables. We demonstrate how researchers can create these weights using inverse-propensity score reweighting. Although neither of these methods is new, they have rarely been applied individually or together in empirical papers using linked historical data.

This paper illustrates the value of these two strategies using the 1850–1930 Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS), a well-known and frequently used dataset in historical research. In section I, we review the linking and weighting methodology used to create the IPUMS-LRS dataset, emphasizing the components of its construction that are relevant to our later analysis. In section II, we demonstrate two examples of validation variables: name commonness (which can be used in almost all historical samples) and parent birthplace disagreement (which is specific to the IPUMS-LRS). Using a new hand-linked dataset, we show that both validation variables produce subsamples with fewer observations that human reviewers code as incorrect. In section III, we show how generating custom weights can improve the representativeness of the IPUMS-LRS, even relative to the provided weights available in the linked data. In contexts where weights are not available, analyzing representativeness and generating custom weights are even more important. The value of these strategies for the IPUMS-LRS—a highly curated dataset—demonstrates their potential to improve research with other linked datasets.

I. A Brief Overview of the IPUMS-LRS

The IPUMS-LRS consist of roughly 500,000 individuals for seven pairs of years: 1850–1880, 1860–1880, 1870–1880, 1880–1900, 1880–1910, 1880–1920, and 1880–1930 (the 1890 Census was excluded, because most of the original manuscripts were destroyed in a fire). These samples were created by the Minnesota Population Center (MPC), which linked

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the full-count 1880 Census (which was digitized by the Church of Jesus Christ of Latter-Day Saints) to the one-percent samples of the 1850, 1920 and 1930 Censuses, the 1.2 percent samples of 1860, 1870 and 1900 Censuses, and the 1.4 percent sample of the 1910 Census (Ruggles, 2006). Our analysis focuses on linked men from these samples.

To link men from one Census to the 1880 Census, the MPC produced a cross product of individuals across the two Censuses (e.g. 1850 and 1880). Using the Freely Extensible Biomedical Record Linkage software (FEBRL), the MPC kept each potential match from the cross product if the two observations had names that met a string similarity threshold, shared the same birthplace (state or country), and had ages that fell in a specified window.[3] They then trained a support vector machine (SVM) classifier using a set of hand-matched Census data, and applied the SVM to the non-training data in the cross product. Using these results, they kept all potential matches that had a predicted match probability that exceeded a match "quality" threshold and dropped all matches that had multiple potential links to 1880.

The MPC used two strategies to create representative samples. First, like many modern linking projects, they linked observations using theoretically time-invariant characteristics such as name, age, and birthplace rather than characteristics like place of residence, occupation, and family structure that may change over time. The use of these time-invariant characteristics limits selection bias in creation of links (Ruggles, 2006). For instance, linking individuals by using information on state of residence could make the sample much less geographically mobile than the population of interest to researchers.

Second, because different population subgroups might have different likelihoods of being linked, the MPC created weights to balance the representation of observed characteristics for the "linkable" population. Linkable men are those who were alive in both years and resided in the U.S. and could, therefore, be enumerated by the Census in both years. To determine the population of linkable men, the MPC took the final year Census and dropped men younger than the gap in years between Censuses (e.g. for 1880 in the 1860–1880 data, they drop everyone 20 years and younger). Because Census data do not specify when foreign-born men immigrated to the U.S., the MPC estimated the share of these foreign-born men who were present in the first year using life tables.

The MPC created weights for the linkable population using an iterative process. To start, the MPC assigned each observation a weight that was the inverse match rate for the relevant birth and race group (with denominators described by the linkable population). Then, they applied these weights and calculated weighted inverse match rates for other covariates, including relationships to head of household, individual birthplaces, 5-year age groups, and categories for size of place and occupation. They used these new inverse match rates to iteratively alter the weights until arriving at a final weight.

The IPUMS-LRS weights were designed to allow researchers to adjust the characteristics of the linked sample to resemble a simple random sample from the linkable population and, therefore, make inferences about this population's characteristics. The MPC is careful to

[3]FEBRL is a record linking software developed by the *ANU Data Mining Group* and the *Centre for Epidemiology and Research* in the *New South Wales Department of Health*. See Christen and Churches (2005) for more information.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

note the potential limitations of these weights, saying “researchers must decide whether the constructed weights are appropriate for their specific samples” (Goeken, Huynh, Lynch, & Vick, 2011).

II. Validation Variable as a Method to Improve Match Quality

The first method that we suggest for improving inference in historical linked data is to use one (or many) “validation variables.” A validation variable is a variable that is correlated with whether a link is correct but was not used *deterministically* in the linking process. Consequently, a researcher can condition on a validation variable to obtain a subsample with a smaller Type I error rate. Additionally, researchers can use validation variables to examine the links where the validation variable fails (i.e., links that are expected to have a higher Type I error rate) to investigate the performance of their algorithm by applying more scrutiny to a subset of more questionable records.

To motivate the purpose and practice of validation variables, we first lay out some basic theory. Consider the full dataset of links observed, L_i, and let whether or not a given link is correct be described by the following function:

$$C_i = f(Y(X_i), X_i, Z_i)$$

where C_i is an indicator variable equal to 1 if the link is correct, $Y(X_i)$ is the impact of the linking algorithm, which considers the information in X_i, and lastly Z_i, or variables that were not included in the linking process. Note that X_i impacts C_i through the process of the linking function and independently of the linking function, A validation variable, V_i, is a variable that satisfies the following properties:

1. $cov(C_i, V_i | L_i = 1) > 0$, and
2. $var(V_i | L_i = 1) \neq 0$

The first condition assures that the validation variable contains relevant information on whether the links are correct. The second condition ensures that the validation variable varies after conditioning on the observed links, which means that the validation variable is adding information beyond what is in the linking algorithm. If the validation variable agrees with all linking decisions, this condition will not be met. Note that a validation variable could be either a variable that was not included in the linking process (e.g. Z_i) or a variable that was included in the linking process but is used differently than it was in the linking process (e.g. X_i). Good validation variables may be more or less difficult to find depending on the linking setting, but our next section provides several examples hiding in plain sight.

A. Examples of Validation Variables

We use two different validation variables to demonstrate how these variables may reduce incorrect links: name commonness and disagreements in parents’ place of birth. We chose these two variables because the first is available in almost all historical linking contexts, but

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the second is specific to the IPUMS-LRS. Here we describe these variables and offer intuition for why they might be effective as validation variables.

Our first example of a validation variable, name commonness, is a broadly applicable validation variable. Name commonness is available in many linking situations and is intuitively correlated with whether a link is correct. More common names, for example "John Smith," have more possible matches than less common names. Therefore, measurement error in other features (age or birthplace) may lead an algorithm to select an incorrect match more frequently for more common names. Observations with uncommon names, on the other hand, have fewer potential matches available, so measurement error in other linking variables are less likely to cause an algorithm to choose an incorrect link. Bailey et al. (2019) provide empirical support for this intuition and show that eliminating more common names from the linking process significantly reduces incorrect links, or Type I errors, in some algorithms.

