NBER WORKING PAPER SERIES

FINDING JOHN SMITH:
USING EXTRA INFORMATION FOR HISTORICAL RECORD LINKAGE

Ran Abramitzky
Leah Platt Boustan
Harriet M. Brookes Gray
Katherine Eriksson
Santiago Pérez
Hannah M. Postel
Myera Rashid
Noah Simon

Working Paper 33999
http://www.nber.org/papers/w33999

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2025

We thank Frankie Fan for excellent research assistance, Joe Price for useful conversations, and Kasey Buckles, Alvaro Calderon, Bill Collins, Jacob Conway, Gabriele Cristelli, Will Dobbie, Joe Price, and Ariell Zimran for detailed comments. We appreciate suggestions and conversations at the NBER Development of the American Economy Summer Institute. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Finding John Smith: Using Extra Information for Historical Record Linkage
Ran Abramitzky, Leah Platt Boustan, Harriet M. Brookes Gray, Katherine Eriksson, Santiago Pérez, Hannah M. Postel, Myera Rashid, and Noah Simon
NBER Working Paper No. 33999
July 2025
JEL No. N31, N32

ABSTRACT

We introduce a new rule-based linking method for historical Census records. We augment earlier algorithms based on name, age and place of birth (Abramitzky, Boustan, Eriksson, 2012, or "basic ABE"), with five matching characteristics – middle initial, county of residence, and spouse and parents' names. Relative to basic ABE, ABE-Extra Information ("ABE-EI") greatly increases match rates, improves accuracy and is similarly representative of the population on most attributes, with geographic mobility being one important exception. Relative to machine learning algorithms, ABE-EI has somewhat lower match rates, improved representativeness, and offers full replicability. We also create the first ABE-based links for women.

Ran Abramitzky
Stanford University
Department of Economics
and NBER
ranabr@stanford.edu

Leah Platt Boustan
Yale University
Department of Economics
and NBER
leah.boustan@yale.edu

Harriet M. Brookes Gray
Yale University
harriet.brookesgray@yale.edu

Katherine Eriksson
University of California, Davis
Department of Economics
and NBER
kaeriksson@ucdavis.edu

Santiago Pérez
University of California, Davis
Department of Economics
and NBER
seperez@ucdavis.edu

Hannah M. Postel
Duke University
Sanford School of Public Policy
hannah.postel@duke.edu

Myera Rashid
Northwestern University
Department of Economics
mrashid@u.northwestern.edu

Noah Simon
Princeton University
ns4240@princeton.edu

1. Introduction

The recent digitization of historical complete-count population censuses has enabled social scientists to create the first large individual-level panel datasets for the past.[1] These historical datasets, which link millions of individual records, allow scholars to study central questions about intergenerational mobility, the persistent effects of childhood environment, and the assimilation of immigrant populations, among other topics.[2]

The main challenge for linking historical records is that most historical datasets lack unique identifiers such as Social Security numbers. Researchers have instead developed linking algorithms based on a small set of pre-determined matching attributes such as first and last name, birthplace, and birth year (see, for instance, Ferrie 1996, Long and Ferrie 2013, and Abramitzky, Boustan, and Eriksson 2012, 2014, 2019 which we refer to throughout the paper as "basic ABE"). However, because a substantial fraction of the population cannot be uniquely identified by these attributes alone, these algorithms can only link around a quarter of the male population.[3] Matching a relatively small share of the population (16-30%) raises concerns about representativeness of linked samples and limits the ability to study questions related to smaller subpopulations.

In this paper, we introduce and assess the performance of a new rule-based, fully automated linking algorithm that adds five matching characteristics – middle initial, county of residence, spouse's name, mother's name, and father's name – to the basic ABE matching approach. We refer to this algorithm as "ABE with Extra Information" (or "ABE-EI"). We post these new links for all Census pairs on the Census Linking Project website and provide a Stata command to implement this approach, allowing researchers the flexibility to customize the method to their research applications. We also create and test the first links for women using the ABE and ABE-EI

[1] Linked historical census datasets are posted for researcher use by the Census Linking Project (Abramitzky et al. 2020), the Multigenerational Longitudinal Panel project (Helgertz et al. 2022), and the Census Tree (Buckles et al. 2023).

[2] Abramitzky et al.'s (2021) survey of historical linking methods cites more than 50 such research articles published through 2020 and has been cited by nearly 400 additional papers (as of June 2025).

[3] Studies using linked data have traditionally focused on men, because women tend to change their last name upon marriage. Standard algorithms can link women who remain married or unmarried across two census pairs. More recent studies have started making progress in linking women across the marriage transition by using other data sources (for instance, marriage certificates) that include women's maiden names (see, for instance Eriksson, Niemesh, Rashid, and Craig 2023).

algorithms. As with any algorithm, we focus on women who do not change their marital status across census periods and therefore retain their surname. Note that women who change their marital status can only be matched using genealogical data or with external sources such as marriage certificates.

Relative to the basic ABE algorithm, ABE-EI produces samples with higher match rates (25-37%) and improved accuracy for both men and women, as measured by accordance with hand-made links from amateur genealogists on the online Family Tree (97-99% accordance). For comparison, ABE basic "only" agrees with the Family Tree 90-97% of the time.[4] For most attributes, these improvements do not come at a cost of population representativeness relative to the basic ABE algorithm, although we caution that ABE-EI matches fewer individuals who move between locations *and* change their household structure (e.g., men who move out of their childhood homes and move away from their childhood county).

We summarize algorithmic performance with a "matching frontier" illustrating the tradeoff between increasing match rates (True Positive Rate or TPR) and improving accuracy (Positive Predictive Value or PPV). By this metric, adding extra information *shifts out* the matching frontier by both increasing sample size and improving accuracy at the same time. We emphasize that improvements in match rates (and thus sample size) are largest for ABE conservative variants that require that a match be unique within a five-year age band. Improvements in accuracy are largest for the ABE "standard" variants, an iterative algorithm that prioritizes exact age matches before considering matches that are one (or two) years older/younger.

Although no linked sample is fully representative of the population, adding extra information to ABE does not reduce sample representativeness of the population relative to basic ABE algorithms on most attributes. ABE-EI is consistently more representative than basic ABE on one attribute (married) and consistently less representative on two attributes (living in group quarters, living in state of birth).

In addition to benchmarking ABE-EI against the basic ABE algorithm, we also compare ABE-EI to two widely-used machine learning algorithms for linking historical census records: the IPUMS

[4] We note that accuracy can only be measured in this setting for links both made by ABE and present on the Family Tree (around 70% of ABE links).

Multigenerational Longitudinal Panel ("MLP") and the machine learning links that are included in the data released by CensusTree.org ("CT-ML") (Helgertz, et al. 2022, Price, et al. 2021). Note that the CT-ML links are only a subset of the links provided by the Census Tree; the full Census Tree dataset also includes genealogical links from the Family Tree and links from other sources. The MLP and CT-ML algorithms use hand-linked training data and train probit models to determine the statistical relationship between individual attributes and the probability of linkage. This approach implicitly uses *all* available information about an individual – including residential location and names of household members – to generate matches.[5]

Machine learning algorithms achieve higher match rates than ABE-EI (between 36-45%).[6] Linked samples produced with ABE-EI are similarly representative of the population to CT-ML and somewhat more representative than MLP. ABE-EI is substantially more representative of the population than the genealogical links from the Family Tree, the composition of which depends on who chooses to research their ancestors and add their details to the tree. Concerns about representativeness are accentuated for racial and ethnic subpopulations.

The choice between ABE-EI and machine learning algorithms can be a matter of researcher preference. Although machine learning algorithms provide higher match rates, procedural algorithms like ABE-EI have two important benefits. First, ABE-EI is fully replicable, in the sense that the linking process can be reproduced from start to finish without any human decision making. By contrast, machine learning methods rely on training data. Creating training samples requires using human coders to determine likely links and this process is fundamentally impossible to replicate: Two different human coders often make different decisions, and even the same human coder might not be fully internally consistent and might code differently in another time. Reliance on training data is particularly problematic in the context of historical record linkage because we do not have ground truth data to adjudicate between different potential matches whenever human coders disagree. Furthermore, the final links selected by machine learning models can change as

[5] Some recent work relies on hand-linked samples directly for analysis of specific sub-populations. Recent examples include studies of the recipients of mothers' or Union Army pensions, participants in a mobility program, winners of a land lottery, roommates at Harvard College, or samples of US Congressmen (Aizer, et al. 2020; Abramitzky, et al. 2024; Bleakley and Ferrie 2016; Eli 2015; Michelman, et al. 2022; and Querubin and Snyder 2013). However, creating specific hand-linked samples is prohibitively expensive for many purposes.

[6] The full Census Tree dataset, which combines CT-ML with the hand links from the Family Tree, achieves much higher match rates than any algorithm can achieve on its own (64-69%).

new training data become available, leaving the underlying sample fundamentally unstable. Second, ABE-EI can be easily customized by researchers who want to allow matches made by one attribute but not another, or who want to link other historical datasets to the census. Re-running an ABE-EI algorithm takes only a few hours on a standard laptop, whereas re-running a customized machine learning model requires substantial time on a supercomputer.

Finally, we investigate how the use of various linking algorithms affects inference in a common research context – namely, estimates of father-son intergenerational mobility. For all linked samples, whether from ABE or machine learning algorithms, and whether using a basic or expanded set of matching attributes, we estimate a narrow range of father-son rank-rank correlations in the full population. Specifically, we estimate rank-rank correlations of 0.38-0.42 in unweighted regressions and 0.35-0.38 in regressions that are reweighted to match observed population characteristics. The lower persistence rates (higher mobility) observed in the weighted samples points to the importance of reweighting linked samples for inference.

We end the paper by providing suggestions for researchers using historical linked census data. In many applications, ABE basic can be replaced with ABE-EI with little downside because ABE-EI improves on both match rates and link accuracy with little effect on representativeness. However, we emphasize that researchers should select between algorithms depending on their question of interest. For instance, ABE-EI may be particularly well-suited for questions involving small subpopulations like residents in rural counties, specific urban neighborhoods, or members of ethnic minorities. However, this algorithm may be less appropriate for questions about geographic mobility because it uses county of residence as one matching attribute.

Given that linking is inevitably imperfect, we recommend that researchers test the robustness of their results across linking algorithms. We also encourage end users of historical linked datasets to familiarize themselves with the details of how the data was constructed and to make informed decisions about which sample(s) are appropriate to their research question.

This paper contributes to the growing methodological literature examining best practices in historical record linking. These studies include the introduction of rule-based algorithms (Abramitzky, Boustan and Eriksson 2012, 2014, 2019), a statistically optimized approach (Abramitzky, Mill and Perez 2019, Protzer et al. 2024), and the use of machine learning techniques with hand-linked training data (Feigenbaum 2016, Helgertz et al. 2022 and Price et al. 2021).[7] New work by Althoff et al. 2024, Bailey and Lin 2024, Buckles et al. 2023 and Eriksson et al. 2023 are applying matching techniques to social security data, marriage licenses, and genealogical links to match the many women who changed their name at marriage. Bailey et al. 2020 and Abramitzky et al. 2021 offer comparisons across linking approaches. We add to this literature by proposing a linking approach that allows for the incorporation of extra information for matching, but does not require a hand-linked training sample and hence is still fully automated, easily replicable, and can be readily used to match the census to other historical sources. We also add the first links for women with the ABE basic and new ABE-EI algorithms.

More broadly, the longitudinal datasets arising from these matching methods have provided new evidence on social mobility in the past, including questions of intergenerational mobility, immigrant assimilation, the long-run effects of government programs, and the effects of childhood environment on economic outcomes. New work includes papers on the long-run effect of the 1918 flu published after the COVID pandemic (Beach, et al. 2022a, 2022b, Ager et al. 2021), papers on the long-run consequences of wartime shocks (Ager et al. 2021, Feigenbaum et al. 2022, Caprettini and Voth 2023) and studies of the long-run effects of governmental or non-profit programs (Abramitzky, Boustan and Connor, 2024, Modrek, et al. 2022), among others. By increasing match rates, the approach we describe in this paper would enable researchers to investigate topics requiring large sample sizes, particularly questions relating to smaller or underrepresented groups.[8]

2. The Abramitzky, Boustan and Eriksson (ABE) matching algorithms

This section describes our new extra information procedural algorithm in detail. Before doing so, we review the basic algorithm introduced in Abramitzky, Boustan, and Eriksson (2012, 2014, 2019) and outlined in detail in Abramitzky, et al. (2021).

[7] Perez 2017, Zimran 2020, and Bailey et al. 2020 emphasize the importance of population weighting when conducting analysis with linked samples. Mattheis (2024) and Sasaki and Zimran (2025) introduce methods to correct estimates in the presence of false positives. Ghosh, et al. 2023 raise concerns about the legibility of the handwriting of census enumerators, leading to transcription errors.

[8] Postel (2023) and Postel et al. (2025) suggest refinements of these matching techniques that improve their success rates for sub-groups with distinctive names (e.g., immigrants born in China).

2.1. ABE basic

The core steps of the algorithm when matching individuals from dataset A to dataset B (for instance, from the 1900 Census to the 1910 Census) are diagrammed in Figure 1a and can be described as follows:

1. Clean and standardize names in both datasets for common nicknames.[9] Restrict the sample to people who are unique by first and last name, implied birth year calculated from calendar year and age, and place of birth (either state or country) in dataset A.
2. For each record in dataset A, look for records in dataset B that match on first name, last name, place of birth, and birth year. At this point there are three possibilities:
    a. If there is a *unique* match, this pair of observations is considered a match.
    b. If there are multiple potential matches in dataset B with the same year of birth, the observation is discarded.
    c. If there are no matches by exact year of birth, the algorithm searches for matches within ± 1 year of reported birth year, and if this is unsuccessful, it looks for matches within ± 2 years. In each of these steps, only unique matches are accepted. If none of these attempts produces a unique match, the observation is discarded.
3. Repeats Steps 1 and 2 but now matching from dataset B to dataset A. Our sample is comprised of the intersection between observations that uniquely link from dataset A to dataset B and observations that uniquely link from dataset B to dataset A.[10]

These steps describe the basic ABE algorithm (“standard”); this terminology refers to the original iterative approach for matching on age. However, the algorithm also has two commonly used (and often preferred) variants:

---

[9] To clean and standardize names for ABE basic, we remove any non-alphabetic characters and account for common misspellings and nicknames. This allows the algorithm to recognize Ben and Benjamin as the same name.

[10] An earlier version of the algorithm performed the matching in only one direction (from A to B). However, the more recent implementations, including the *abematch* Stata command and the links available from the Census Linking Project, perform the matches in both directions as described above.

ABE conservative: This variant modifies the uniqueness requirement described in Step 1 above by further imposing that names are unique within ± 2 years of the implied birth year. By requiring uniqueness within a wider age band, this algorithm reduces the number of false positive matches between individuals with the same name but slightly different ages. We often recommend that researchers use ABE conservative variants because they exhibit higher accuracy.

ABE NYSIIS (versus exact): For this variant, names are first standardized using the New York State Identification and Intelligence System (NYSIIS) algorithm, rather than using the exact names as reported in the census. This approach helps deal with differences in spelling and pronunciation across different sources.

2.2. ABE with extra information (ABE-EI)

2.2.1. Conceptual issues and variable selection

All versions of the ABE algorithm described above use first and last name, place of birth, and year of birth as the only matching features. These characteristics share three valuable properties: (1) they are typically "immutable" throughout an individual's lifetime, (2) they are "always observed" regardless of an individual's economic decisions (e.g., whether to live with parents or to get married), and (3) they are available in every census survey and for all respondents.[11]

A drawback of using only these basic matching attributes is that a sizeable proportion of individuals in the data are non-unique on the basis of these characteristics alone. Linked samples based on a small set of matching attributes will have lower match rates and may be unrepresentative of the population because they are less able to match individuals with common names.

We propose that adding pieces of extra information to "break ties" between individuals with common names will increase match rates and improve accuracy because the likelihood that two unrelated people share all attributes falls as the number of attributes rises.