Some papers use name commonness restrictions in the matching process or as a robustness check, implicitly treating it as a validation variable. Abramitzky et al. (2012, 2014) use such a strategy, verifying that their results from their main dataset hold for links that have name-birth place combinations that are unique in a two-year age band. For our exercise, we similarly create a validation variable equal to 1 if a name-birthplace combination has only one observation within a two-year band of the individual's name.[4] The validation variable would be equal to zero for very common names and equal to one for less common names. As an example, the validation variable for "John Smith" born in Ohio aged 30 in the 1880 Census would be equal to zero, if multiple "John Smiths" ages 28 to 32 born in Ohio appeared in the 1880 Census.

Our second validation variable, parent birthplace disagreement, is specific to the IPUMS-LRS. When matching the 1850, 1860 and 1870 Census samples to the 1880 full count Census, the MPC did not include parent birthplaces in the linking process.[5] If parent birthplaces are correctly recorded for an individual in the Census, they should be consistent over time. Although some parent birthplaces may be measured with error (Goeken, Lynch, Lee, Wellington, & Magnuson, 2017), limiting attention to matches that agree in parent birthplaces would intuitively tend to select matches that are more likely to be correct.[6]

B. Examining the Effectiveness of Validation Variables

Bailey et al. (2019) recommend that researchers create training data (hand-links) for some of their observations in order to document the performance of their algorithm and similarly

[4]We are performing this restriction on the data *ex post* as we only have access to the finished IPUMS-LRS matches. However, Abramitzky et al. (2012, 2014) as described in Bailey et al. (2019), perform this restriction before engaging their matching algorithm.
[5]The MPC did use parental birthplace when linking the 1900, 1910, 1920 and 1930 Census samples to the 1880 full count Census.
[6]Data quality issues prior to 1880 are the reason that the MPC did not use this variable in the matching process for 1850–1870. For these years, parent birthplaces can only be inferred from individuals living at home with their parents. Furthermore, relationships within a household in those years are not listed by Census takers, and need to be inferred from the order in which individuals are listed in the Census and the ages of individuals. In Appendix I, we demonstrate that, although parent birthplace is clearly measured with error, patterns of parental birthplace disagreement between individuals living at home with their parents and those not living at home are similar in the years after 1880. Therefore, assuming that the imputed household relationships are accurate in the years prior to 1880, this evidence suggests that parent birthplace disagreement patterns for children living at home might be similar to parent birthplace disagreements for people who are not living at home with their parents.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

defend their choice of validation variables. We follow this advice and link a subsample of the 1850–1880 IPUMS-LRS to directly examine the quality of the IPUMS-LRS and the performance of our validation variables. To link these data, we randomly selected 653 IPUMS-LRS linked men who were aged 0 to 25 and living at home with their parents in 1850. An experienced group of genealogical linkers at the Family History and Technology Lab at Brigham Young University (BYU) then linked these observations by hand to the 1880 full count Census, without knowledge of the IPUMS-LRS links. The team at BYU used all the information available to the MPC and used additional information available to them through Ancestry.com and FamilySearch.org's databases. For the purpose of our exercise, we treat BYU's links as the truth and use these links to examine the performance of our validation variables.[7]

Table 1 summarizes the differences between the 1850–1880 IPUMS-LRS links and BYU's links.[8] The resulting share of links rejected by hand linkers is 10.0 percent, which is higher than the Type I error rate estimated by the MPC but is still low relative to machine-linked datasets analyzed in Bailey et al. (2019). Seventy percent of the differences come from cases where BYU determined that there was not enough data to reliably state a link. This outcome often occurred when a record had several possible matches, and genealogists were unsure about which possible match was correct. The remaining 30 percent of differences come from matches where BYU identified a link that disagreed with the IPUMS-LRS link.

Columns 3 and 4 of Table 2 examine the usefulness of our first validation variable, keeping only records that are unique for a given name, birthplace and age within a two-year band.[9] The first row under column 3 shows that 627 out of the 653 links considered by BYU make this cut using exact names. Given that many linking papers use phonetic cleaning to alter names for matching, columns 5 and 7 summarize the number of links that are unique in terms of NYSIIS or Soundex cleaned name and age combinations for the same age band.[10] The results show that requiring uniqueness of first and last name within the two-year age-radius lowers the rate of disagreement with hand linkers slightly, by 4 to 16 percent (0.4 to 1.6 percentage points on a base of 10.0 percentage points) depending on the name cleaning used. The drop in disagreements is likely small in part due to the fact that error rates are lower in the IPUMS-LRS data than in many other linked data. In other datasets, Bailey et al. (2019) show that a similar restriction in the Abramitzky et al. (2012, 2014) algorithm reduces rates of disagreement with hand linkers by as much as 10 percentage points.

Table 2 repeats this exercise using the validation variable for parent birthplace disagreement. As was the case for common names, genealogists are more likely to disagree with IPUMS-LRS links when parent birthplaces disagree. Dropping observations with a disagreement in father's birthplace drops the discrepancies with genealogists by 20 percent, a reduction of

[7]Appendix I provides more indirect evidence to demonstrate the relevance of parent birthplace disagreement as a validation variable without using hand-linked data.

[8]It is worth noting that hand-linked data are not "true" matches. Human error in matching may also produce false matches or fail to capture all 'true' matches. Given the dearth of longitudinal historical data, we have no direct test of the effectiveness of matching by hand.

[9]For completeness, we also considered other age bands, including a one-year and three-year age band in addition to the two-year age band in Table 1. The larger the band, the more observations tend to be dropped from consideration, but the Type I error rate also falls.

[10]Researchers use name cleaning algorithms to adjust exact names for errors in transcription, recording and changes in phonetic spelling. For more background on these algorithms, see Bailey et al. (2019).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2.0 points relative to a base of 10.0 percentage points. Dropping observations with a disagreement in mother's birthplace reduces disagreements by 18 percent, a reduction of 1.8 percentage points, and dropping observations with a disagreement in both mother and father birthplaces drops the error rate by 16 percent, a reduction of 1.6 percentage points.

If one takes records linked by genealogists as the truth, both sets of results suggest that conditioning on validation variables could reduce incorrect links. As a final test, we further probe the strength of the relationship between our validation variables and the determination by linkers that a link is incorrect. Specifically, we regress BYU's determination that a link is incorrect on our two validation variables as well as other data characteristics measuring a match's quality, including differences in age, own birthplace, and differences in recorded name using Jaro-Winkler similarity scores. This regression tests whether the validation variable contains information beyond that already present in these other features of the matches.

Table 3 shows the results from this regression using validation variables for name commonness and parent birthplace disagreement. Columns 1 and 4 show the unadjusted difference in error rates between observations that meet the validation variable and those that fail, demonstrating that the validation variables predict disagreements. Columns 2 and 5 show the correlation between the validation variables—after adjusting for the similarity of the individual's first and last name, difference in expected age, and own birthplace disagreement. Records with a higher similarity in first and last names or a smaller difference in expected age are negatively associated with BYU's determination that the link is incorrect, which is consistent with these record features partially determining matches. However, the inclusion of these covariates barely alters the partial correlation of the validation variables with link correctness. Similarly, the correlation between the validation variables with the likelihood of a link being judged incorrect by a reviewer is nearly unchanged by the inclusion in columns 3 and 6 of additional covariates, including indicator variables for living in an urban area, being in school, being born abroad, having a mother born abroad, having a father born abroad, residence on a farm, race, and Census region of residence. Across specifications, our validation variables remain a sizable and statistically significant predictor of the IPUMS-LRS link agreeing with hand-linked records.

Overall, our findings suggest the value of using a validation variable to diagnose and potentially increase link quality. Even though name commonness and discrepant parent birthplaces are noisy determinants of link quality, they appear to help diagnose errors and select higher quality links without having to examine the entirety of a dataset by hand.

Here we have only considered two validation variables, and other validation variables may be more or less effective in other settings depending on the matching process that produced the linked data. When selecting and implementing validation variables, researchers should consider the strength of the correlation of a validation variable with whether links are incorrect, and the effect of restricting on a validation variable on missed matches, called Type II errors. For instance, imposing restrictions on name commonness using exact names produces a limited decrease in Type I errors, but match rates drop non-trivially, resulting in increases in Type II errors. This limited decrease in Type I error likely reflects the fact that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the MPC considered some variation of name commonness in their linking. On the other hand, imposing restrictions on name commonness using NYSIIS- and Soundex-cleaned names produces a larger drop in Type I errors and also a larger increase in Type II errors, because these cleaned variables contain different information than that which was used in the algorithm. Thus, name commonness in our setting is more similar to an X_i variable, using the terminology of the linking example before: some part of this information was included in the MPC algorithm, but using a different part of the information still impacts incorrect link rates.

Parental birthplace was not explicitly used in the MPC's linking process for the 1850 Census data and is, therefore, more similar to the Z_i variable in our framework. We see a large drop in Type I errors from using this information as a validation variable, with the drop again potentially reflecting that the information from this validation variable was not captured by the other variables in X_i. Thus, selecting validation variables relies on knowledge of how the sample was initially constructed, and researchers will want to balance improvements in link quality from drops in Type I error rates against (sometimes) non-trivial increases in Type II errors.

III. Increasing the Representativeness of Linked Samples

Validation variables can help purge samples of lower quality links, but their effect on Type II errors raises concerns about sample representativeness. This concern motivates a second and complementary strategy for improving inference with linked samples: generating customized weights for the analytic sample. Generating custom weights may be important *even* in high quality linked data that contain weights (such as the IPUMS-LRS), as problems with representativeness may occur when researchers select certain subsamples for which weights do not balance covariates or because the relevant covariates were not used in the creation of weights. Consequently, weights may not create representative samples (Andrews & Oster, 2017; Angrist & Pischke, 2009; Caliendo & Kopeinig, 2008; Solon, Haider, & Wooldridge, 2015).

There are many ways to generate customized weights. Here, we document a simple, two-part procedure. First, we recommend that researchers document the degree to which their linked data are representative of the reference population using a regression test. Note, this investigation can be implemented in a manner similar to balance tests in randomized control trials (Duflo, Glennerster, & Kremer, 2007). Some papers currently do this check by reporting means of covariates of interest for the linked population and the reference population in the style of a covariate balance test. While this approach is valid, a regression provides a more concise *joint* test of representativeness. Second, we recommend that researchers construct and use custom weights using inverse propensity-score matching and report weighted results alongside unweighted results. While applying custom weighting may be especially important when using restrictions like validation variables, this strategy can also be used with nearly all historical linked data, as most historical linked samples have problems with non-representativeness.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Testing the representativeness of linked data requires establishing the relevant population for comparison—the reference population of interest. Consider a linking setting like IPUMS-LRS where links are between two Census years. The reference population would be the set of individuals who were alive and present in the U.S. in the earlier year and was still alive and present in the U.S. in the later year. That is, some of the observations present in the earlier year would not be linkable to the later year due to mortality and migration. Some of the observations in the later year would not be linkable to the earlier year if they had not been born yet, or if they had immigrated into the U.S. between the Censuses. Depending on the research questions, either year could be used for testing representativeness, so researchers would need to decide which is the relevant reference population for their analysis.

When testing representativeness in the IPUMS-LRS samples, we follow the MPC and identify the reference population as the individuals alive in the second Census: the 1880 full count Census for the 1850–1880, 1860–1880, and 1870–1880 samples. We examine the 1910 Census for the 1880–1910 sample, 1920 for the 1880–1920 sample, and 1930 for the 1880–1930 sample. Following the MPC, we identify as the reference population the potentially linkable individuals within this Census who would have been alive in the previous year by dropping all individuals who (given their reported age) would not have been alive in the earlier Census year (e.g. men younger than 30 in the 1880 Census in the 1850–1880 IPUMS-LRS). Unlike the MPC, we make two further restrictions on the sample of links to simplify our analysis. First, we drop from consideration all men born outside the U.S. The MPC included these individuals and created weights for them using life tables to account for the fact that some of the foreign-born men present in the later year may have immigrated into the U.S. between the two Census years. For simplicity, we avoid these adjustments by isolating attention to U.S. born men. Second, we drop all non-white men from our analyses. The MPC included these individuals, but given issues with counting African-American men in the 1850 and 1860 Censuses, we wanted to limit attention to men who could have been counted in the previous Census.[11] Thus, for our analysis, we restrict attention to matches within the population of white U.S.-born men present in the final year of the Census. Note that here we are not imposing any restrictions related to our validation variables–we are considering the representativeness of the IPUMS-LRS data overall.

A. A Simple Regression Test of Representativeness

Our representativeness test uses a simple regression method proposed in Bailey et al. (2019). Specifically, we recommend that researchers take the reference population data, create a dummy variable equal to 1 if an observation is linked, and then regress the dummy variable on a series of covariates describing the reference population. If using a linear probability model, we recommend researchers use Huber-White standard errors to account for the fact that errors of a linear probability model are heteroskedastic (Huber 1967, White 1980). Our representativeness test-statistic is a heteroscedasticity-robust Wald test of joint significance of the covariates. Under the null hypothesis of representativeness of the linked sample, there should be no relationship between the covariates and the likelihood an observation is linked.

[11] In 1850 and 1860, African-American slaves were enumerated separately under a slave schedule.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The advantage of this test over variable-by-variable balance test of means is that it accounts for the correlations among covariates and the *joint relationship* of the group of covariates with the likelihood of being linked, aggregating all information in the relevant covariates into a single test statistic. Furthermore, the magnitudes of the regression coefficients conveniently quantify which characteristics are more or less likely to result in a linked observation after controlling for other record characteristics. Note, however, that this technique is only a diagnostic test of the null hypothesis of representativeness, and rejecting the null hypothesis is not an indication that inference estimates are necessarily biased for two reasons. First, statistical significance of differences in covariates does not imply scientific significance, as magnitude of the bias may be slight (McCloskey, 2005). Moreover, if the relationship of interest (e.g. job mobility) is homogeneous for all groups in the population, selecting a non-representative sample would not bias estimates.

Table 4 summarizes the results of the representativeness tests for all of the IPUMS-LRS samples. Since the MPC provides weights to adjust for the non-representativeness of linked data, we compare the sample characteristics using both unweighted and weighted data. The first two columns present the results of a regression of a binary dependent variable (=1 if the observation is linked) on a subset of the covariates that the MPC used to construct their weights. These include 11 binary variables for relationship of an individual to the head of household (e.g. spouse, child, etc.); eight binary variables for birthplaces by region (e.g. Northeast, Mid-Atlantic); and up to 14 binary variables for the size of the place the individual currently lives in (see table notes for details). For the unweighted results in column 1, the p-values show that the Wald test easily rejects the null-hypothesis of representativeness. After we apply IPUMS-LRS weights in column 2, we fail to reject representativeness at the 5-percent level in this subset of characteristics for three samples, which suggest the IPUMS-LRS weights largely work as intended. However, for the other four samples, applying the weights results in p-values that reject representativeness at conventional levels of significance.