---

[11] Even the basic matching attributes may change over time due to name changes, mis-reported ages or changes in the boundaries of country of birth. The potential instability of basic characteristics is an inherent limitation of linking census records, given the lack of unique numerical identifiers. Race is another characteristic that likely satisfies these properties. However, due to the phenomenon of "racial passing", self-reported race might also change from census to census (see, for instance, Mill and Stein 2016 and Dahis, Nix and Qian 2019).

Table 1 illustrates the potential for disambiguation when using one such additional characteristic: mother's first name. This example shows two 7-year-old individuals named John Smith who were born in Alabama. Although we would not be able to uniquely match these individuals using just the baseline matching features, we may be able to do so once we observe extra information such as mother's name in this case (Mary vs. Ellen).

Appendix Table 1 lists a set of potential matching variables from the census. We take this set from the matching variables used in the MLP algorithm. The first six variables (variants of name, age and birthplace) are used in ABE basic, as marked in column 2. The remaining 16 variables are candidates for use in an extra information algorithm. As we discuss in detail below, we select five of these variables – middle initial, mother's and father's name, spouse's name and county of residence, as marked in column 3 – using the following criteria: algorithmic performance, coding simplicity, and broad availability in the census. Notably, the variables used in ABE basic, which were intended to be immutable throughout an individual's lifetime, are only somewhat more likely to be persistent across time in family hand-links (73-96%) than are the variables selected for ABE-EI (76-86%); see column 4.

The logic behind our variable selection is as follows:

We opt not to use the names of other household members beyond spouse and parents (siblings, children, etc.) in ABE-EI as doing so substantially increases computational time with very little gains in terms of additional matches. When we experimented with including these attributes as matching variables, we found that they broke very few ties after already matching on parents' names and spouse's name: Individuals living at home with siblings (children) are also almost always living at home with parents (spouses). However, using siblings (or children) names has a substantial toll on computational time as it requires comparing *all* siblings (children) in one dataset to *all* siblings (children) in the other dataset.[12]

We also choose not to use variables like age of first marriage, duration of current marriage or mother tongue because they are only available in a few census years and so they cannot be consistently applied to create matches (see Appendix Table 1, column 1 for details).

12 For instance, if someone has two siblings of the same gender in one dataset and two in the other, using siblings' names requires performing four comparisons. By contrast, using the name of the father and the name of the mother requires just two comparisons.

Finally, we exclude four variables – race, year of immigration, parental nativity and state of residence – from the algorithm based on poor performance. Table 2 compares the accuracy rates, as measured by agreement with hand-made links on the Family Tree, for ABE basic (column 1) and our preferred ABE-EI (column 2) relative to two possible versions of ABE-EI: an ABE-EI that uses only "immutable" attributes (middle initial, race, year of immigration and parental nativity) (column 3), and an ABE-EI that includes all nine matching variables (the five in the actual algorithm, as well as race, immigration year, parental nativity and state of residence) (column 4). Our preferred ABE-EI algorithm agrees with the hand-made Family Tree links 97-98% of the time. Using only immutable attributes or adding three of these immutable attributes and state of residence lowers accuracy (accordance rate = 90-96%).

Appendix Table 2 documents that each of these four excluded variables – as well as matches selected only by the basic ABE linking variables – contribute matches with poor accuracy. The first panel reports matches made by the basic linking variables or one of the excluded variables *as well as* at least one of the variables in the preferred ABE-EI algorithm (these matches may be overlapping). The second panel contains matches that are made only by the basic linking variables or one of the four excluded variables that are not selected by one of the preferred algorithm variables. Matches in the first panel are highly accurate (95-99%), but matches in the second panel have much lower accuracy: 61-70% accuracy for conservative variants and 43-48% accuracy for standard variants. We conclude that these variables have poor performance when at least one piece of extra information does not contribute to selecting the match, so we exclude them from the algorithm (Note that the overall accuracy rates for ABE basic reported in Table 4 are a weighted average of the high accuracy of the 5-7 million links also selected by a piece of extra information and the low accuracy of the 1-2 million links that are not).

Appendix Table 3 provides specific examples of matches that would be made by one of these excluded variables but are not selected by one of the core five variables in the preferred ABE-EI. These candidate matches are not included in our ABE-EI samples. The first case, Rupert Fritz, would be linked by ABE basic because he is the only man with this name, birth year and birthplace in both 1900 and 1910. However, other pieces of extra information do not select this match (e.g., Rupert's mother's name is Mary in 1900 and Catherine in 1910). The second case, John Scholz, would be matched using basic linking variables and parental nativity, but again is not selected by other pieces of extra information (e.g., John's mother's name is Margaret in 1900 and Susanna in

1910).

Table 3 documents two attributes of our preferred linking variables: share of records that are unique when adding extra information and share of records for which this variable is present. 78% of individual records are unique by basic matching characteristics (first and last name, age and place of birth). Adding middle initial or parents' or spouse's name as matching attributes increases the proportion unique to 89-91%, whereas adding county of residence increases this proportion to 97%. The last row in Table 2 reports the share of observations with a non-missing field for each extra information attribute. Extra information is only useful in creating a match when it is reported in both census years. County of residence is particularly useful because it is always available. The other pieces of extra information are available between 36% and 45% of records. Note that, conditional on the attribute being non-missing, the uniqueness share rises further.[13]

2.2.2. ABE with "extra information" algorithm (ABE-EI)

We now describe the algorithm that incorporates extra information into ABE-style census linking. The algorithm is diagrammed in Figure 1b and has the following steps:[14]

1. Create candidate matches using each piece of extra information separately. In particular, conduct five concurrent rounds of ABE-style matching using the baseline set of matching characteristics plus one additional matching feature. For instance, when using county of residence as the additional matching feature, we require uniqueness on the basis of first name, last name, place of birth, birth year, *and* county of residence and then look for a possible match on these attributes, ignoring other pieces of extra information. If the piece of extra information considered in a given round is missing in the census (e.g., mother's name is blank), we do not make a match in that round. At this point, each individual from dataset A could generate from zero to five possible matches to dataset B. Note that possible matches made in this step may be different from each other or may be overlapping.[15]

[13] For example, *conditional* on observing mother's name in 1900, around 93% (rather than 91%) of individuals are unique on the basis of this variable.

[14] Middle initial has been incorporated into ABE variants in the past, but is not included in the most commonly used ABE algorithms (see Abramitzky, et al. 2021).

[15] Note that we do *not* add a round for the basic ABE matching variables only because we deemed the matches made by basic matching variables but not selected by any piece of extra information to have low rates of accuracy (see Appendix Table 2).

2. Choose valid matches from this set of candidate matches based on the following uniqueness criterion: The individual from dataset A (call them "A1") must be linked to only one unique individual from dataset B ("B1"). For example, if the father's name "matching round" yields a link from A1 to B1, but the spouse's name "matching round" links A1 to B2, then we drop all matches made from A1 to individuals in dataset B.
3. Repeat Steps 1 and 2, but now matching from dataset B to dataset A. As in ABE basic, the sample is comprised of the intersection between observations that uniquely link from dataset A to dataset B and observations that uniquely link from dataset B to dataset A.

ABE-EI makes one further change beyond the published versions of ABE basic. We replace 'country of birth' with 'region of birth' for individuals born in Eastern Europe for the 1900-10 and 1910-20 census pairs to address changes in country boundaries in Europe in this period (for example, Poland is not offered as a 'country of birth' option in the 1910 Census).[16] We recommend making this change for ABE basic as well.

Appendix Table 4 subdivides the 1900-10 census links made with ABE-EI into groups based on the set of linking variables that selected the match. To reduce dimensionality, we collapse mother's and father's names into a single category (parents' names) and report matches made by the seven possible combinations of: (1) parent or spouse's name, (2) county of residence and (3) middle initial. 54% of links are made by either two or three of these options. In this case, accuracy rates, as measured by agreement with hand-made Family Tree links, are 99.3-99.9%. 45% of links are made by only one piece of extra information. Links made by county of residence or parental or spouse's names alone are 89-95% accurate. The only links that exhibit relatively low accuracy (75-81%) are links made by middle initial alone, which represent only 5% of the sample.

In Table 4, we provide two examples from the 1900 census of individual records that are excluded from the linked sample because they fail the uniqueness screen in step 2 of the algorithm. The first example, Edward Hoffman, was three years old in 1900, living with his mother Wilhelmina and his father Charles. According to the mother's name round, Edward matches to another Edward Hoffman (age 13 in 1910) with a mother named Wilhelmina and a father named Herman.

[16] The region of birth correction for 1910 applies to the following country and region of birth designations; Austria, Bulgaria, Czechoslovakia, Germany, Hungary, Poland, Romania, Yugoslavia, Central Europe, Eastern Europe, Estonia, Latvia, Lithuania, Baltic States, Other USSR/Russia.

According to the father's name round, Edward matches to a different Edward Hoffman (age 13 in 1910) with a mother named Mary and a father named Charles. We thus drop Edward from the sample. The second example, Daniel Lehr, aged 32 in 1900, matches to one Daniel living in the same county in 1910 by the county of residence round but a different Daniel with the same father's name in 1910 by the father's name round. Both observations and others like them are excluded.

3. Measuring performance

We assess the performance of ABE-EI for linking historical US census records along three dimensions: match rates (or the share of potential links that are created), accuracy (or the share of created links that match to the "right" person), and representativeness of the linked sample relative to the full population. Our measures of accuracy are only suggestive because we lack a true measure of "ground truth" in this setting. We consider the effect of adding extra information to a variety of ABE algorithms, from the most conservative (ABE exact conservative) to the most lenient (ABE NYSIIS standard). We first focus on men for comparability with our previous work and then introduce a new sample of women linked across censuses.

3.1 Match rates

Figure 2 shows that match rates increase when using ABE-EI to link between the 1900 and 1910 census compared to basic ABE for all variations of the algorithm. This pattern holds for all census pairs from 1900-10 to 1930-40, the match rates for which are reported in Appendix Table 5.

We present match rates for four variants of the ABE algorithm: standard versus conservative and NYSIIS versus exact. For example, when using the NYSIIS name correction, ABE-EI increases the match rates from 16% to 31% from 1900-10 for ABE conservative and from 26% to 33% for ABE standard. More broadly, ABE-EI increases match rates substantially for ABE conservative variants (9-15 percentage points) and mildly for the "standard" ABE approach (2-6 points). ABE-EI can increase match rates by enabling matches for individuals with common names who would otherwise be non-unique, but can also decrease match rates by eliminating any false matches that are created with basic attributes but are dropped when adding extra information. The second force is stronger for the ABE standard variants, and therefore the improvement in match rates is not as large in this case.

Figure 2 also compares ABE-EI's match rates to those of datasets from MLP and the Census Tree. MLP 'Round 1' is based on individual attributes alone. MLP-Full includes matches from both 'Round 1' and 'Round 2,' which links additional household members of individuals matched in the first round. For Census Tree links, ABE-EI is most comparable to links made by the machine learning algorithm (*xgboost*), which we call CT-ML. For completeness, we also report match rates for CT's Family Tree (the hand-made links, which are also used as training data) and for the full CT dataset, which includes the Family Tree, machine learning links, and various other components.

The highest ABE-EI match rate (33%) is lower than machine learning algorithms that incorporate extra information by a few percentage points, including MLP Round 1 (38%) and CT machine learning (36%). The match rate for the full MLP sample that includes Round 2 matches is 44% and the full CT sample that includes hand links is 64%.[17]

For comparison, the match rate for samples that are linked *by hand* using only available census variables range from 50% to 69%, with the highest value in the range representing a combination of hand links and machine learning (Aizer, et al. 2024; Catron, et al. 2024; Michelman, et al. 2022; Helgertz, et al. 2022).[18] Hand-linked samples provide a sense of the "ceiling" of possible links using only census variables, and these rates are well below 100 percent due to mortality, return migration, under-enumeration and transcription error. Taking the midpoint of this range as a metric of possible matches (60%), ABE-EI can match 55% of *possible* matches and machine learning algorithms can match up to 63%.

3.2 Agreement with hand-made links ("accuracy") and machine learning

[17] In Appendix Figure 1, we compare the match rates of ABE-EI links over 30 years (1900-30) relative to ten years (1900-10, as in Figure 2). The gain in match rates associated with extra information attenuates over longer periods when household structure and geographic location is less likely to be stable. For example, match rates for ABE-EI conservative NYSIIS falls from 31% to 16%. The same pattern holds for links made with machine learning algorithms, which rely on a similar set of extra information. However, the accuracy of these longer-range links are still high (Appendix Table 6).

[18] We calculate the implied match rates for Aizer et al. 2024 from summary statistics and observation counts reported in Appendix Table 2 in the Aizer paper. For example, of the 13,383 mothers in the original sample, 7,515 can be followed to the 1910 Census.

Although extra information increases match rates, higher match rates might come at the cost of accuracy, i.e., a higher rate of false positive links. For example, matching based on characteristics that can change over time, like state or county of residence, might result in a false match if the original resident moves out of state or across county lines. On the other hand, the use of extra information may *reduce* the rate of false positives. It is possible that Person A shares the same basic name and age as Person B, leading a false match to be made. However, such possibility becomes less likely as we increase the number of matching attributes. Ultimately, whether adding extra information improves accuracy depends on whether the "marginal" links that are made by doing so are more or less accurate than those made using only the basic information, which is an empirical question.

To assess the "accuracy" of matches made by ABE-EI, we compare links made by ABE-EI to high-quality genealogical links created through the Family Tree. Many of these manual links are made by members of the public searching for their own ancestors. Others are made by trained student linkers in the Family History lab at Brigham Young University. Often, these individuals have access to private family knowledge and so they can adjudicate between similar records and find the correct match. However, we note that even these links cannot be considered ground truth because we cannot determine with certainty whether the pairing is indeed true or false. Without ground truth, we report comparisons with the Family Tree links as an imperfect proxy for accuracy.

The first column of Table 5 compares samples generated by various ABE algorithms for 1900-10 census links to the genealogical matches created through the Family Tree. In all cases, we report the 'conditional agreement rate,' or the share of matches created by *both approaches* that link to the same record. We report agreement rates for other census pairs in Appendix Table 7. Note that we are only able to assess accuracy for the subset of matches that are included on the Family Tree. Links made only by ABE but not by the user of Family Tree represent individuals who did not have any descendants or whose descendants chose not to use the Family Tree.

For all variants, we find that adding extra information improves match accuracy. For ABE basic, the conditional agreement rate ranges from 90-94% for standard variants to 96-97% for conservative variants (Abramitzky, et al. 2021 also reported this pattern). By adding extra information, the conditional agreement rate increases to 97-99% for standard variants and 98-99% for conservative variants. The largest improvements in agreement rates are for ABE standard variants, which do not impose uniqueness within a 2-year age band and hence introduce relatively more false positive matches between individuals of similar ages.

We cannot use the Family Tree genealogical links to compare the accuracy of ABE-EI to that of machine learning algorithms because the machine learning algorithms use the Family Tree as a component of their training data and hence will (by construction) always agree with records on the Family Tree. Instead, the second and third columns of Table 5 compare ABE-EI to machine learning algorithms directly. We note that any disagreement between ABE and machine learning-based matches does not mean that the ABE-based link is incorrect, but instead merely reports that the two linking approaches disagree.

We find that adding extra information improves the already-high conditional agreement rate between ABE and machine learning algorithms, particularly for standard variants. For ABE basic, standard variants enjoy 91-94% conditional agreement with CT-ML and MLP and conservative variants agree 97-99% of the time. After adding extra information, these conditional agreement rates jump up to 98% for standard variants and 98-99% for conservative variants. Essentially any match that is made both by ABE-EI and a machine learning algorithm links to the same individual.

3.3 The Match Rate-Accuracy Frontier

Adding extra information to the ABE algorithm improves both match rates and accuracy, meaning there is *no tradeoff* between the gain in match rates and a loss in accuracy. As a result, ABE-EI represents a shift out in the frontier of possible matches, rather than a move along the frontier expressing the tradeoff between match rates and accuracy.