Columns 4 and 5 consider the entirety of the covariates that the MPC used in their weighting procedure (all previous variables from columns 1 and 2 as well as binary variables for five-year age groups and four categories for occupations) both with and without weights. Unsurprisingly, we reject representativeness in the unweighted samples at the 1-percent level in all cases. After we apply IPUMS-LRS weights in column 5, we fail to reject representativeness for this full set of weighting covariates at the 5-percent level for the 1850–1880, 1870–1880, 1880–1910, and 1880–1920 samples.

Finally, columns 7 and 8 consider all variables that were used by the MPC to calculate weights *and additional variables that were not*. These additional variables include binary variables for whether or not a man lives with his parents, whether that man's parents were born in the U.S., the region of the country that man lives in, his marital status, farm status, the number of co-resident siblings, and an indicator variable for whether or not an individual lives in the same state as birth. In both weighted and unweighted samples across all years, the p-values show we reject representativeness at the 1-percent level for each sample. This result is less surprising, as the IPUMS-LRS weights might only be expected to achieve balance in covariates used to create these samples.[12] Similarly, in other settings, weights

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

may not create representative samples for every research question or purpose and may not work well when isolating attention to specific subgroups (Caliendo and Kopeinig 2008, Angrist and Pischke 2009, Solon et al. 2015, Andrews and Oster 2017).

Looking beneath the test of *statistical* significance, this lack of representativeness may have consequences for inference. For brevity, Table 5 presents a subset of estimates for the 1860–1880 sample from the regressions underlying Table 4. We report the full set of regression results for all samples in Appendix IV for the interested reader. As a complement to these findings, Table 6 presents more standard mean comparisons for a subset of covariates in the 1860–1880 sample (the full set of mean comparisons for all samples are reported in Appendix III). The IPUMS-LRS weights improve representativeness with respect to some variables, especially those used in the construction of the weights, including age categories, size of place categories and current location of residence categories. As one might expect, however, the weights do little to balance the representation of characteristics that were not included in their construction. Moreover, some categories that were *included* in the weighting process remain unbalanced. For example, some IPUMS-LRS samples after applying weights over-represent heads of household while others underrepresent them. These patterns could be important for inference for a variety of research questions on family structure, particularly those relating to structure of intergenerational co-residing families (Ruggles, 2011).

In terms of migration and nativity outcomes, the weighted IPUMS-LRS often produce unrepresentative samples of Census region of residence and parental birthplaces. The weighted IPUMS-LRS samples over-represent individuals from the Northeast in five of six samples, including the 1860–1880 data reported in Table 4. All samples underrepresent U.S.-born children with foreign-born parents—a finding that could affect inferences about U.S. immigration from Asia (Hatton, 2011) and Europe (Abramitzky et al., 2012). Furthermore, all IPUMS-LRS samples, including the 1860–1880 sample shown in Tables 4 and 5, over-represent individuals living in the same state as where they were born. Living in the same state as birth increases the probability of being linked among U.S.-born white men by 4 to 6 percentage points across all samples *after* applying IPUMS-LRS weights. This suggests that the linked IPUMS-LRS sample appears less geographically mobile, which could affect inferences about intergenerational occupational mobility, occupation selection, and generational household structure.

Thus, overall, even in datasets like the IPUMS-LRS that have weights that work as intended for adjusting the covariates that were included in the weighting process, these weights may not be effective when considering different subsamples of the data, or other covariates that were not included in the weighting process. This lack of representativeness may create biases in inference from over or under-representation of specific groups if heterogeneous effects are present (Bailey et al. 2019).

[12]It is worth noting that these findings hold up in more traditional t-tests as well. Notably, we reject the null hypothesis of equality of means among the variables not included by the MPC roughly 63 percent of the time across all samples. See Appendix III for the full set of results. Note also that if the weights addressed all issues with representativeness of the data that there should not be these issues with other variables.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. Creating Weights Customized to a Sample or Question of Interest

If non-representativeness or imbalance in certain characteristics is a concern, researchers should report weighted results that adjust for that imbalance in addition to traditional unweighted estimates. If weights are not available, or the weights do not adjust sufficiently for non-representativeness, then researchers may construct their own using an application-specific inverse propensity (IP) score reweighting technique.

This approach requires that (1) the propensity of being linked is properly specified and can be consistently estimated (often described as unconfoundedness assumption) and that (2) the distribution of the propensity of being linked spans the same support as the reference population (often described as a common support assumption). It is impossible to test assumption (1) directly and it could be violated in a linking situation where the probability of being linked depends on unobservable features of an observation that are correlated with the variables included in the weight estimation process. However, theory can guide the selection of variables for (1). Assumption (2), on the other hand, can be tested directly by examining the estimated link propensities of linked records and the reference population.

This method can be implemented using the following steps:

1. Append the data for the linked sample to the population which the researcher wants the reweighted sample to represent.
2. Create a dependent variable, L_i, equal to 1 for each observation, i, in the linked sample and 0 for each observation in the reference population. Using this dependent variable, estimate a probit model on covariates of record characteristics, $\boldsymbol{X}_i$ (for instance, the variables used in columns 7–9 in Table 4).
3. Using the results from the probit, predict the conditional probability of being linked, $P_{\square}(L_i = 1|\boldsymbol{X_i})$, for each observation.
4. To check assumption (2) regarding common support, plot the probabilities of being linked for the linked and unlinked observations. The overlap in the two distributions provides information on which individuals can be compared. Also, Crump, Hotz, Imbens, and Mitnik (2009) recommend trimming extreme probabilities, which is another easy-to-implement strategy for improving inference.
5. Using the predicted probabilities, researchers may calculate weights as $W_i = \left(1 - P_{\square}(L_i = 1|\boldsymbol{X_i})\right)/P(L_i = 1|\boldsymbol{X_i}) * q/(1-q)$, where q is the share of records that are linked. If a certain set of characteristics is underrepresented in the linked sample relative to the population of interest, this weight will increase the influence of this particular observation. The second component normalizes these probabilities to fit the size of the linked and unlinked samples.

We implement this procedure for each sample using the covariates in columns 7–9 in Table 4 and find evidence that the common support assumption holds. Intuitively, the common support assumption requires that there is sufficient overlap in the characteristics of links and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the reference population, as summarized by the propensity score, so that the former can be reweighted to look like the latter.

Applying these weights to the IPUMS-LRS samples makes a meaningful difference in our representativeness calculations. Although only a handful of means in Appendix III remain statistically significant after reweighting, column 3 of Table 4 shows that coefficient estimates from the regression are very close to zero for a large number of covariates in the 1860–1880 IPUMS-LRS. This finding is substantively different from the unweighted (column 1) and IPUMS-LRS weighted results (column 2). Moreover, columns 3, 6 and 9 of Table 4 show that we fail to reject representativeness for all of the IPUMS-LRS samples (p-values very close to one) after applying IP-weights for these covariates of interest. Of course, if we omit certain variables when constructing the IP-weights and then test for representativeness in these same variables after applying IP-weights, we also tend to reject representativeness, just as we did when considering the MPC's weights with variables they had not included in the reweighting process. It is important, therefore, that researchers specify the propensity score equation in step 3 with covariates to achieve balance in characteristics relevant for answering a particular research question.