Figure 3 visualizes this algorithmic performance by comparing the True Positive Rate (TPR) – a measure of match rates – and the Positive Predictive Value (PPV) – a measure of accuracy – for variants of ABE basic and ABE-EI algorithms produced for different census pairs (1900-10 through 1930-40). For this figure, we consider hand-made links on the Family Tree to be "ground truth." TPR indicates the share of matches on the Family Tree that are also made by ABE (one measure of match rates), and PPV indicates the share of matches made by both ABE and Family Tree that agree (one measure of accuracy).

In earlier work, we showed that ABE basic algorithms exhibited a tradeoff between higher match rates and improved accuracy (Abramitzky, et al. 2021). Figure 3 documents two notable improvements in performance with ABE-EI. First, all ABE-EI variants have higher match rates and improved accuracy relative to all ABE basic samples, as depicted by a shift to the upper right of the quadrant. All basic ABE variants (represented with open shapes) have lower TPR (28-42%) and lower PPV (90-97%), whereas all ABE-EI variants (represented with colored shapes) have higher TPR (45-57%) and higher PPV (97-99%). Second, the tradeoff between match rates and accuracy (downward slope) which is apparent in ABE basic is nearly eliminated with ABE-EI. Standard variants of ABE basic offer higher match rates but at a cost of 4-5 points lower accuracy relative to conservative variants of ABE basic. By contrast, the ABE-EI variants with the highest match rates can be achieved with only a 0.5-1.0 point cost in accuracy. By disambiguating individuals with common names, using extra information can both add new links to the sample and, at the same time, improve certainty in the match.

3.4 Are linked samples representative of the population?

Thus far, we have documented that adding extra information to the ABE algorithm improves match rates and accuracy. However, adding extra information can have conflicting effects on the representativeness of the sample. On the one hand, a selected sample of individuals will remain in the same county or remain married to the same spouse, leading to bias. On the other hand, adding extra information allows the algorithm to "break ties" for individuals with common names like John Smith, who otherwise might be different from the population as a whole.

Figure 4 compares the representativeness of matched samples created with various linking algorithms for 1900-10 relative to the full male population in the base year (1900 census), the population at risk to be matched (Appendix Figure 2a-2c present similar graphs for all census pairs relative to the base year, and Appendix Figure 3 instead compares the 1900-10 matched sample to the 1910 census).[19] In particular, Figure 4 reports coefficients from a series of regressions that compare each linked sample to the full population. Each regression takes the form:

$$Y_i = \alpha + \beta\,(=1 \text{ if in linked sample}_i) + \varepsilon_i \qquad (1)$$

[19] Overall, representativeness patterns look similar whether comparing to the base year (Figure 4) or to the end year (Appendix Figure 3). One difference is that the linked sample is somewhat younger than the full population when comparing to the base year because older men in 1900 are more likely to die by 1910 and thus remain unmatched. By contrast, the linked sample is somewhat *older* than the full population in 1910. Because of this age gap, the linked sample is more likely to be married or live with children when compared to the full 1910 population.

For each linked sample, we consider a full set of dependent variables (Y_i = an indicator for being white, an indicator for being US-born, etc.), and the right-hand side variable is an indicator for being in the linked sample in question. If the linked sample is more likely to include individuals with a particular attribute, the coefficient is greater than zero; otherwise, it is less than zero. Coefficients are reported in standard deviation units. Appendix Table 8 contains regression coefficients corresponding to the figure but without standardization so that differences can be observed in 'natural' units.

We document three patterns about representativeness. First, no matching algorithm is fully representative of the population. Men in linked samples tend to be 0.1-0.2 standard deviations away from the full population on attributes such as race (white or Black), US-born, literate, living in group quarters, residing in one's state of birth and living with parents. Linked samples also "underrepresent" older individuals, which is a mechanical effect of higher mortality rates at higher ages. Yet algorithms produce samples that are largely representative of being married, living with children, urban or farm status. Note that, for most attributes, the crowd-sourced data on the Family Tree is less representative of the population than any matching algorithm. The lack of representativeness on the Family Tree likely reflects the set of users who choose to participate in online family history projects.

Second, adding extra information to the ABE algorithm leaves population representativeness unchanged for most attributes. ABE-EI is consistently more representative than ABE basic on one attribute (married status) and less representative on two attributes (living in group quarters and residing in state of birth). Third, ABE-EI achieves similar representativeness to CT-ML, but is more representative of the population than the MLP algorithm on most attributes.

3.5 Using ABE and ABE-EI to match women

We create and test the first links for women using the ABE and ABE-EI algorithms. We focus on women who do not change their marital status across census periods and therefore retain their surname. This sample restriction also holds for machine-learning based algorithms; women who change their marital status can only be matched using genealogical data. The match rates and accuracy of women matched by ABE are similar to those for men.

Table 6 reports the match rates and accuracy of links created for women between 1900-10 using various ABE algorithms. We focus on women who report being "never married" or "ever married" in both census years. These two groups represent most women at the youngest and oldest ages. However, a sizeable group of women in their teens and twenties transition from unmarried to married.[20]

For women, the ABE-EI algorithm with the highest match rates (NYSIIS-standard) can link 28% of women, compared to 30% of women with the CT machine learning algorithm, 34% of women with MLP, and 33% of men with ABE-EI. The 28% match rate is a weighted average of a somewhat lower match rate for never married women (27%) and a somewhat higher match rate for ever married women (29%). Higher match rates for ever married women is found for all ABE and machine learning algorithms.

Women's matches created by ABE-EI are highly accurate when compared to the Family Tree. Accuracy rates by algorithm vary between 96% and 97.5%, with slightly lower rates for never married than for ever married women (e.g., 95.5% for never married vs. 97% for ever married for NYSIIS-standard). Accuracy rates for men range from 97 to 98%.

3.6 Can ABE-EI match men who move between counties?

One concern about adding information like county or state of residence to linking algorithms is that men who move between geographic areas will be more difficult to match. Indeed, samples linked with ABE-EI contain a higher share of men living in their state of birth relative to the population and is somewhat more unrepresentative on this dimension than ABE basic.

Despite the use of county of residence as one matching variable, ABE-EI is still able to capture some movers using middle initial or parents' and spouse's name (for men who move with family members). Appendix Table 9 reports the share of links selected by each matching variable, documenting that different combinations of extra information are useful for linking movers of

[20] Appendix Figure 4c shows that 40-70% of women in core ages (ages 11-20 in the base year) cannot be linked because of a marriage transition. For example, for 11-year-old women in 1920, 100% are unmarried in 1920 (Appendix Figure 4b) and 45% of this group transition to marriage by age 21 (Appendix Figure 4a), suggesting that 45% of the age group go through a marital transition. For 20-year-old women in 1920, 60% of this age group are unmarried in 1920 and 75% of these unmarried 20-year-olds transition to married by age 30 (= 0.6 x 75% = 45%). The share of women who cannot be linked due to marriage transition quickly falls off at older ages because 80% of women are ever married by age 27 and, of those who remain unmarried, the probability of transitioning into marriage also falls.

different ages. Young children (age 0-9 in base census) who move between counties often move with their parents and are linked primarily on the basis of mother's and father's names. Older adults (age 20+) often move with their spouse and are linked on the basis of spouse's name. Middle initials also help to link teenagers (age 10-19) and older adults.

Appendix Table 10 provides an example of a possible mover who is selected as a match by ABE-EI and another possible mover who is excluded by ABE-EI. George Joynson moved between Minnesota (county 1050) and Oklahoma (county 710). Yet he has the same middle initial and mother's and father's names. In this case, three pieces of extra information select the same individual record as a match in 1910 and two other pieces of information (spouse's name and county) do not select a match at all. Thus, George Joynson is included in the ABE-EI sample. By contrast, Frank Hanley appears to have moved from Iowa to Washington State. Yet his mother's and father's names are different (Frances vs. Harriet and Henry vs. James) in 1900 and 1910. Thus, no pieces of extra information select a match for Frank Hanley and so he is not included in the ABE-EI sample.

The choice of ABE-EI versus ABE basic entails a tradeoff: ABE-EI captures "too few" movers but the movers included in the sample are more likely to be accurate. We illustrate this tradeoff in Appendix Table 11. First, the table provides suggestive evidence that ABE basic includes "too many" movers (i.e., some of the movers are false positive matches) whereas ABE-EI includes "too few" movers. Ground truth data on migration behavior from the 1940 census implies that 11% of young children, 13% of teenagers, and 11% of adults move across counties between 1935-40. Comparing the sample counts of movers and stayers in linked samples results in migration rates that are too high in ABE basic (12% of young children, 17% for teenagers and 17% for adults) and too low in ABE-EI (8-9% of all age groups).

Although ABE-EI matches "too few" movers, the movers that are included in the ABE-EI sample appear to be substantially more accurate than the movers in ABE basic. Accuracy problems in ABE basic become particularly severe for adult movers. In this case, only 79% of conservative matches and 55-57% of standard matches accord with the hand-made family links, compared to 85-91% of all movers in the ABE-EI samples.

Hence, ABE-EI matches are less likely to capture men who move between counties (particularly for teenagers who leave their childhood home) but the movers that are included in the sample are more likely to be accurate. This tradeoff raises the question of whether one can improve ABE-EI's performance in capturing movers by modifying the algorithm. We consider two modifications: dropping county of residence as a linking variable or replacing county of residence with state of residence as a linking variable.

We summarize the performance of these two geographic variants relative to ABE-EI with county of residence in Appendix Table 12. We note two patterns. First, county of residence is an important variable for differentiating between individuals with common names; therefore, dropping county as a matching variable substantially reduces match rates from 25% to 18%. Second, replacing county of residence with state of residence as a matching variable improves population representativeness on the share of the sample living in their state of birth, but does so at a cost to accuracy. The share of the sample that accords with hand-made Family Tree links falls from 98% to 95%. (Appendix Figure 5 compares population representativeness on all attributes observed in the census). Overall, we do not recommend modifying ABE-EI to try to capture additional movers.

4. Inference using ABE-EI

In this section, we consider the effect of using various extra information census linking algorithms on inference, focusing on intergenerational mobility, a research topic of great interest. In particular, we estimate historical rates of intergenerational mobility in samples constructed with ABE-EI and then compare point estimates to samples generated by basic ABE approaches or machine learning algorithms.

We argued above that adding extra information to ABE algorithms increased both the match rate and match accuracy with little cost to representativeness. An increase in accuracy (reduction in the "false positive" rate) could reduce attenuation bias, increasing coefficient estimates in absolute value. In the specific context of intergenerational mobility, note that reducing attenuation bias would result in *higher* estimated intergenerational persistence and thus *lower* rates of intergenerational mobility.

A central question in economics and economic history is: what is the rate of intergenerational mobility in income, wealth or other outcomes between parent and child? A common approach to answer this question is to observe young men living at home with their fathers in a childhood census (here: 1910 Census) and to link these men to an adulthood census (here: the 1940 Census). With an intergenerational dataset of this type, one can then estimate the correlation between a son's rank in the national income distribution and his father's rank. We estimate the following equation:

$$rank(Y^{son})_{1940} = \alpha + \beta rank(Y^{father})_{1910} + \varepsilon \quad (1)$$

where $rank(Y)$ denotes the individual's position in the national income distribution of their cohort and β estimates the persistence between the economic rank of a father and his son (β is inversely related to the level of social mobility). Because information on earnings is not available in historical censuses before 1940, we use a predicted income score measure based on occupation, age and state of residence (details on this measure can be found in Abramitzky, et al., 2021).

The MLP samples are only available for adjacent census pairs (1910-20, 1920-30 and 1930-40). For comparison, we create 1910-1940 linked samples for all algorithms by chaining together samples from three adjacent census pairs.[21] We note that, when adding extra information, it may be preferable to create this type of "chained" samples so that the sample is not constrained to only include men who have stable attributes (e.g., county of residence or spouse name) over a longer time window.

Table 7 compares estimates of the intergenerational rank-rank coefficient (β) across samples based on various ABE algorithms, with and without extra information, and on the machine learning algorithms from the Census Tree and MLP. Despite sample size changes across columns (from 620,000 to 4 million records), the coefficients remain stable. Starting with unweighted regressions (columns 1-2), we show that ABE basic samples produce estimated intergenerational elasticities between 0.38-0.40. Using ABE-EI instead results in estimate that are a point higher (0.39-0.41). The Census Tree samples produce similar point estimates (0.40-0.41) and the MLP estimate is a point higher yet (0.42).

Columns 3-4 report the same set of coefficients from regressions that have been weighted to represent the full Census population. Coefficients for all methods are around 3 points lower than their unweighted counterparts, suggesting that all linked samples select for families that have

[21] In particular, we match every two adjacent census pairs between the 1910 and 1940 Censuses (i.e. 1910-1920, 1920-1930 and 1930-1940) and only keep sons that can be followed in every census between 1910 and 1940. For ABE, a sample based on matching the end points (1910-40) produce nearly identical estimates.

higher rates of persistence (lower mobility), perhaps because of higher rates of farm families or households with literate heads. Coefficients for ABE basic are 0.35-0.37, for ABE-EI are 0.36-0.37 and for machine learning methods are 0.37-0.38. Coefficients remain 1 percentage point higher for ABE-EI than for ABE basic, likely reflecting the improved accuracy, and thus reductions in attenuation bias.

5. Linking racial and ethnic subgroups

Most of the American population in this period were white and US-born (around 75%). However, many research projects are focused on specific ethnic and racial sub-groups. This section compares the match rates, accuracy and representativeness of ABE and machine-learning based algorithms by subgroup.

Table 8 reports match rates and accuracy for all subgroups in 1900-10 for one preferred ABE basic and EI algorithm (NYSIIS-conservative) and for CT and MLP. We report results for other ABE algorithms and census years in the appendix.

For all algorithms, match rates are highest for white US-born residents, followed closely by immigrants from English-speaking countries (e.g., 36% for white US-born individuals using ABE-EI). Match rates are around half as large (18%) for Black US-born residents and for certain immigrant groups (e.g., Irish and Russian). Other immigrant groups have even lower match rates (Italians at 9% and Chinese at 3%).

Shifting from ABE basic to ABE-EI doubles match rates for most subgroups. For the Irish and Swedish, two subgroups with particularly common first or last names, ABE-EI triples match rates. For the Italians, match rates are not much affected by extra information. Match rates for almost every subgroup are 4-5 points higher with machine learning algorithms relative to ABE-EI.

Accuracy rates are high for all subgroups when using ABE-EI. All US-born groups and several immigrant groups (e.g., England, Germany, Sweden) agree with Family Tree links 97-98% of the time. These high agreement rates represent a 2-6 point gain in accuracy relative to ABE basic. Other immigrant groups (Ireland, Italy, Russia) have somewhat lower accuracy (95-96%), but these values represent a substantial gain relative to ABE basic, which had accuracy rates of around

88% for conservative variants and only 70-79% for standard variants. The Chinese are the only subgroup that do not have high agreement rates with the Family Tree even with ABE-EI (83%); however, there are very few Chinese-born individuals on the Family Tree and, from our hand checks of these links, their status as “ground truth” is uncertain.

The representativeness of linked samples differs by subgroup, particularly for characteristics like household composition that are incorporated into the extra information algorithm or the genealogical training data compiled by living ancestors. We focus on two attributes, marital status and living with children, in Figure 5 to document this point and report representativeness on all attributes in Appendix Figure 6. ABE basic, ABE-EI and MLP are highly representative of the population on living with children for white and Black US-born. By contrast, links made on the Family Tree are 0.3-0.4 standard deviations more likely to live with children, and this bias in the training sample has some effect on the CT machine-learning based sample. For immigrant subgroups, both ABE-EI and MLP become less representative than ABE basic (by around 0.3 standard deviations). The degree of unrepresentativeness in the Family Tree sample also gets more severe for immigrant subgroups, with individuals on the Family Tree being 0.75-1.0 standard deviations more likely to live with children. Furthermore, relative to ABE basic, any extra information algorithm is less representative for immigrant groups.