Although we have only been considering the overall representativeness of the IPUMS-LRS data, we find the same results regarding lack of representativeness of linked data and effectiveness of IP weights after imposing restrictions using our two validation variables. We omit those results here for brevity.

Lastly, it is important to note that, even though this reweighting procedure produces a sample very similar in *observed* characteristics, the resulting data may still be unbalanced in terms of *unobserved* characteristics, and reweighting will only accurately address bias from non-representativeness if the unconfoundedness assumption described earlier holds. That is, reweighting's effectiveness ultimately depends on the assumptions specified earlier, although the hope is that reweighting at least mitigates the problem of non-representativeness of linked data (DiNardo, Fortin, & Lemieux, 1996; Heckman, 1979).

IV Recommendations and Conclusions

Many important questions relate to how individuals, families, and communities changed over time, and new linked samples are critical in facilitating new research on these questions. As documented in Bailey et al. (2019), measurement error induced by linking algorithms may have substantial implications for inference. In light of this evidence, this paper suggests two complementary strategies to improve inference with linked samples.

First, we recommend using a validation variable that is *correlated* with link quality and not deterministically used in the linking process in order to improve inferences. These two conditions imply that the validation variable will contain additional information about link quality. These variables allow researchers to perform robustness tests by purging links more likely to be incorrect from their analysis samples without the high cost of hand linkage. For our case study using the 1850–1880 IPUMS-LRS, we use name uniqueness and parental birthplaces to identify a set of links more likely to be correct. Although both of these

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

variables are noisy indicators of linking errors, regression evidence demonstrates that name commonness and discordance in both parents' birthplaces are nevertheless powerful predictors of incorrect links—even in a high quality sample like the IPUMS-LRS. Purging samples of links with common names reduces the error rate in the pre-1880 IPUMS by up to 15 percent, and dropping observations with discordant parent birthplaces, reduces the error rate by up to 20 percent. We have only examined two examples of variables but other contexts may lead to other potential validation variables.

Limiting samples by purging potentially false links may also increase problems with non-representativeness, an issue with almost all linked data. This problem leads us to suggest a second, complementary strategy for improving inferences with linked records. Like many surveys and historical samples, the IPUMS-LRS (even with weights) are not generally representative of the reference population of potentially linkable individuals. However, applications of inverse probability weighting can substantially improve representativeness. To this end, we describe a simple inverse propensity score reweighting approach similar to that proposed by DiNardo et al. (1996) and demonstrate its effectiveness for the IPUMS-LRS. This method is easily adaptable to various applications and will generally produce representative samples catered to specific research objectives under the assumptions we specify. A close examination of the value of these weights also informs researchers about where more time-intensive genealogical or clerical review methods may increase the representation of hard-to-link groups. Used in combination with validation variables, custom reweighting may help improve inference with linked data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This project was generously supported by the National Science Foundation under grant SMA 1539228, the National Institute on Aging under grant R21 AG05691201 and R01 AG057704, the University of Michigan Population Studies Center Small Grants under grant R24 HD041028, the Michigan Center for the Demography of Aging under grant P30 AG012846-21, the University of Michigan Associate Professor Fund, and the Michigan Institute on Research and Teaching in Economics (MITRE). We gratefully acknowledge the use of the services and facilities of the Population Studies Center at the University of Michigan under grant R24 HD041028. During work on this project, Cole was supported by the NICHD under grant T32 HD000733 as a UM Population Studies Center Trainee. We are grateful to George Alter, Trent Alexander, Katie Genadek, Alfia Karimova, Maggie Levenstein, Evan Roberts, and Steve Ruggles for their helpful suggestions and comments. We are also grateful to Sarah Anderson, Garrett Anstreicher, Ali Doxey, Meizi Li, and Mike Ricks for their many contributions to the LIFE-M project and assistance with this analysis.

V. References

Abramitzky R, Platt Boustan L, & Eriksson K (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. [PubMed: 26594052]

Abramitzky R, Platt Boustan L, & Eriksson K (2013). Have the Poor Always been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration. Journal of Development Economics, 102, 2–14. [PubMed: 26609192]

Abramitzky R, Platt Boustan L, & Eriksson K (2014). A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. Journal of Political Economy, 122(3), 467–506. [PubMed: 26609186]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Aizer A, Eli S, Ferrie J, & Lleras-Muney A (2016). The Long Term Impact of Cash Transfers to Poor Families. American Economic Review, 106(4), 935–971. [PubMed: 28713169]

Alsan M, & Goldin C (2015). Watersheds in Infant Mortality: The Role of Effective Water and Sewage Infrastructure, 1880 to 1915. NBER Working Paper 21263.

Andrews I, & Oster E (2017). Weighting for External Validity. NBER Working Paper 23826.

Angrist JD, & Pischke J-S (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press.

Bailey MJ, Anderson S, Karimova A, & Massey CG (2016). Creating LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database. University of Michigan Working Paper. Retrieved from http://sites.lsa.umich.edu/life-m/

Bailey MJ, Cole C, Henderson M, & Massey CG (2019). How Well Do Automated Linking Methods Perform in Historical Samples? Evidence from New Ground Truth. NBER Working Paper 24019.

Bleakley H (2007). Disease and Development Evidence from Hookworm Eradication in the American South. Quarterly Journal of Economics, 122(1), 73–117. [PubMed: 24146438]

Bleakley H, & Ferrie J (2014). Land Opening on the Georgia Frontier and the Coase Theorem in the Short- and Long-Run. University of Michigan Working Papers. Retrieved from http://www-personal.umich.edu/~hoytb/Bleakley_Ferrie_Farmsize.pdf

Bleakley H, & Ferrie J (2016). Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations. Quarterly Journal of Economics, 131(3), 1455–1495. [PubMed: 28529385]

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

Boustan LP, Kahn ME, & Rhode PW (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.

Caliendo M, & Kopeinig S (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys, 22(1), 31–72.

Christen P, & Churches T (2005). Febrl - Freely extensible biomedical record linkage. Release 0.3.1. Retrieved from http://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/manual.html

Clay K, Lewis J, & Severnini E (2016). Canary in a Cola Mine: Infant Mortality, Property Values, and Tradeoffs Associated with Mid-20th Century Air Pollution. 1–61.

Collins WJ, & Wanamaker MH (2016). Up from Slavery? African American Intergenerational Economic Mobility Since 1880. University of Michigan Economic History Seminar. Retrieved from http://www.nber.org/papers/w23395.pdf

Crump RK, Hotz V Joseph Imbens, Guido W, & Mitnik OA (2009). Dealing with Limited Overlap in Estimation of Average Treatment Effects. Biometrika, 96(1), 187–199.

Cutler DM, & Miller G (2005). The Role of Public Health Improvements in Health Advances: The 20th Century United States. Demography, 42(1), 1–22. [PubMed: 15782893]

DiNardo J, Fortin NM, & Lemieux T (1996). Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach. Econometrica, 64(5), 1001–1044.

Duflo E, Glennerster R, & Kremer M (2007). Chapter 61 Using Randomization in Development Economics Research: A Toolkit In Schultz TP & Strauss JA (Eds.), Handbook of Development Economics (Vol. 4, pp. 3895–3962): Elsevier.

Eli S, Salisbury L, & Shertzer A (2016). Migration in Response to Civil Conflict: Evidence from the Border of the American Civil War. NBER Working Paper 22591.

Goeken R, Huynh L, Lynch TA, & Vick R (2011). New Methods of Census Record Linking. Historical Methods, 44(1), 7–14. [PubMed: 21566706]

Goeken R, Lynch T, Lee YN, Wellington J, & Magnuson D (2017). Evaluating the Accuracy of Linked U. S. Census Data: A Household Approach. Retrieved from https://pop.umn.edu/sites/pop.umn.edu/files/1.working_paper17_0.pdf

Hatton T (2011). The Cliometrics of International Migration: A Survey In Oxley L (Ed.), Economics and History: Surveys in Cliometrics (pp. 187–216). London: Wiley-Blackwell.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Heckman JJ (1979). Sample Selection Bias as a Specification Error. Econometrica, 47(1), 153–161. doi:10.2307/1912352

Hornbeck R, & Naidu S (2014). When the Levee Breaks: Black Migration and Economic Development in the American South. American Economic Review, 104(3), 963–990.

Mazumder B, & Aaronson D (2011). The Impact of Rosenwald Schools on Black Achievement. Journal of Political Economy, 119(5), 821–888.

McCloskey D (2005). The Trouble with Mathematics and Statistics in Economics. History of Economic Ideas XIII, 3, 85–102.

Mill R (2013). Record Linkage across Historical Datasets. Inequality and Discrimination in Historical and Modern Labor Markets. Stanford University Dissertation. Retrieved from https://searchworks.stanford.edu/view/10232417

Mill R, & Stein LC (2016). Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America. Retrieved from http://www.public.asu.edu/~lstein2/research/mill-stein-skincolor.pdf

Rhode P, & Olmstead AL (2015). Arresting Contagion: Science, Policy and Conflicts over Animal Disease Control. Cambridge, MA: Harvard Univ. Press.

Ruggles S (2006). Linking Historical Censuses: A New Approach. History and Computing, 14(1–2), 213–224.

Ruggles S (2011). Intergenerational Coredisence and Family Transitions in the United States, 1850–1880. Journal of Marriage and the Family, 73(1), 138–148. [PubMed: 22039309]

Ruggles S, Genadek K, Grover J, & Sobek M (2015). Integrated Public Use Microdata Series (Version 6.0) [Machine-Readable database] In Minnesota U. o. (Ed.). Minneapolis: University of Minnesota.

Solon G, Haider SJ, & Wooldridge JM (2015). What are We Weighting For? Journal of Human Resources, 50(2), 301–316.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Name Uniqueness in IPUMS-LRS and Linking Errors from Comparison to Genealogically Linked Sample

	All IPUMS Observations		Uniqueness in Exact Name in Two Year Radius		Uniqueness in NYSIIS Name in Two Year Radius		Uniqueness in Soundex Name in Two Year Radius	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Total IPUMS Observations	653	100.00%	627	100.00%	573	100.00%	489	100.00%
Total IPUMS-LRS Correct (Matched by BYU, Link is Same)	588	90.05%	567	90.43%	525	91.62%	446	91.21%
Total IPUMS-LRS Incorrect	65	9.95%	60	9.57%	48	8.38%	43	8.79%
A) Matched by BYU, Link Different	19	2.91%	16	2.55%	13	2.27%	12	2.45%
B) Not Matched by BYU	46	7.04%	44	7.02%	35	6.11%	31	6.34%

Notes: This table uses a hand-linked sample of the 1850–1880 censuses produced by the BYU Family History and Technology Lab. When IPUMS-LRS agrees with BYU, we call the link “correct.” When IPUMS-LRS differs from BYU, we call the link “incorrect.”

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2.

Parent Birthplace Disagreements in IPUMS-LRS and Linking Errors from Comparison to Genealogically Linked Sample

	All IPUMS Observations		Observations without Father Birthplace Disagreement		Observations without Mother Birthplace Disagreement		Observations without Father and Mother Birthplace Disagreement	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Total IPUMS Observations	653	100.00%	514	100.00%	475	100.00%	547	100.00%
Total IPUMS-LRS Correct (Matched by BYU, Link is Same)	588	90.05%	473	92.02%	436	91.79%	501	91.59%
Total IPUMS-LRS Incorrect	65	9.95%	41	7.98%	39	8.21%	46	8.41%
A) Matched by BYU, Link Different	19	2.91%	13	2.53%	13	2.74%	15	2.74%
B) Not Matched by BYU	46	7.04%	28	5.45%	26	5.47%	31	5.67%

Notes: This table uses a hand-linked sample of the 1850–1880 censuses produced by the BYU Family History and Technology Lab. When IPUMS-LRS agrees with BYU, we call the link “correct.” When IPUMS-LRS differs from BYU, we call the link “incorrect.”

Hist Methods. Author manuscript; available in PMC 2021 January 01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Hist Methods. Author manuscript; available in PMC 2021 January 01.

Table 3.

Regression-Adjusted Measurement of Validation Variable Correlation with Wrong Matches

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	**1=Incorrect Link**					
	IPUMS-LRS 1850–1880 Male Links					
Uniqueness of NYSIIS Name and Birthplace in 2-Year Age Radius	−0.13**	−0.12**	−0.11**			
	(0.047)	(0.046)	(0.048)			
No Disagrement in Both Father and Mother Birthplaces				−0.13**	−0.13**	−0.12**
				(0.050)	(0.051)	(0.051)
Jaro Winkler – Own Last Name		−1.08*	−1.04*		−1.24**	−1.09*
		(0.581)	(0.596)		(0.588)	(0.590)
Jaro Winkler – Own First Name		−0.23*	−0.21		−0.27**	−0.24*
		(0.127)	(0.137)		(0.130)	(0.137)
Difference in Expected Age		0.06**	0.06**		0.06**	0.06**
		(0.024)	(0.025)		(0.025)	(0.025)
Own Birthplaces Disagree		0.14	0.21		0.14	0.22
		(0.226)	(0.229)		(0.228)	(0.230)
Constant	0.21***	1.46**	1.18*	0.21***	1.66**	1.09*
	(0.046)	(0.600)	(0.625)	(0.049)	(0.628)	(0.615)
Additional covariates	N	N	Y	N	N	Y
R-squared	0.006	0.010	0.022	0.011	0.051	0.088

Notes: The regression results are obtained from regressing a binary dependent variable (=1 if a record incorrect, 0 if the link is correct) on the indicated covariates (N = 618). In columns 1–3, additional covariates are measured in the 1850 Census and include an indicator variable for living in an urban area, being in school, whether born abroad, mother born abroad, father born abroad, and farm status, a variable indicating whether the veteran was white, and region fixed effects. In columns 4–6, there are 1,421 observations and the additional covariates include whether the veteran was white and region fixed effects. Heterskedasticity-robust standard errors are reported beneath each estimate, and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4.

Regression Test of Representativeness of IPUMS-LRS of White Men Born in the U.S.