5.1 Case study of the Irish born

Ó Gráda et al. (2023) raise questions about the accuracy of the ABE basic algorithm when linking a sample of Irish-born men who participated in an ethnic savings bank from the 1860 to the 1870 Census. They argue that the rates of geographic and occupational mobility in the linked sample are “too high,” suggesting that men are matched to the wrong individual.

We consider this argument with our measure of accuracy, comparing ABE basic and EI links to the Family Tree. We conclude that the Ó Gráda et al. paper selects as its test case both a time period (1860-70) and an ethnic group (Irish immigrants) for which ABE basic has particularly low accuracy. Appendix Table 13 shows that when we consider other groups in 1860-70 or the Irish born in the other years (e.g., 1930-40), accuracy greatly improves. Agreement with the Family Tree is particularly low for the Irish born in 1860-70, ranging from 55-78%. Agreement is much higher for native-born whites in 1860-70 (87-95%) and for the Irish born by 1930-40 (82-95%).

Furthermore, using extra information improves the accuracy rate for the Irish substantially. When using ABE-EI, accuracy rates for the Irish born in 1860-70 increase by nearly 30 percentage points (to 82-89%) and are even higher by 1930-40 (to 95-97%).

The most we can learn from the Ó Gráda et al. paper is that ABE basic has weak performance *for some groups (e.g., the Irish) in the mid-19th century*. Yet, performance is strong for other groups in those years and for the Irish in the early 20th century, and using extra information particularly helps improve accuracy for the Irish.[22]

6. Conclusion

Over the past decade, digitization of historical data and the development of record linking algorithms have transformed economic history research, allowing scholars to create large historical panel datasets that shed light on central topics such as social mobility, inequality, immigration, the long-run effects of government programs, and the effects of childhood environment on economic outcomes.

In the absence of Social Security numbers, earlier linking algorithms relied on a small set of pre-determined matching attributes such as reported name, birthplace, and age (see, for instance, Ferrie 1996, Long and Ferrie 2013, and Abramitzky, Boustan, and Eriksson 2012, 2014, 2019). More recently, Helgertz, et al. (2022) and Price, et al. (2021) developed machine learning algorithms that include a broader set of matching characteristics, such as place of residence and the names of spouses, parents and other family members. These methods result in substantially increased match rates without reducing match quality. However, a disadvantage of these methods is that they rely on hand-linked training samples and so they are not fully automated, cannot be fully replicated from scratch and cannot be customized to different research settings.

---

[22] We generously received the data from the Ó Gráda et al. co-author team to compare agreement rates between their genealogy-quality hand links and matches made by ABE basic or ABE-EI. The dataset starts with ~1200 bank accounts, of which 600 are linked by the co-authors to both the 1860 and 1870 Census but only 397 could be linked to 'histids.' The linked data provided to us by the co-author team did not include 'histids.' The co-authors told us that we should be able to find most of these people by using variables like ward, dwelling number and family number. We were able to find around 400 of these individuals using this approach. Of these 397 cases, 78/397 (20%) are matched by some ABE basic algorithm and 156/397 (40%) are matched by some ABE-EI algorithm. For these matches, agreement rates ranged from 50-68% for ABE basic and from 60-76% for ABE-EI. These agreement rates are in the same range as the accuracy rates (compared to the Family Tree) reported above for the 1860-70 census pair.

We introduce and assess the performance of a new rule-based, fully automated linking algorithm that adds five matching characteristics – middle initial, county of residence, spouse's name, mother's name, and father's name – to the basic Abramitzky, Boustan, and Eriksson (2012, 2014, 2019) matching approach ("ABE"). We find that, relative to the basic ABE algorithm, this "ABE with Extra Information" (or "ABE-EI") algorithm produces samples with higher match rates and improved accuracy. These improvements do not come at a cost of population representativeness on most attributes, although ABE-EI is less able to link men who move out of their childhood home.

Relative to machine learning algorithms that include extra information, ABE-EI has the key advantages of being fully replicable and enabling researchers to flexibly adjust the set of characteristics used for matching. We provide a Stata command and an R code to implement this method.

The choice of linking algorithm, whether ABE basic, ABE-EI or machine learning approaches, has a minor effect on regression analysis in our test case. Our estimates of father-son rank-rank elasticities are higher with ABE-EI (or any approach with extra information) than with ABE basic but these differences are small (around 5%).

Given improvements in both match rate and accuracy, we conclude that ABE-EI will be preferable to ABE basic for many research applications. ABE-EI is particularly useful for research topics requiring large sample sizes, particularly questions relating to smaller geographic areas or specific population groups. However, all extra information algorithms (including ABE-EI) may be less well-suited for questions about geographic mobility, given that they rely in part on geography to make matches and thus may underrepresent men who move without other household members. However, we also caution that ABE basic includes more movers but at the cost of allowing for a greater number of false matches.

Choosing between procedural and machine-learning algorithms with extra information is a matter of taste: ABE-EI is more representative of the population, and is fully automated and replicable, while MLP and the machine learning algorithm from the Census Tree (CT-ML) have higher match rates. Using the full CT sample, including the family-based hand links, has the largest sample size, but is also the least representative of the population. The hand links are the only matched sample that can link women across the marital transition, when surname usually changes from maiden name to married name.

As in earlier work, we continue to recommend testing the robustness of results across linking approaches and weighting linked samples to address differences in observable attributes relative to the population. When using extra information, we also suggest linking adjacent census pairs, rather than linking across longer time spans, given that household structure and county of residence is less likely to change across ten years than at longer intervals. Longer linked datasets can be created by “chaining together” adjacent census pairs.

We are currently working to assess whether adding extra information can help improve match rates for specific subpopulations, including racial minorities and immigrant groups (Postel, et al. 2025). We note that the extra information algorithm can be extended to include information on the names of other household members, like siblings and children, and to prioritize “rare” (and thus more likely) matches based on name and county frequencies. The use of extra information may also give rise to linking algorithms that allow links to be created even if some pieces of identifying information do not match as long as a “sufficient” number of features do. We leave these and other extensions for future work.

References

Abramitzky, R., & Boustan, L. (2017). "Immigration in American Economic History". *Journal of Economic Literature*, *55*(4), 1311–45.

Abramitzky, R., Boustan, L., & Connor, D. (2020, June). *Leaving the Enclave: Historical Evidence on Immigrant Mobility from the Industrial Removal Office* (NBER Working Papers No. 27372). National Bureau of Economic Research.

Abramitzky, R., Boustan, L., & Connor, D. S. (2024). "Leaving the enclave: Historical evidence on immigrant mobility from the industrial removal office". *The Journal of Economic History*, *84*(2), 352–394.

Abramitzky, R., Boustan, L., & Eriksson, K. (2019). "To the New World and Back Again: Return Migrants in the Age of Mass Migration". *ILR Review*, *72*(2), 300–322.

Abramitzky, R., Boustan, L., & Eriksson, K. (2020). "Do Immigrants Assimilate More Slowly Today Than in the Past?" *American Economic Review: Insights*, *2*(1), 125–41.

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). "Automated Linking of Historical Data". *Journal of Economic Literature*, *59*(3), 865–918.

Abramitzky, R., Boustan, L., Eriksson, K., Rashid, M., & Pérez, S. (2022). *Census Linking Project* [Dataset].

Abramitzky, R., Boustan, L., Jacome, E., & Pérez, S. (2021). "Intergenerational Mobility of Immigrants in the United States over Two Centuries". *American Economic Review*, *111*(2), 580–608.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration". *American Economic Review*, *102*(5), 1832–56.

Abramitzky, R., Boustan, L. P., & 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.

Abramitzky, R., Mill, R., & Pérez, S. (2018, February). *Linking individuals across historical sources: A fully automated approach* (Working Paper No. 24324). National Bureau of Economic Research.

Ager, P., Boustan, L., & Eriksson, K. (2021). "The Intergenerational Effects of a Large Wealth Shock: White Southerners after the Civil War". *American Economic Review*, *111*(11), 3767–94.

Aizer, A., Early, N., Eli, S., Imbens, G., Lee, K., Lleras-Muney, A., & Strand, A. (2024). "The lifetime impacts of the new deal's youth employment program". *The Quarterly Journal of Economics*, *139*(4), 2579–2635.

Aizer, A., Eli, S., Lleras-Muney, A., & Lee, K. (2020, June). *Do youth employment programs work? evidence from the new deal* (Working Paper No. 27103). National Bureau of Economic Research.

Althoff, L., Brookes Gray, H., & Reichardt, H. (2025). *America's rise in human capital mobility* (CESifo Working Paper Series No. 11729) (CESifo Working Paper No. 11729). CESifo.

Bailey, M. J., Cole, C., Henderson, M., & Massey, C. (2020). "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data". *Journal of Economic Literature*, *58*(4), 997–1044.

Bailey, M. J., & Lin, P. Z. (2024, December). Marital Matching and Women's Intergenerational Mobility in the Late 19th and Early 20th Century US. In *The Economic History of American Inequality: New Evidence and Perspectives* (pp. 165–196). National Bureau of Economic Research, Inc.

Beach, B., Brown, R., Ferrie, J., Saavedra, M., & Thomas, D. (2022b). "Reevaluating the Long-Term Impact of In Utero Exposure to the 1918 Influenza Pandemic". *Journal of Political Economy*, *130*(7), 1963–1990.

Beach, B., Clay, K., & Saavedra, M. (2022a). "The 1918 Influenza Pandemic and Its Lessons for COVID-19". *Journal of Economic Literature*, *60*(1), 41–84.

Bleakley, H., & Ferrie, J. (2016). "Shocking behavior: Random wealth in antebellum georgia and human capital across generations". *The Quarterly Journal of Economics*, *131*(3), 1455–1495.

Buckles, K., Haws, A., Price, J., & Wilbert, H. E. (2023, September). *Breakthroughs in historical record linking using genealogy data: The census tree project* (Working Paper No. 31671). National Bureau of Economic Research.

Caprettini, B., & Voth, H.-J. (2023). "New Deal, New Patriots: How 1930s Government Spending Boosted Patriotism During World War II". *The Quarterly Journal of Economics*, *138*(1), 465–513.

Catron, P., Vignau Loría, M., & Farr, S. (2024). "Contextual boundaries: Skin tone stratification and skill transferability among mexicans in the age of mass migration". *Demography*, *61*(5), 1377–1402.

Dahis, R., Nix, E., & Qian, N. (2019, November). *Choosing Racial Identity in the United States, 1880-1940* (NBER Working Papers No. 26465). National Bureau of Economic Research, Inc.

Eli, S. (2015). "Income effects on health: Evidence from union army pensions". *The Journal of Economic History*, *75*(2), 448–478.

Eriksson, K., Rashid, M., Craig, J., & Niemesh, G. (2023). *Marriage and the intergenerational mobility of women: Evidence from marriage certificates 1850–1920* [Working paper].

Feigenbaum, J. (2016). *Automated census record linking: A machine learning approach* [Working Paper].

Feigenbaum, J., Lee, J., & Mezzanotti, F. (2022). "Capital Destruction and Economic Growth: The Effects of Sherman's March, 1850–1920". *American Economic Journal: Applied Economics*, *14*(4), 301–42.

Ferrie, J. (1996). "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules". *Historical Methods*, *29*(4), 141–156.

Ghosh, A., Hwang, S. I. M., & Squires, M. (2023). "Links and legibility: Making sense of historical u.s. census automated linking methods". *Journal of Business & Economic Statistics*, *42*(2), 579–590.

Goeken, R., Lee, Y. N., Lynch, T., & Magnuson, D. (2017). *Evaluating the accuracy of linked u.s. census data: A household linking approach* [Working

Paper No. 2017-1, University of Minnesota and Bethel University. Presented at The Systematic Linking of Historical Records, University of Guelph, May 10–13, 2017].

Helgertz, J., Price, J., Wellington, J., Thompson, K., Ruggles, S., & Fitch, C. (2022). "A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel". *Historical Methods*, *55*(1), 12–29.

Long, J., & Ferrie, J. (2013). "Intergenerational Occupational Mobility in Great Britain and the United States since 1850". *American Economic Review*, *103*(4), 1109–37.

Mattheis, R. (2024). *Spurious mobility in imperfectly linked historical data* [Working Paper].

Michelman, V., Price, J., & Zimmerman, S. D. (2022). "Old Boys' Clubs and Upward Mobility Among the Educational Elite". *The Quarterly Journal of Economics*, *137*(1), 845–909.

Mill, R., & Stein, L. C. (2016). *Race, skin color, and economic outcomes in early twentieth-century america* [Working Paper].

Modrek, S., Roberts, E., Warren, J. R., & Rehkopf, D. (2022). "Long-Term Effects of Local-Area New Deal Work Relief in Childhood on Educational, Economic, and Health Outcomes Over the Life Course: Evidence From the Wisconsin Longitudinal Study". *Demography*, *59*(4), 1489–1516.

Ó Gráda, C., Anbinder, T., Connor, D., & Wegge, S. A. (2023). "The problem of false positives in automated census linking: Nineteenth-century new york's irish immigrants as a case study". *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *56*(4), 240–259.

Pérez, S. (2017). "The (south) american dream: Mobility and economic outcomes of first- and second-generation immigrants in nineteenth-century argentina". *The Journal of Economic History*, *77*(4), 971–1006.

Postel, H., Abramitzky, R., Boustan, L., & Eriksson, K. (2025). *Subgroup disparities in automated census record linkage* [Unpublished manuscript].

Postel, H. M. (2023). "Record linkage for character-based surnames: Evidence from chinese exclusion". *Explorations in Economic History*, *87*, 101493.

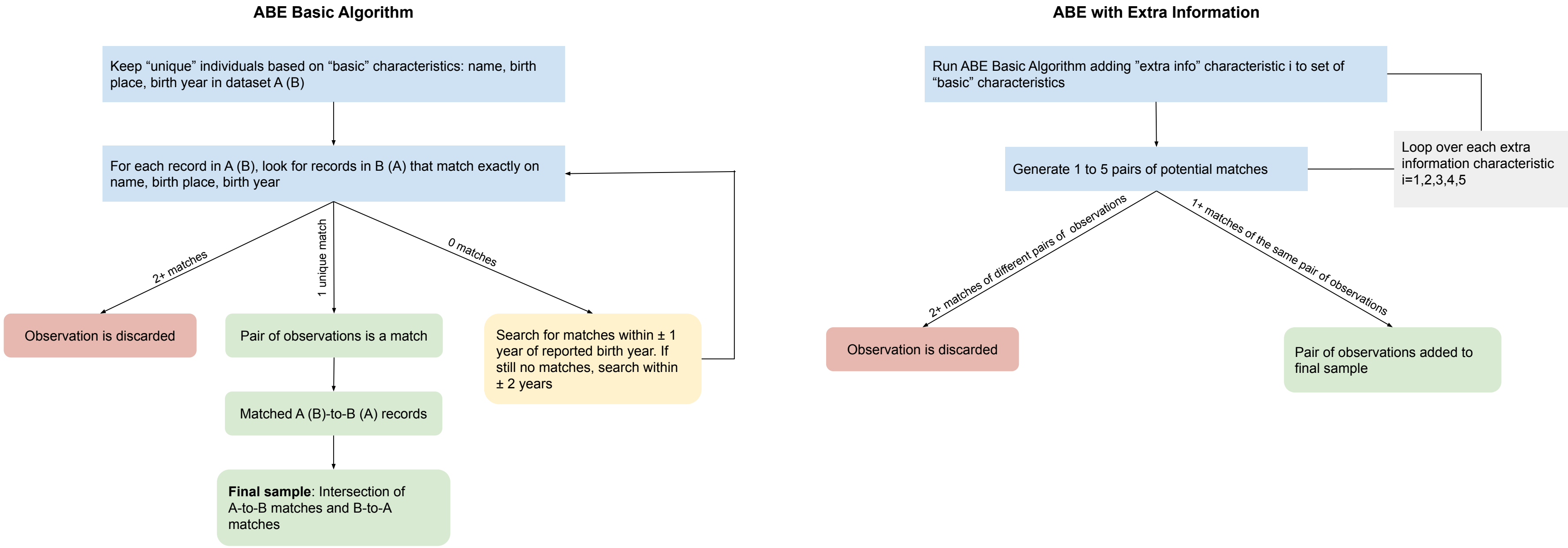
Price, J., Buckles, K., Van Leeuwen, J., & Riley, I. (2021). "Combining family history and machine learning to link historical records: The Census Tree data set". *Explorations in Economic History*, *80*(100).

Protzer, E., Orazbayev, S., Gomez-Lievano, A., Hartog, M., & Neffke, F. (2024). *A new algorithm to efficiently match u.s. census records and balance representativity with match quality* (Working Paper). Growth Lab, Harvard Kennedy School of Government.

Querubin, P., & Jr., J. M. S. (2013). "The control of politicians in normal times and times of crisis: Wealth accumulation by u.s. congressmen, 1850–1880". *Quarterly Journal of Political Science*, *8*(4), 409–450.

Sasaki, Y., & Zimran, A. (2025, May). *Using discrepancies to correct for false matches in historical linked data* (Working Paper No. 33881). National Bureau of Economic Research.

Zimran, A. (2020). "Recognizing sample-selection bias in historical data". *Social Science History*, *44*(3), 525–554.



Panel A: Flow Chart of ABE Basic Algorithm

Panel B: Flow Chart of ABE-EI Algorithm

FIGURE 1: **Flow Chart of ABE Algorithms**

Note: This flow chart shows the matching procedure of ABE-EI algorithm.

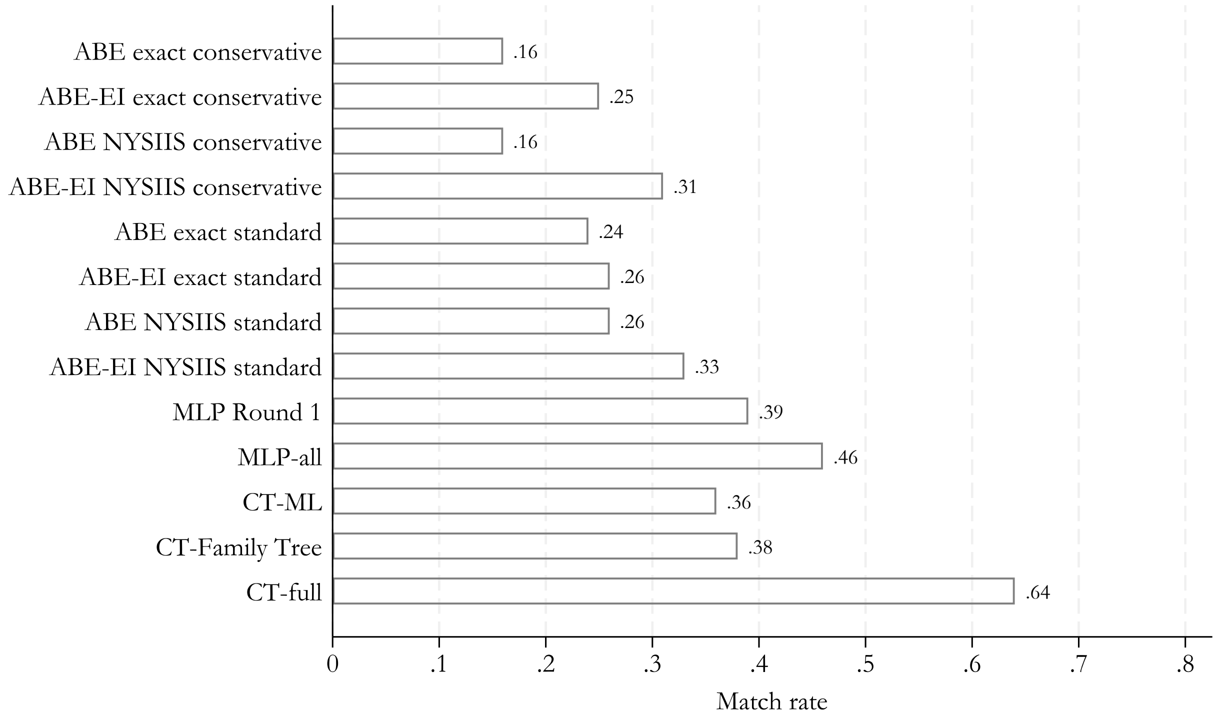


Figure 2: **Match rates**

Note: This figure shows the match rates (1900-1910) of various matching algorithms, where match rate is defined as the percentage of linked individuals out of the full count 1900 Census (men only). The ABE-Extra Information method uses first name, last name, place of birth, age, middle name initial, mother's first name, father's first name, county of residence, and spouse's first name as matching variables. This is in contrast to the basic ABE method that only uses first name, last name, place of birth and age.

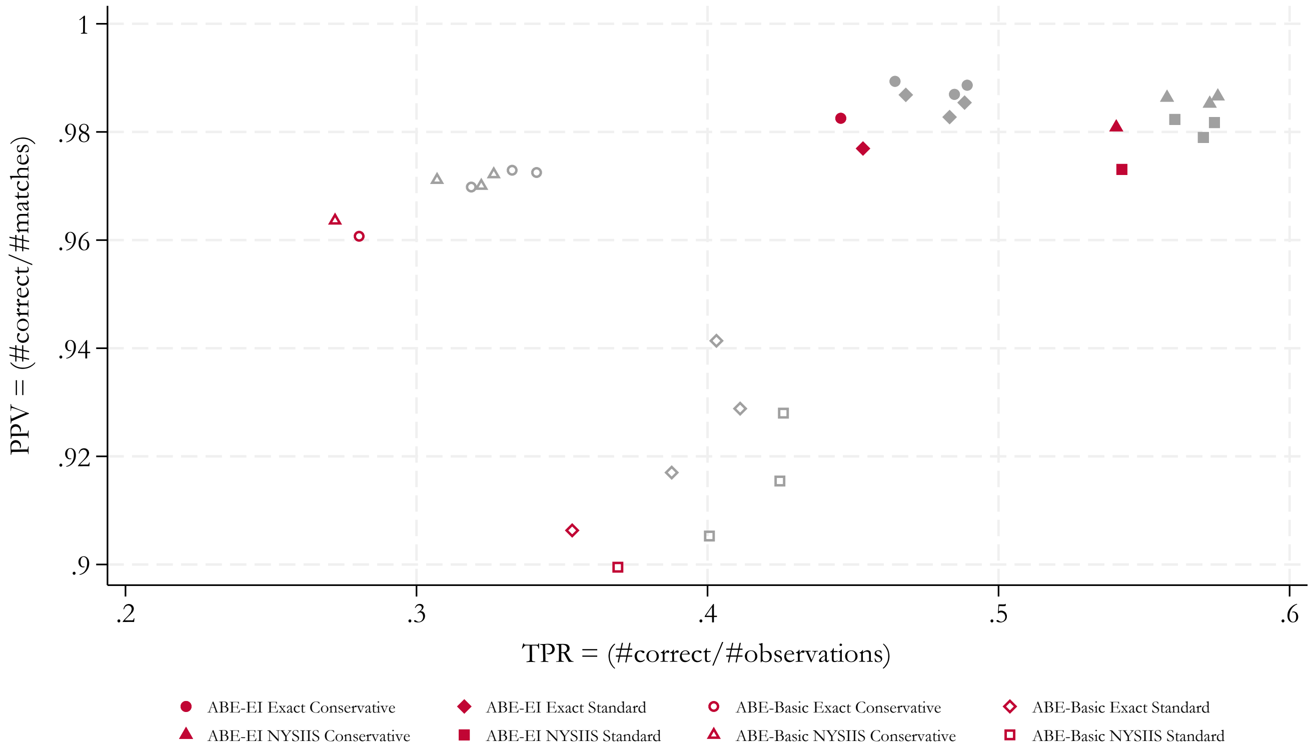


FIGURE 3: **Accuracy versus efficiency - comparing fully automated linking algorithms**

Note: This figure displays PPV (#correct/#matches) and TPR (#correct/#observations) for the various ABE-EI and ABE-Basic matching algorithms, where #matches is the number of individuals matched both by the ABE method in question and the Census Tree Family Tree, #correct is the subset of #matches for which the ABE method and Census Tree Family Tree agree on the match, and #observations is the number of individuals matched by the Census Tree Family Tree. The red symbols represent 1900-1910, whereas the gray symbols represent the adjacent decadal pairs from 1910 to 1940.

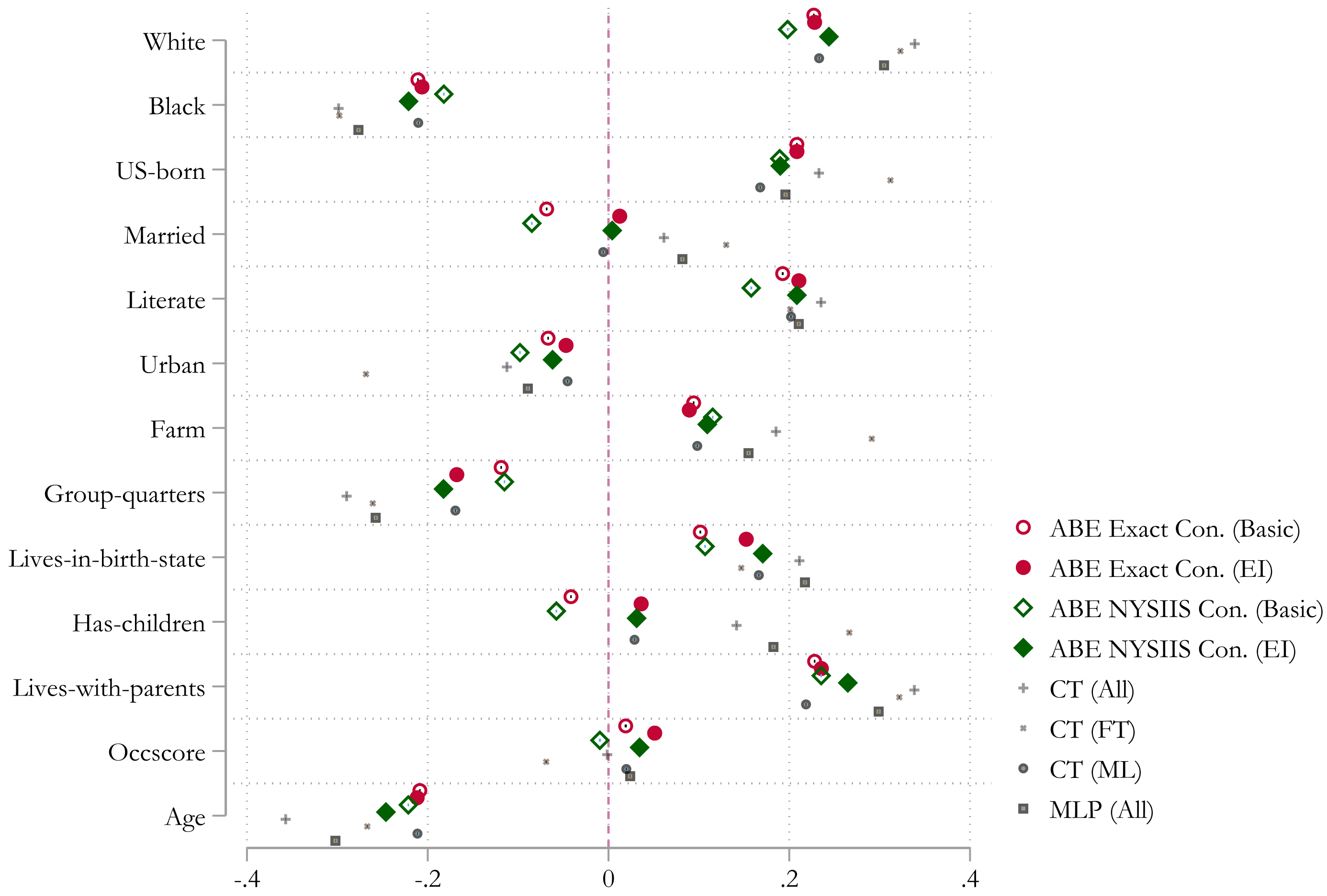
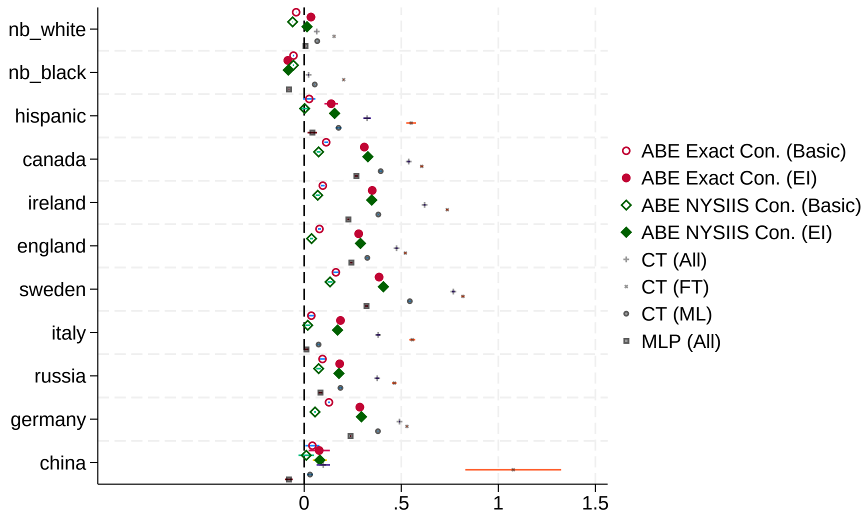
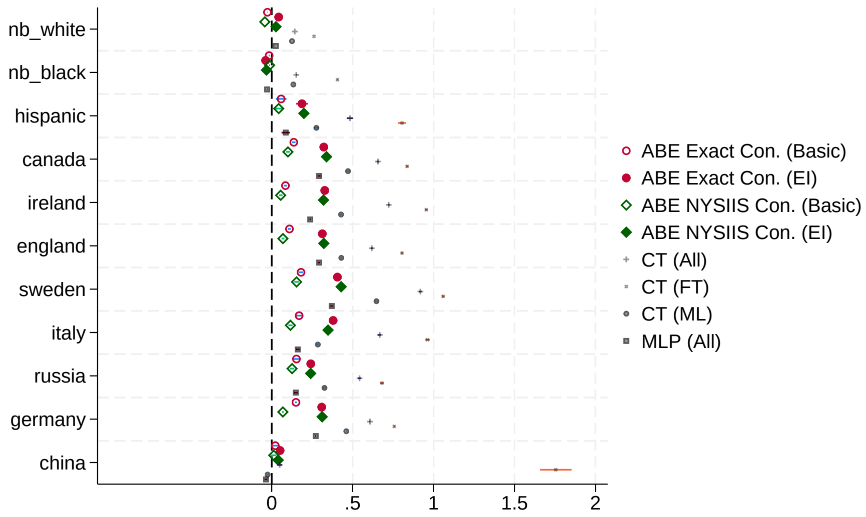


FIGURE 4: **Representativeness relative to base year Census male population (1900)**

Note: This figure presents a coefficient plot of the representativeness of 1900-1910 matches generated by the ABE, Census Tree, and MLP algorithms compared to the 1900 full count Census. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.



Panel A: Married



Panel B: Has Children

FIGURE 5: **Representativeness relative to base year Census male population (1900), subgroups**

Note: This figure presents a coefficient plot of the representativeness of 1900-1910 matches by subgroup generated by the ABE, Census Tree, and MLP algorithms compared to the 1900 full count Census. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.