	Restricted IPUMS Covariates			All IPUMS Covariates			Al IPUMS Covariates and Other Covariates		
Years Matched	**Unweighted**	**IPUMS Weighted**	**IP Weighted**	**Unweighted**	**IPUMS Weighted**	**IP Weighted**	**Unweighted**	**Weighted**	**IP Weighted**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
1850–1880	1287	13678	5	1529	15043	14	1714	1094	20
	0.00	*0.00*	*1.00*	*0.00*	*0.00*	*1.00*	*0.00*	*0.00*	*1.00*
1860–1880	1819	92	8	2149	100	19	2503	332	24
	0.00	*0.00*	*1.00*	*0.00*	*0.00*	*1.00*	*0.00*	*0.00*	*1.00*
1870–1880	3122	60	7	3560	71	22	4301	540	33
	0.00	*0.00*	*1.00*	*0.00*	*0.03*	*1.00*	*0.00*	*0.00*	*1.00*
1880–1900	1970	25	2	2603	79	9	3275	355	13
	0.00	*0.73*	*1.00*	*0.00*	*0.00*	*1.00*	*0.00*	*0.00*	*1.00*
1880–1910	1512	46	18	1939	52	20	2355	293	25
	0.00	*0.03*	*0.95*	*0.00*	*0.26*	*1.00*	*0.00*	*0.00*	*1.00*
1880–1920	961	44	23	1190	53	27	1390	204	29
	0.00	*0.05*	*0.80*	*0.00*	*0.17*	*0.98*	*0.00*	*0.00*	*1.00*
1880–1930	772	130	18	937	335	18	1070	455	18
	0.00	*0.00*	*0.96*	*0.00*	*0.00*	*1.00*	*0.00*	*0.00*	*1.00*
Covariates included	C,B,H	C,B,H	C,B,H	C,B,H,O,A	C,B,H,O,A	C,B,H,O,A	C,B,H,O,A,R,C,X	C,B,H,O,A,R,C,X	C,B,H,O,A,R,C,X

Notes: Each estimate is a heteroscedasticity-robust Wald-test from a separate regression of a binary dependent variable (=1 for linked record) for samples described in the text. the associated p-values is printed beneath in italics. “IP Weights” refers to inverse-propensity score weighted estimates. Covariate abbreviations are as follows. *C* denotes dummy variables for size of local city (under 1,000 or unincorporated; 1,000 to 2,499; 2,500 to 3,999; 4,000 to 4,999; 5,000 to 9,999; 10,000 to 24,999; 25,000 to 49,999; 50,000 to 74,999; 75,000 to 99,999, 100,000 to 199,999; 200,000 to 299,999; 300,000 to 499,999; 500,000 to 599,999; 600,000 to 749,999; 750,000 to 999,999; 1 million to 1.99 million and 2 million and up). *B* denotes dummy variables for birth location (Northeast, Mid-Atlantic Region, East North Central Region, West North Central Region, South Atlantic Region, East South Central Region, West South Central Region, Mountain Region, and born outside U.S.). *H* denotes dummy variables for relationship of individual to household head (head/householder, spouse, child, child-in-law, parent, parent-in-law, sibling, sibling-in-law, grandchild, other relatives, parent friend or visitor). *O* denotes dummy variables for occupation (white collar occupation, farming occupation, semi-skilled occupation, unskilled occupation), and *A* denotes age category variables (dummy variables for five-year categories of ages). *R* denotes dummies for region of residence (Northeast, Midwest, West), *E* is a set of dummy variables for whether an individual lives with his mother, lives with his father, or lives with both parents. *X* is a set of dummy variables for whether one’s father was born abroad, mother was born abroad, marital status, or farm status and whether they were living in the same state as birth. It also includes number of siblings in the household. Weights for BCM are described in text

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5.

Regression Estimates of Representativeness of 1860–1880 IPUMS-LRS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables Included	**Subset of Covariates**	**Subset of Covariates**	**All Weighting Covariates**	**All Weighting Covariates**	**All Covariates**	**All Covariates**	**All Covariates**
Weights	**None**	**IPUMS-LRS**	**None**	**IPUMS-LRS**	**None**	**IPUMS-LRS**	**IP**
Born in North East	0.05***	−0.05	0.04**	−0.06	0.01	−0.10**	−0.01
	(0.016)	(0.045)	(0.016)	(0.045)	(0.016)	(0.045)	(0.043)
Born in Mid-Atlantic Region	−0.06***	−0.05	−0.06***	−0.05	−0.08***	−0.09**	−0.01
	(0.015)	(0.044)	(0.015)	(0.044)	(0.016)	(0.043)	(0.042)
Born in East North Central Region	−0.04**	−0.04	−0.04**	−0.05	−0.05***	−0.06	−0.01
	(0.015)	(0.043)	(0.016)	(0.044)	(0.016)	(0.042)	(0.040)
Born in West North Central Region	−0.02	−0.04	−0.02	−0.04	−0.03**	−0.05	−0.01
	(0.016)	(0.044)	(0.016)	(0.044)	(0.016)	(0.042)	(0.041)
Born in South Atlantic Region	−0.02	−0.07	−0.02	−0.08*	−0.03*	−0.08*	−0.01
	(0.016)	(0.044)	(0.016)	(0.045)	(0.016)	(0.044)	(0.043)
Born in East South Central Region	−0.05***	−0.06	−0.05***	−0.07	−0.06***	−0.07*	−0.01
	(0.016)	(0.044)	(0.016)	(0.045)	(0.016)	(0.044)	(0.042)
Born in West South Central Region	−0.01	−0.03	−0.01	−0.04	−0.02	−0.04	−0.01
	(0.017)	(0.047)	(0.017)	(0.047)	(0.018)	(0.046)	(0.044)
Born in Mountain Region	−0.06***	−0.03	−0.05**	−0.03	−0.05**	−0.02	−0.02
	(0.021)	(0.091)	(0.021)	(0.092)	(0.021)	(0.094)	(0.089)
Head/Householder	0.05***	−0.02	0.04***	−0.03***	0.03***	−0.06***	0.00
	(0.003)	(0.012)	(0.003)	(0.013)	(0.004)	(0.015)	(0.015)
Spouse	−0.01	−0.32***	−0.02	−0.33***	−0.03	−0.35***	−0.05
	(0.026)	(0.092)	(0.025)	(0.091)	(0.025)	(0.090)	(0.147)
Child	0.04***	0.02	0.05***	0.02	0.03***	−0.02	0.01
	(0.003)	(0.014)	(0.003)	(0.014)	(0.007)	(0.024)	(0.024)
Child-in-law	0.02**	−0.06	0.02*	−0.06*	0.01	−0.09**	0.01
	(0.010)	(0.036)	(0.010)	(0.037)	(0.011)	(0.038)	(0.038)
Parent	0.09***	−0.04	0.01	−0.06	0.01	−0.07**	0.01
	(0.014)	(0.033)	(0.015)	(0.035)	(0.015)	(0.035)	(0.034)
Parent-in-Law	0.08***	−0.04	0.01	−0.06	0.00	−0.08*	0.01
	(0.018)	(0.042)	(0.018)	(0.043)	(0.018)	(0.043)	(0.044)
Sibling	0.02***	−0.06**	0.02***	−0.07**	0.02**	−0.07***	0.01
	(0.007)	(0.027)	(0.007)	(0.027)	(0.008)	(0.027)	(0.027)
Sibling-in-Law	0.02*	−0.08*	0.02*	−0.08*	0.02	−0.09**	0.00
	(0.011)	(0.041)	(0.011)	(0.041)	(0.011)	(0.041)	(0.041)
Lives with Mother					−0.00	−0.00	0.00