Table 1: Additional matching features create uniqueness

First Name	Last Name	Birth-place	Age	Mother's name	Unique in ABE basic	Unique in ABE-EI
John	Smith	Alabama	7	Mary	No	Yes
John	Smith	Alabama	7	Ellen	No	Yes

Note: This table displays two individuals with identical first and last names, the same place of birth, and ages within a 5-year band. While unique matching based solely on these baseline features is not feasible, the inclusion of mother's name as a matching variable demonstrates the potential for successful matching by incorporating extra information.

TABLE 2: Comparing accordance rates with Family Tree hand-made links for preferred ABE-EI and two possible alternate ABE-EI algorithms, 1900-10

	ABE Basic		ABE-EI		ABE-EI + 4		Only immutable	
	N	Acc. Rate	N	Acc. Rate	N	Acc. Rate	N	Acc. Rate
Exact Conservative	6,357,902	96.07	9,713,326	98.25	11,309,341	95.66	8,388,520	95.03
NYSIIS Conservative	6,325,878	96.36	12,097,709	98.09	13,763,181	95.93	9,331,791	94.83
Exact Standard	9,269,657	90.63	10,185,582	97.69	11,673,774	93.66	10,807,647	90.82
NYSIIS Standard	10,159,055	89.95	12,636,630	97.31	14,193,592	93.73	12,673,366	89.78

Note: This table presents the match counts and accordance rates with the Census Tree Family Tree for ABE Basic and ABE-EI as well as modified versions of ABE-EI where either four additional variables vote (state of residence, parental nativity, year of immigration, and race) or only the 'immutable' variables vote (middle initial, parental nativity, race, year of immigration). In both new cases, unanimity is enforced during the voting.

Table 3: Attributes of matching features in 1900

	(1) Core matching variables	(2) Mother's first name	(3) Father's first name	(4) Spouse first name	(5) Middle initial	(6) County of residence
% records unique	78.18	91.13	90.23	89.37	90.47	97.10
% records not missing	100	44.5	40.54	38.65	35.88	100

Note: This table presents the impact of each matching variable on the percent of men in the full count 1900 Census who are uniquely identifiable, as well as the percent of that population for whom each variable is not missing. Column (1) presents the percentage of men in the Census who can be uniquely identified based on exact first name, exact last name, birthplace, and age within a 5-year band. The remaining columns present the percent of men in the Census who can be uniquely identified based on the variables used in Column (1) and the additional variable described in the column title. The columns are not cumulative (for example, Column (3) does not use mother's first name as a variable). "Not missing" is defined as the variable being present for an individual in the Census, as opposed to being left blank.

TABLE 4: Specific examples of conflicting matches rejected by ABE-EI

Individual	First Name	Last Name	Age	Birth-place	Middle initial	Spouse's name	Mother's name	Father's name	County of residence
1900 source	Edward	Hoffman	3	NY			Wilhelmina	Charles	NY 470
Mother's name match	Edward	Hoffman	13	NY			Wilhelmina	Herman	NY 610
Father's name match	Edward	Hoffman	12	NY			Mary	Charles	NY 610
1900 source	Daniel	Lehr	32	Illinois			Johanna	Henry	IL 1630
County match	Daniel	Lehr	42	Illinois		Wilhelmina			IL 1630
Father's name match	Daniel	Lehr	42	Illinois		Mageretha		Henry	IL 1890

Note: These two examples indicate how the matching variables of ABE-EI can link the same individual to different people in the following Census year. The groups of rows above present the personal characteristics of different individuals from 1910 to which one individual in 1900 was matched. Because of the unanimity requirement of the ABE-EI matching algorithm, these individuals, as well as their 1900 'source', were removed from the linked sample.

TABLE 5: ABE Basic and ABE-EI accordance rates with CT Family Tree hand-made links, and additional comparisons against CT machine learning links and MLP (1900-1910)

		Accuracy proxy	**Additional comparisons**	
	N	CT Family Tree	CT ML	MLP
ABE Basic				
Exact Conservative	6,357,902	96.07	97.90	96.55
NYSIIS Conservative	6,325,878	96.36	99.06	97.00
Exact Standard	9,269,657	90.63	93.22	91.48
NYSIIS Standard	10,159,055	89.95	93.96	90.68
ABE Extra Information				
Exact Conservative	9,713,236	98.25	98.71	98.43
NYSIIS Conservative	12,097,709	98.09	98.88	98.32
Exact Standard	10,185,582	97.69	98.14	97.83
NYSIIS Standard	12,636,630	97.31	98.14	97.47

Note: This table presents the accordance rate of ABE Basic and ABE-EI links with the Census Trees's Family Tree links (CT FT), the Census Tree's Machine Learning links (CT ML), and MLP links. The Census Tree links are obtained from censustree.org/data. The Family Tree links, generated by individual users who manually connect relatives to census records, are considered to be highly accurate. For each version of ABE Basic and ABE-EI, the number of matches made by that method is reported ('N'). The accordance rate is the percent of links made by both methods that are identical (i.e., conditional on both methods attempting to match the same individual from 1900 to 1910, the percent of the time that the two methods pick the same individual from 1910 as the match).

Table 6: ABE-EI women match rates and accordance rates with CT Family Tree hand-made links, 1900-1910

	N	Match rate	**Accuracy proxy**	**Additional comparisons**	
			CT Family Tree	CT ML	MLP
ABE Basic					
Never Married					
Exact conservative	2,774,146	13.52	96.63	98.26	96.35
NYSIIS conservative	2,893,839	14.11	96.8	99.31	96.64
Exact standard	3,890,662	18.97	92.92	95.14	92.37
NYSIIS standard	4,569,766	22.28	92.21	95.73	91.78
Ever Married					
Exact conservative	2,697,527	16.15	94.87	97.59	94.72
NYSIIS conservative	2,759,477	16.52	94.77	98.65	94.7
Exact standard	3,890,940	23.29	87.98	91.98	87.43
NYSIIS standard	4,471,987	26.77	86	91.8	85.48
ABE Extra Information					
Never Married					
Exact conservative	4,231,910	20.63	96.89	98	96.58
NYSIIS conservative	5,288,486	25.78	97.04	98.74	96.85
Exact standard	4,425,871	21.58	95.63	97.02	95.07
NYSIIS standard	5,603,601	27.32	95.5	97.79	95.04
Ever Married					
Exact conservative	3,727,143	22.31	98.3	98.92	98.14
NYSIIS conservative	4,668,146	27.94	98.13	99.02	97.99
Exact standard	3,845,109	23.02	97.56	98.29	97.22
NYSIIS standard	4,865,450	29.12	97.16	98.23	96.79

Note: This table presents the accordance rate of ABE-EI women (never married and ever married) links with the Census Trees's Family Tree links (CT FT), the Census Tree's Machine Learning links (CT ML), and MLP links. The Census Tree links are obtained from censustree.org/data. The Family Tree links, generated by individual users who manually connect relatives to census records, are considered to be highly accurate. For each version of ABE, the number of matches made by that method is reported ('N'). The accordance rate is the percent of links made by both methods that are identical (i.e., conditional on both methods attempting to match the same individual from 1900 to 1910, the percent of the time that the two methods pick the same individual from 1910 as the match). The coverage of the CT Family Tree represents the share of Family Tree links also made by ABE-EI.

Table 7: Coefficients on father's income rank from intergenerational mobility regressions for various matching algorithms

	Unweighted		Weighted	
	Basic	EI	Basic	EI
Exact Conservative	0.4000***	0.4060***	0.3920***	0.3969***
	(620,338)	(887,092)	(620,338)	(887,092)
NYSIIS Conservative	0.4013***	0.4063***	0.3924***	0.3957***
	(709,223)	(1,201,528)	(709,223)	(1,201,528)
Exact Standard	0.3865***	0.3965***	0.3782***	0.3877***
	(884,079)	(853,165)	(884,079)	(853,165)
NYSIIS Standard	0.3827***	0.3940***	0.3740***	0.3844***
	(1,105,261)	(1,129,761)	(1,105,261)	(1,129,761)
CT (All)	0.4009***		0.3922***	
	(4,088,300)		(4,088,300)	
CT (ML)	0.4086***		0.4000***	
	(1,595,579)		(1,595,579)	
MLP	0.4229***		0.3750***	
	(1,661,779)		(1,661,779)	

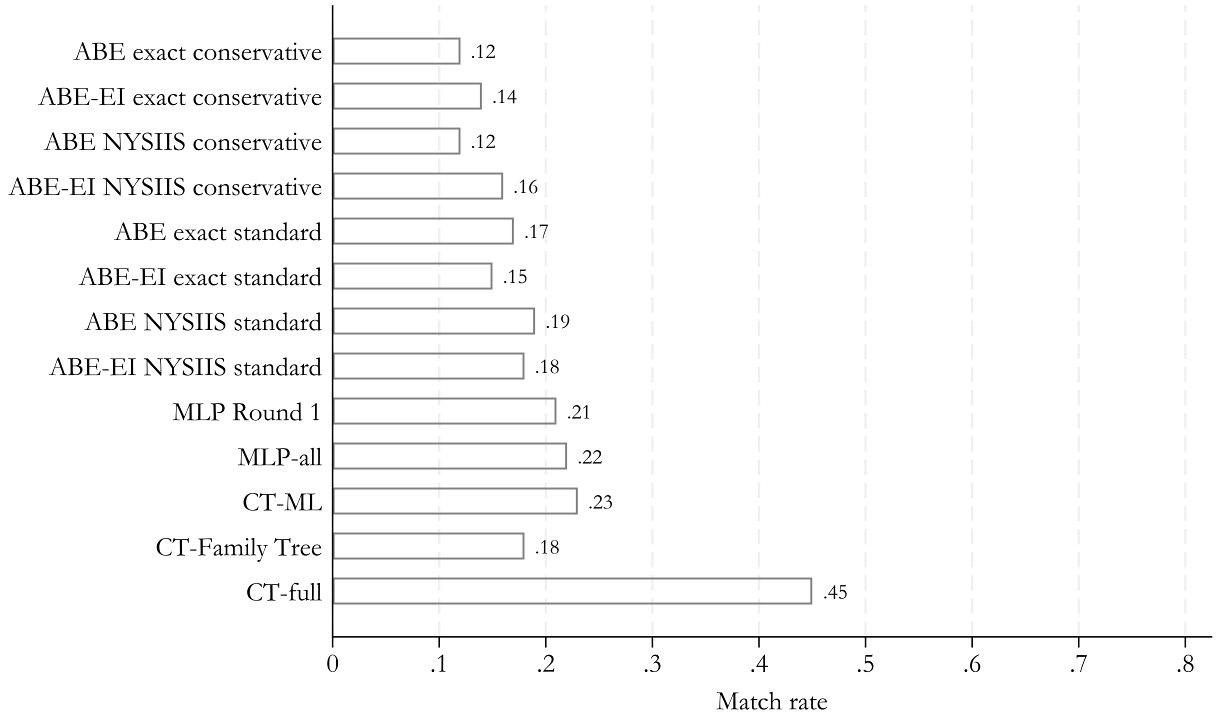
Note: The different datasets use three sets of crosswalks strung together to create a 1910-1940 sample: 1910-1920, 1920-1930 and 1930-1940 crosswalks. We also run the same regressions using the complete Census Tree data, the Census Tree Machine Learning data, and MLP. This sample consists of white men who were below the ages of 17 and whose fathers were between 30 and 50 years of age in 1910. Income rank is generated using income score. Weights are used to match the linked sample to the full count Census demographics across the following characteristics: US-born, married, living in an urbane environment, living on a farm, age (split into deciles), residing in state of birth, parents present in household, and has children. Sample sizes, presented in parentheses, range from 620,338 to 4,088,300. Standard errors range from 0.0005 to 0.0015, and all coefficients are significant at the .01 level. Note that neither CT or MLP has either 'basic' or 'EI' versions, so those results are all presented in the 'basic' column solely for ease of viewing.

TABLE 8: ABE Basic and ABE-EI match rate for subgroups, and Accuracy rates with CT Family Tree hand-made links, 1900-1910

	Total			Native-born White			Native-born Black			Hispanic			Canada			Sweden		
	N	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate
ABE Basic																		
NYSIIS Conservative	6,325,878	16.35	96.36	5,332,217	18.69	96.65	402,716	9.32	93.38	6,747	7.94	91.22	47,805	15.41	96.03	26,604	8.19	93.14
ABE Extra Information																		
NYSIIS Conservative	12,097,709	31.26	98.09	10,091,351	35.37	98.16	772,388	17.87	97.36	5,878	6.92	96.89	88,905	28.65	98.31	77,050	23.75	96.81
MLP																		
Round 1	14,764,543	38.15		12,326,718	43.2		883,086	20.44		9,306	10.95		98,015	31.59		90,279	27.83	
All	17,097,065	44.18		14,115,670	49.47		1,078,439	24.96		12,593	14.82		113,795	36.67		112,707	34.74	
Census Tree																		
All	24,976,021	64.54		19,999,158	70.09		1,954,129	45.22		24,605	28.95		180,850	58.28		191,340	58.98	
Family Tree	14,548,042	37.59		12,587,387	44.12		773,654	17.9		7,051	8.30		89,816	28.95		71,383	22	
Machine Learning	14,075,305	36.37		11,516,242	40.36		978,554	22.65		7,585	8.92		103,763	33.44		104,070	32.08	

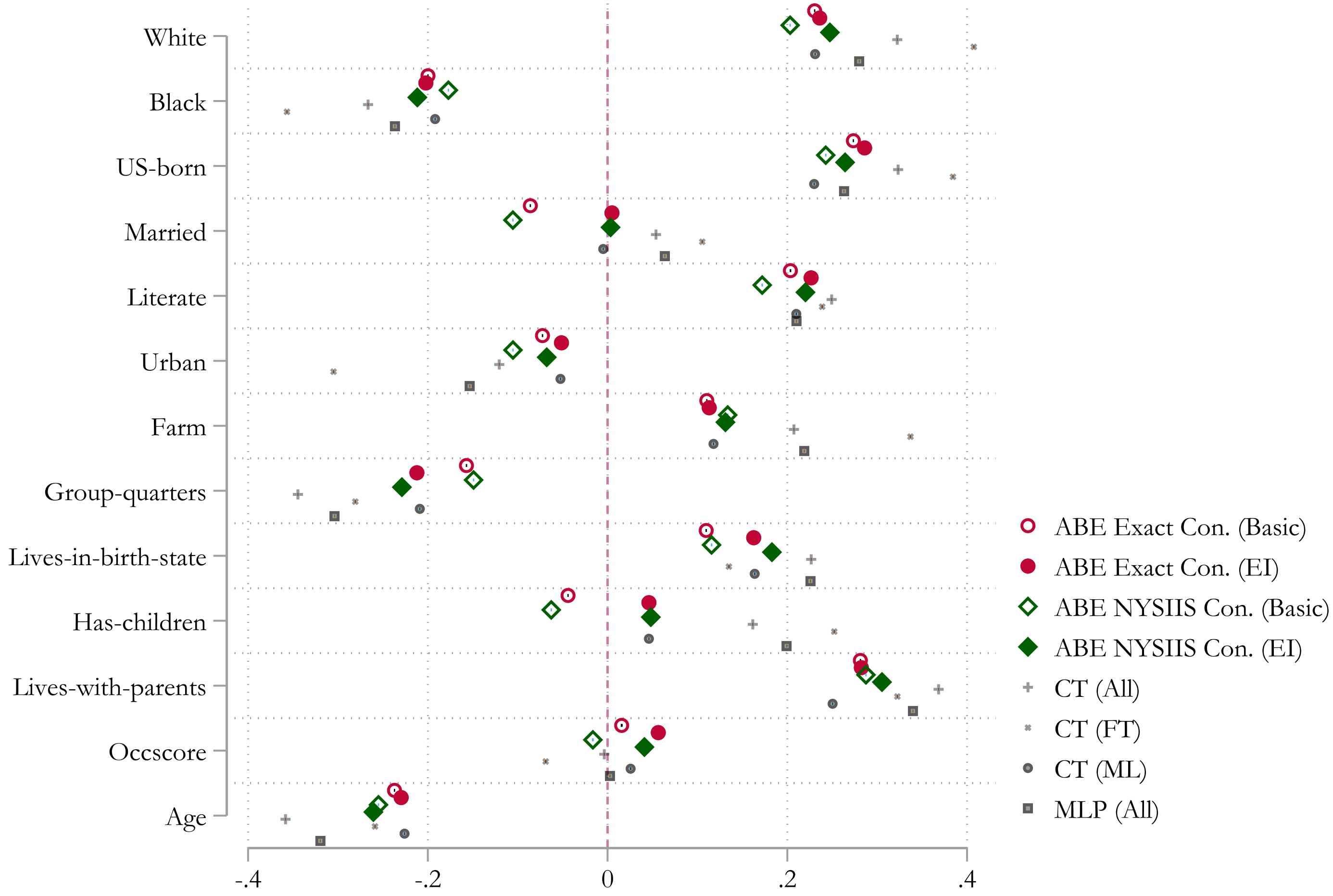
	England			Ireland			Italy			Germany			Russia			China		
	N	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate	*N*	Match rate	Accuracy Rate
ABE Basic																		
NYSIIS Conservative	66,781	14.42	95.99	36,440	4.85	91.38	27,637	8.72	87.51	164,519	11.43	93.90	28,522	12.27	89.44	2,488	2.58	90
ABE Extra Information																		
NYSIIS Conservative	145,476	31.4	98.51	137,851	18.34	95.44	29,424	9.29	95.77	379,742	26.39	98.01	43,301	18.63	95.48	3,036	3.15	83.33
MLP																		
Round 1	162,952	35.18		167,155	22.24		44,711	14.11		521,360	36.23		70,059	30.15		3,112	3.23	
All	184,628	39.85		216,970	28.87		54,518	17.21		603,664	41.95		86,858	37.38		3,399	3.53	
Census Tree																		
All	274,607	59.28		367,519	48.9		119,723	37.79		829,667	57.65		135,501	58.31		21,440	22.26	
Family Tree	131,520	28.39		101,355	13.49		18,626	5.88		369,694	25.69		35,861	15.43		64	.07	
Machine Learning	168,937	36.47		195,226	25.98		44,399	14.01		452,672	31.45		56,920	24.49		10,818	11.23	

Note: This table presents the match rate of Basic and ABE-EI links on the Irish population, as well as their accuracy rate (Acc. rate) with the Census Trees's Family Tree links. The Census Tree links are obtained from censustree.org/data. The Family Tree links, generated by individual users who manually connect relatives to census records, are considered to be highly accurate. For each version of ABE, the number of matches made by that method is reported ('N'). The accordance rate is the percent of links made by both methods that are identical (i.e., conditional on both methods attempting to match the same individual from 1900 to 1910, the percent of the time that the two methods pick the same individual from 1910 as the match).



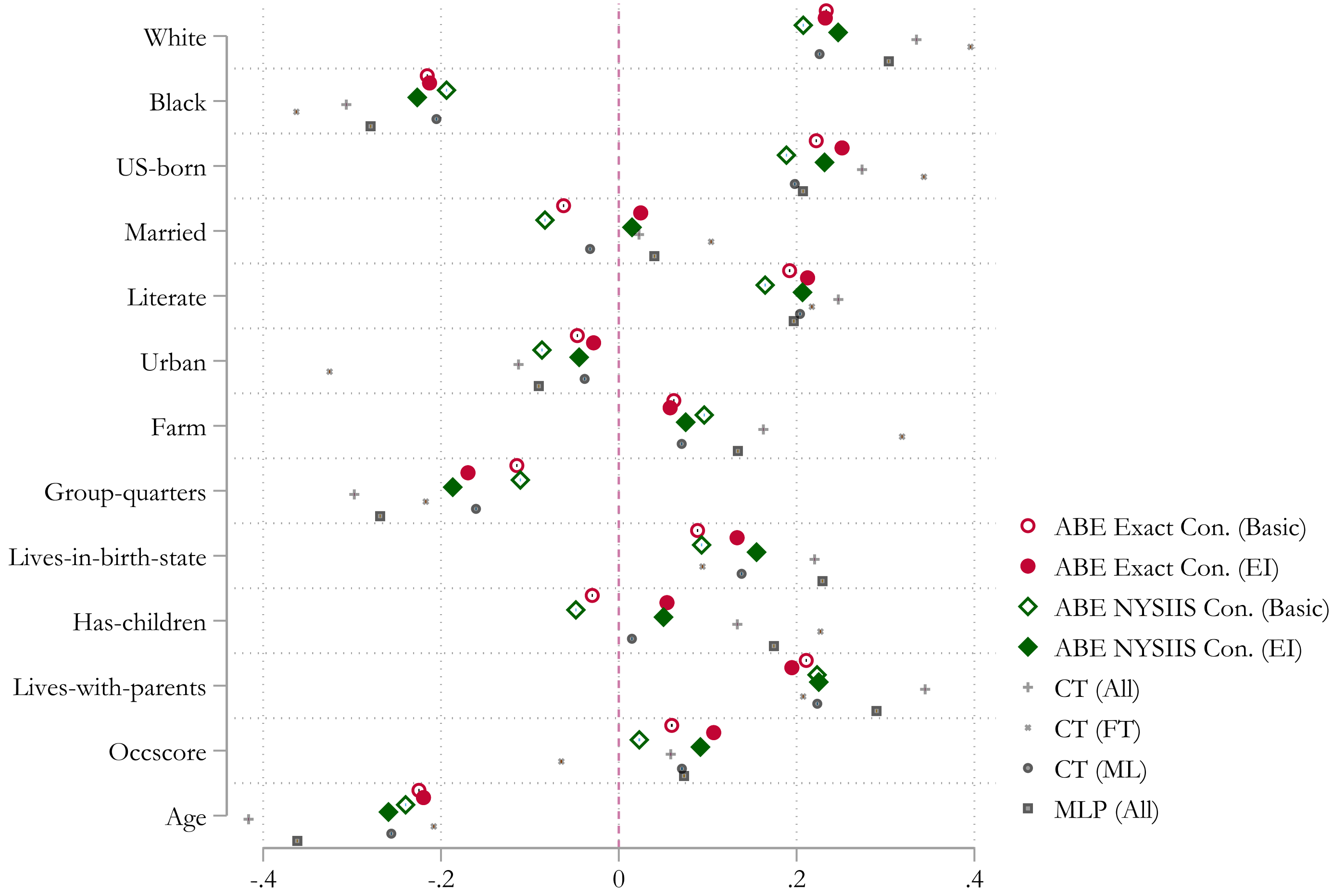
Appendix Figure 1: Match rates (1900-1930)

Note: This figure shows the match rates (1900-1930) of various matching algorithms, where match rate is defined as the percentage of linked individuals out of the full count 1900 Census (men only). The ABE-Extra Information method uses first name, last name, place of birth, age, middle name initial, mother's first name, father's first name, county of residence, and spouse's first name as matching variables. This is in contrast to the basic ABE method that only uses first name, last name, place of birth and age.



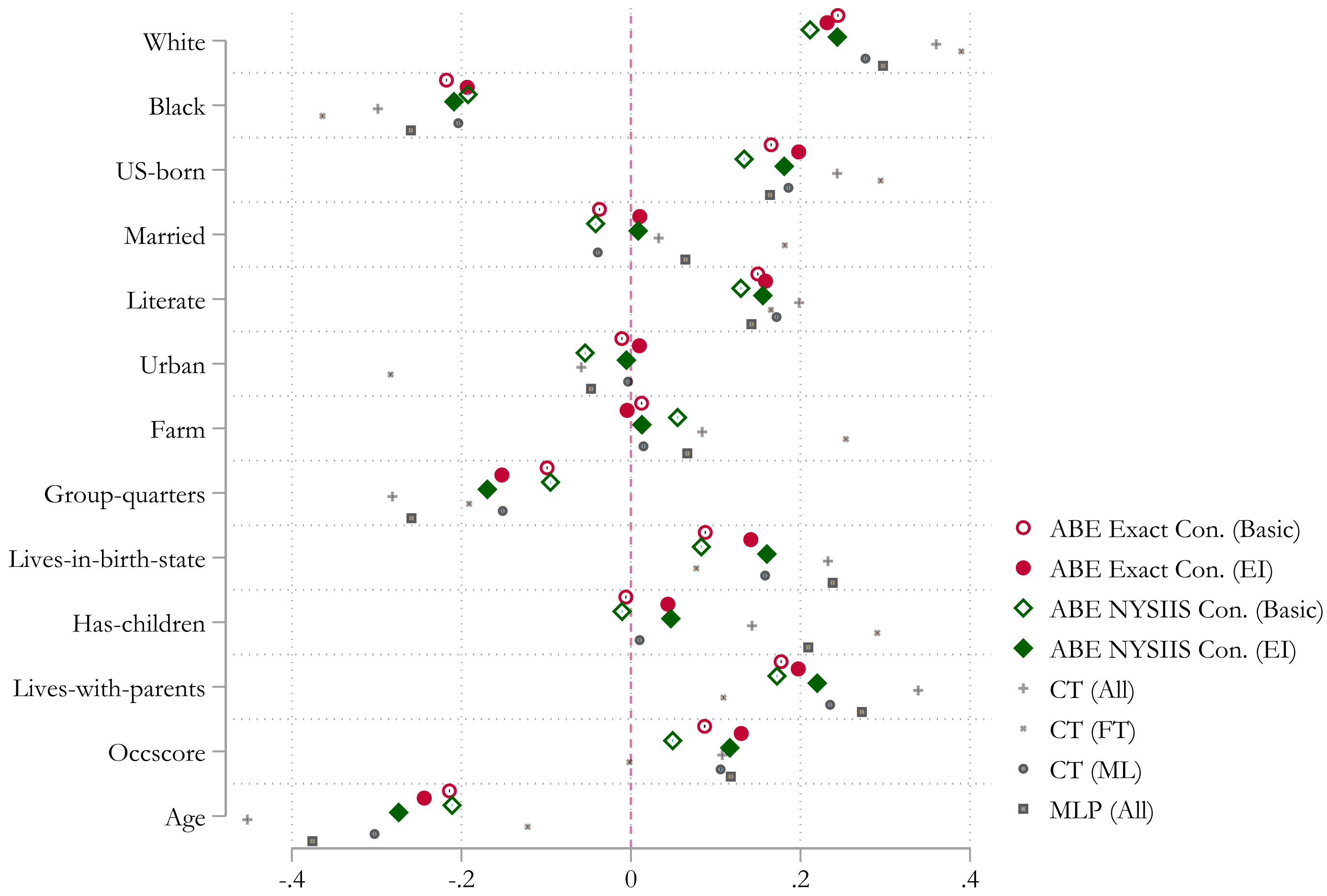
Appendix Figure 2A: Representativeness relative to base year Census male population (1910)

Note: This figure presents a coefficient plot of the representativeness of 1910-1920 matches generated by the ABE, Census Tree, and MLP algorithms compared to the 1910 full count Census. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.



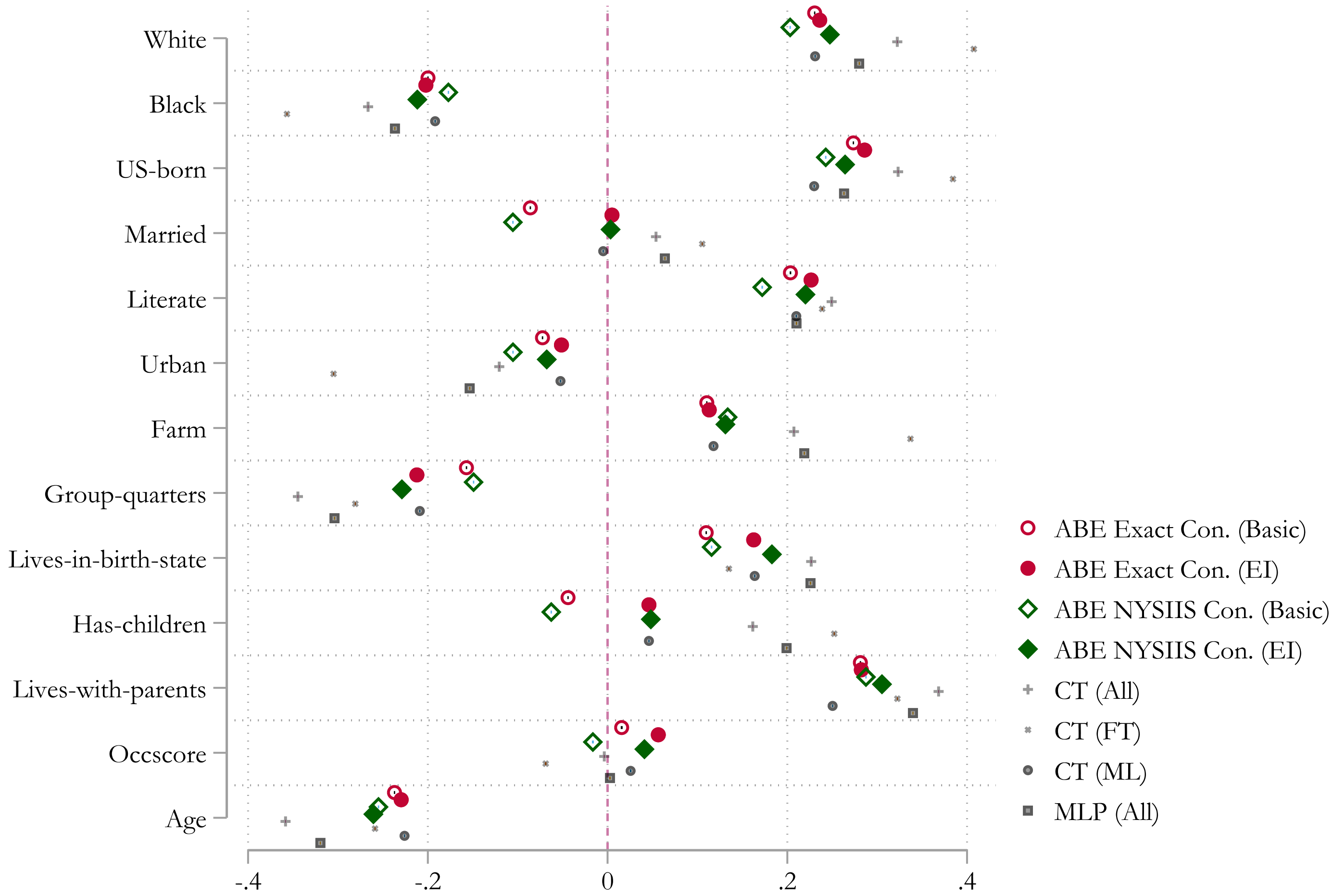
Appendix Figure 2B: Representativeness relative to base year Census male population (1920)

Note: This figure presents a coefficient plot of the representativeness of 1920-1930 matches generated by the ABE, Census Tree, and MLP algorithms compared to the 1920 full count Census. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.



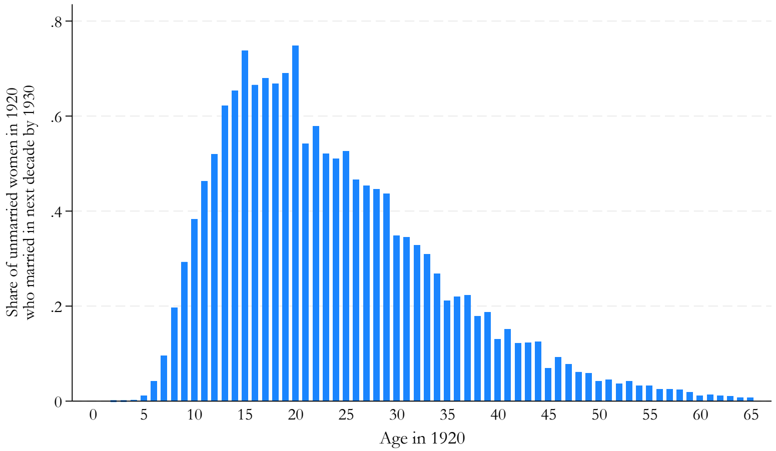
Appendix Figure 2C: Representativeness relative to base year Census male population (1930)

Note: This figure presents a coefficient plot of the representativeness of 1930-1940 matches generated by the ABE, Census Tree, and MLP algorithms compared to the 1930 full count Census. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.