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables Included	**Subset of Covariates**	**Subset of Covariates**	**All Weighting Covariates**	**All Weighting Covariates**	**All Covariates**	**All Covariates**	**All Covariates**
Weights	**None**	**IPUMS-LRS**	**None**	**IPUMS-LRS**	**None**	**IPUMS-LRS**	**IP**
					(0.006)	(0.018)	(0.018)
Lives with Father					0.00	0.03	0.01
					(0.009)	(0.027)	(0.027)
Lives with Both					0.02	0.03	−0.02
Parents					(0.009)	(0.028)	(0.028)
Father: Born Abroad					−0.02***	−0.05***	0.00
					(0.005)	(0.018)	(0.018)
Mother: Born Abroad					−0.02***	−0.08***	−0.00
					(0.005)	(0.019)	(0.019)
Lives in Northeast					0.02***	0.05**	−0.00
					(0.005)	(0.023)	(0.023)
Lives in Midwest					0.01***	0.02	−0.00
					(0.004)	(0.019)	(0.020)
Lives in West					−0.00	−0.00	−0.00
					(0.006)	(0.036)	(0.037)
Number of Siblings					−0.00	0.00	−0.00
					(0.001)	(0.003)	(0.003)
Lives in Same State as Birth					0.02***	0.05***	−0.00
					(0.002)	(0.010)	(0.010)
Constant	0.08***	0.56***	0.18***	0.51***	0.18***	0.51***	0.46***
	(0.019)	(0.053)	(0.053)	(0.102)	(0.053)	(0.110)	(0.112)
Observations	97,123	97,123	97,123	97,123	97,123	97,123	97,123
R-squared	0.017	0.002	0.022	0.003	0.025	0.012	0.000
Wald Statistic	1631	59.2	2029	67.7	2400	255	5.1
Prob > F	0.00	0.00	0.00	0.03	0.00	0.00	1

Notes: The regression results are obtained from regressing a binary dependent variable (=1 if a record is linked, 0 if in the linkable population) on the indicated covariates (N = 8,673,750). Standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6.

T-Tests of Means in the 1860–1880 IPUMS-LRS and the Linkable Population

	Unweighted	IPUMS-LRS Weights	IP Weights
	(1)	(2)	(3)
Age	3.787***	−0.088	0.161
	(0.173)	(0.166)	(0.164)
Born in North East	0.107***	0.001	0.001
	(0.004)	(0.007)	(0.007)
Born in Mid-Atlantic Region	−0.093***	0.002	−0.005
	(0.004)	(0.008)	(0.008)
Born in East North Central Region	−0.021***	0.008	0.001
	(0.004)	(0.007)	(0.007)
Born in West North Central Region	−0.000	0.003	0.000
	(0.002)	(0.003)	(0.002)
Born in South Atlantic Region	0.024***	−0.012*	0.002
	(0.004)	(0.006)	(0.006)
Born in East South Central Region	−0.018***	−0.006	0.000
	(0.003)	(0.005)	(0.005)
Born in West South Central Region	0.003*	0.002	0.000
	(0.002)	(0.002)	(0.002)
Born in Mountain Region	−0.001*	0.000	0.000
	(0.001)	(0.000)	(0.001)
Born in Pacific Region	0.000	0.001	0.000
	(0.001)	(0.001)	(0.001)
Relationship to Head: Head/Householder	0.066***	−0.021***	0.001
	(0.005)	(0.006)	(0.006)
Relationship to Head: Spouse	−0.000**	−0.001***	−0.000
	(0.000)	(0.000)	(0.000)
Relationship to Head: Child	−0.003	0.023***	−0.000
	(0.004)	(0.005)	(0.005)
Relationship to Head: Child-in-law	−0.002**	−0.001	0.000
	(0.001)	(0.001)	(0.001)
Relationship to Head: Parent	0.004***	−0.001	0.000
	(0.001)	(0.001)	(0.001)
Relationship to Head: Parent-in-law	0.002**	−0.000	0.000
	(0.001)	(0.001)	(0.001)
Relationship to Head: Sibling	−0.003***	−0.003**	0.000
	(0.001)	(0.001)	(0.002)
Relationship to Head: Sibling-in-law	−0.002**	−0.002*	−0.000

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	Unweighted (1)	IPUMS-LRS Weights (2)	IP Weights (3)
	(0.001)	(0.001)	(0.001)
Relationship to Head: Grandchild	−0.000	0.000	0.000
	(0.000)	(0.001)	(0.000)
Relation to Household Head: Other	−0.061***	0.006	−0.002
	(0.003)	(0.006)	(0.005)
In White Collar Occupation	0.013***	−0.000	0.000
	(0.004)	(0.005)	(0.005)
In Farming Occupation	0.063***	−0.001	0.002
	(0.005)	(0.007)	(0.007)
In Semi-Skilled Occupation	−0.031***	0.001	−0.000
	(0.004)	(0.006)	(0.006)
In Unskilled Occupation	−0.053***	−0.003	−0.002
	(0.004)	(0.006)	(0.005)
In Other or N/A Occupation	0.009***	0.004	0.000
	(0.003)	(0.003)	(0.003)
Lives with Mother	−0.002	0.022***	−0.001
	(0.004)	(0.005)	(0.005)
Lives with Father	0.005	0.027***	0.000
	(0.004)	(0.005)	(0.004)
Lives with Both Parents	0.006	0.025***	−0.000
	(0.004)	(0.005)	(0.004)
Father: Born Abroad	−0.063***	−0.040***	−0.001
	(0.003)	(0.005)	(0.006)
Mother: Born Abroad	−0.061***	−0.041***	−0.001
	(0.003)	(0.005)	(0.006)
Lives in Northeast	0.042***	0.023**	−0.004
	(0.005)	(0.011)	(0.011)
Lives in Midwest	−0.030***	−0.004	−0.000
	(0.005)	(0.011)	(0.011)
Lives in West	−0.014***	−0.005**	0.003
	(0.002)	(0.005)	(0.005)
Lives in South	0.002	−0.014	0.002
	(0.005)	(0.010)	(0.010)
Currently Married	0.058***	−0.008	0.002
	(0.005)	(0.006)	(0.006)
Farm Status	0.054***	0.010	0.003
	(0.005)	(0.008)	(0.008)
Number of Siblings in Household	−0.031**	0.052***	−0.001

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	Unweighted (1)	IPUMS-LRS Weights (2)	IP Weights (3)
	(0.015)	(0.020)	(0.018)
Living in Same State as Birth	0.044***	0.050***	−0.003
	(0.005)	(0.008)	(0.009)

Notes: A selected set of mean comparisons shows the difference between the means of the linked IPUMS-LRS and the linkable population without IPUMS-LRS weights in column (1); standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). Columns (2) and (3) present the same statistics using IPUMS-LRS and IP weights. The Appendix presents the full set of mean comparisons for 1860–1880 and all other IPUMS-LRS years.