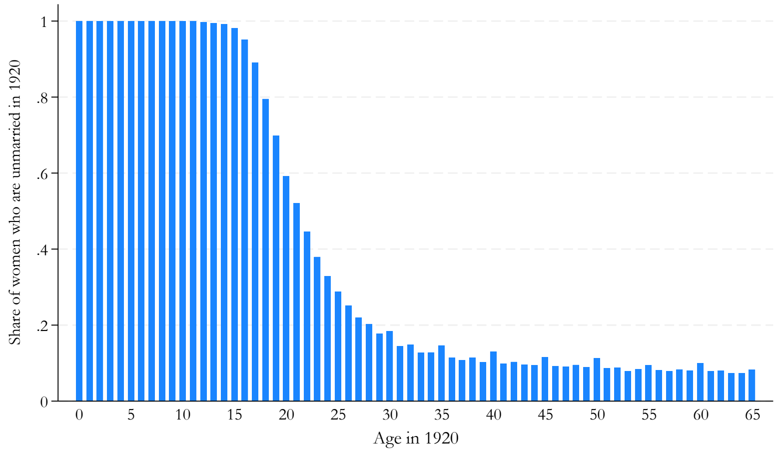


Appendix Figure 3: Representativeness relative to end year Census male population (1910)

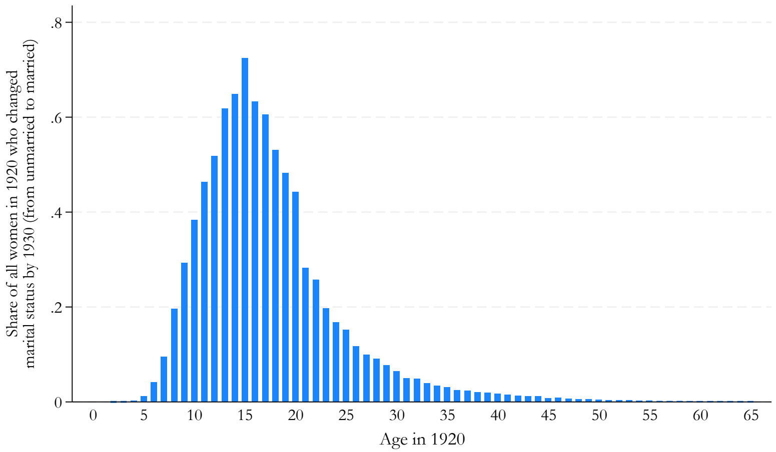
Note: This figure presents a coefficient plot of the representativeness of 1900-1910 matches generated by three geographic variants of ABE-EI: (a) the baseline ABE-EI algorithm version (which includes county), (b) a version with state instead of county county, and (c) a version without county. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.



Panel A: Share of newly married women between 1920-1930 by age



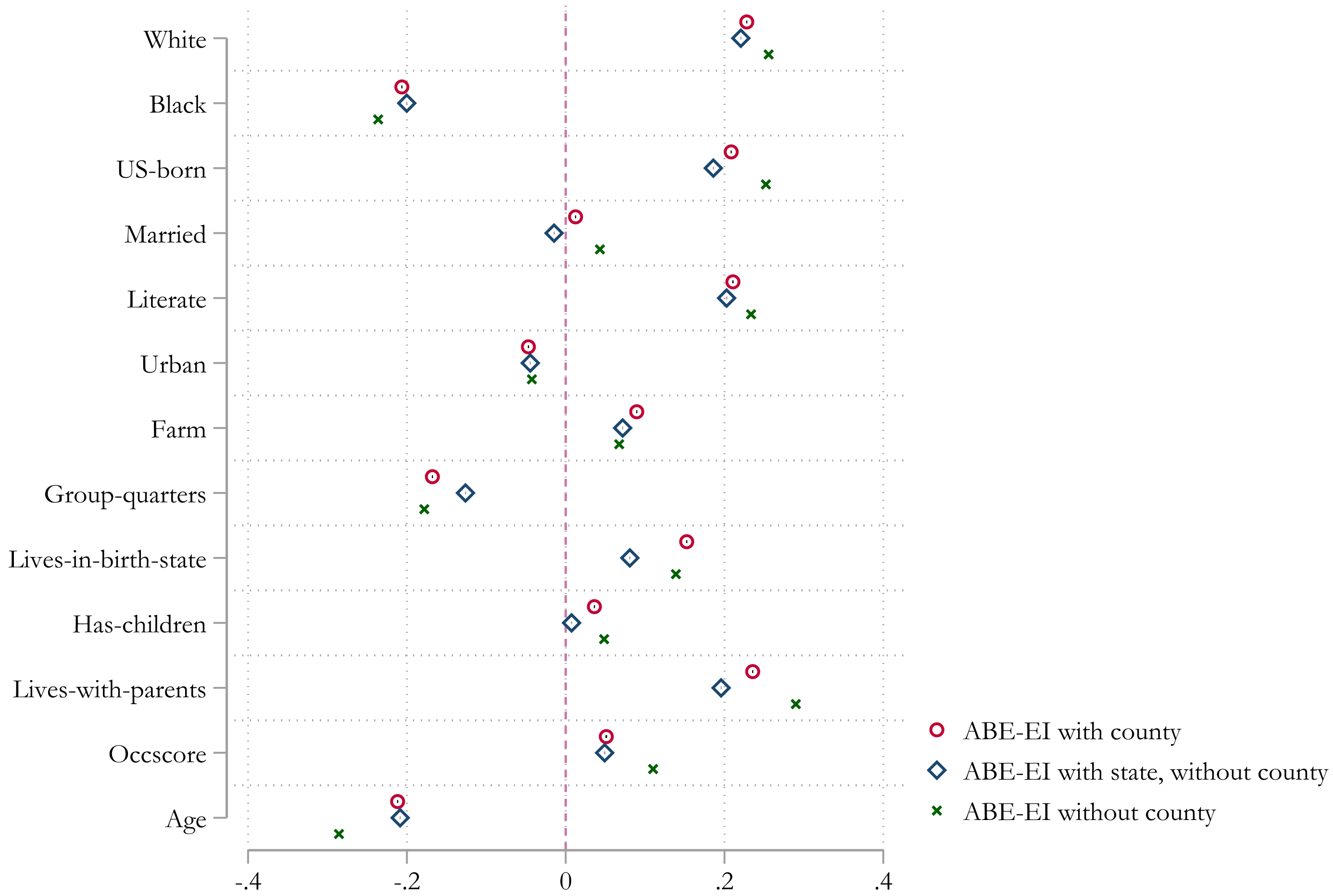
Panel B: Share of unmarried women by age, 1920



Panel C: Share of women who changed marital status by age, 1920–1930

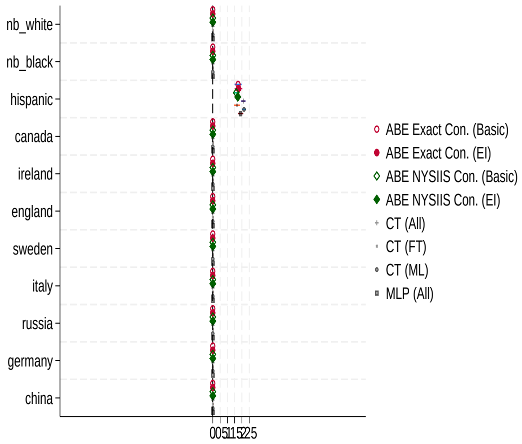
Appendix Figure 4: Marriage Demographics of Women, 1920–1930

Note: Panel A presents the share of unmarried women, by age (truncated for clarity), who married between 1920 and 1930. Panel B shows the share of unmarried womenby age in the 1920 census. Panel C presents the share of women who were not married in the 1920 census, and married by 1930. The data come from the 1920 and 1930 U.S Censuses.

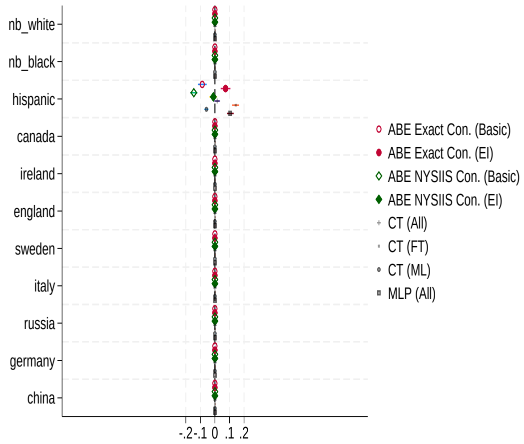


Appendix Figure 5: Representativeness of ABE-EI geographic variants relative to base year Census male population (1900)

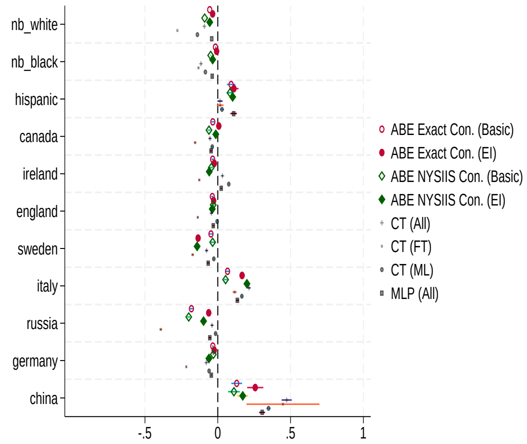
Note: This figure presents a coefficient plot of the representativeness of 1900-1910 matches generated by three geographic variants of ABE-EI: (a) the baseline ABE-EI algorithm version (which includes county), (b) a version with state instead of county county, and (c) a version without county. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.



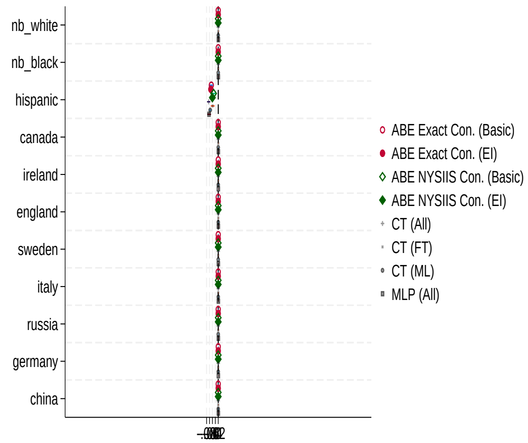
Panel A: White



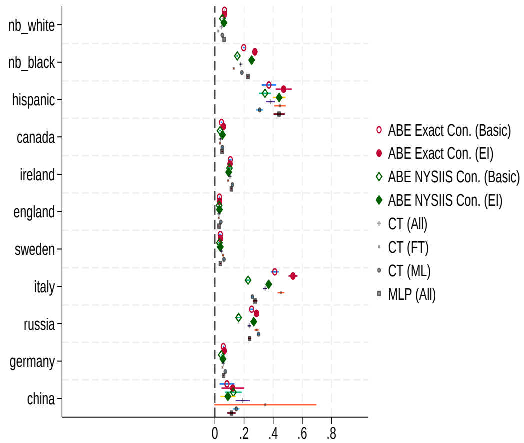
Panel C: US-born



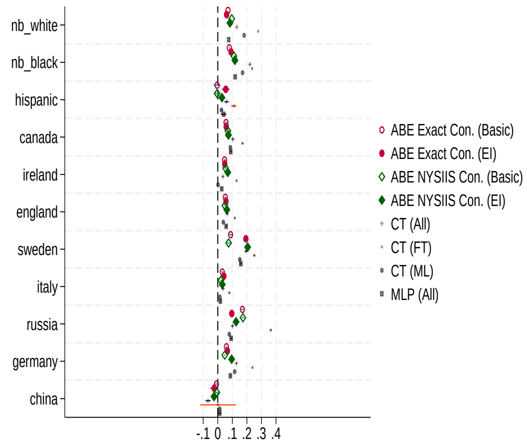
Panel E: Urban



Panel B: Black

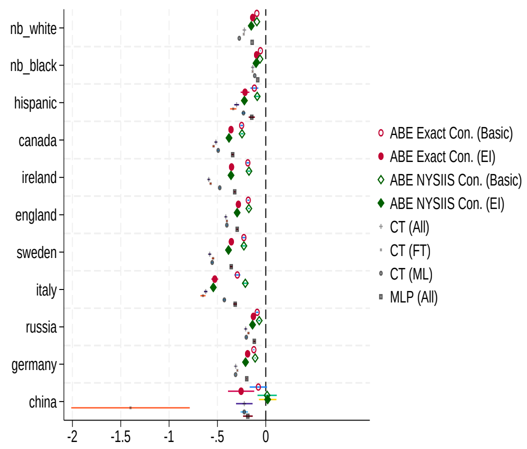


Panel D: Literate

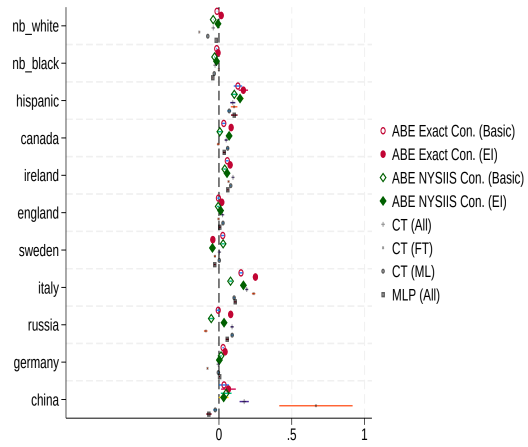


Panel F: Farm

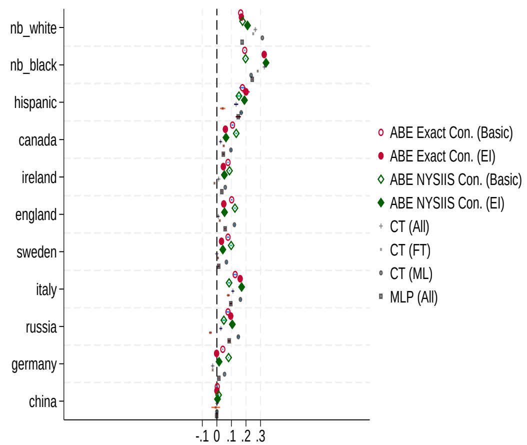
Appendix Figure 6: Representativeness relative to end year Census male population (1910), by subgroups



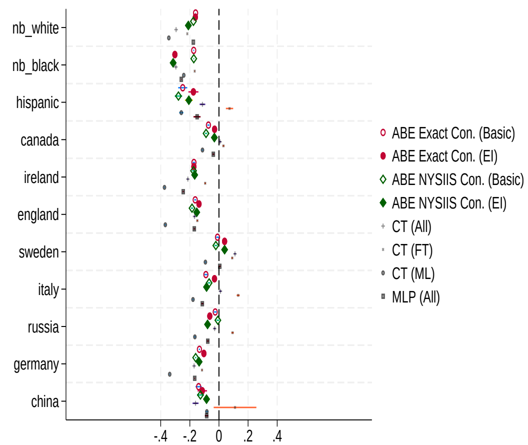
Panel G: Group-quarters



Panel I: Occscore



Panel H: Live with parents



Panel J: Age

Appendix Figure 6: Representativeness relative to end year Census male population (1910), by subgroups

Note: This set of figures presents a coefficient plot of the representativeness of 1900-1910 matches by subgroup generated by the ABE, Census Tree, and MLP algorithms compared to the 1900 full count Census. For each year, we take the full count Census and present the coefficient on a dummy variable representing presence in the matched sample for each characteristic of interest.

Table A1: Matching variables

Matching variable	Census Years	Used by ABE Basic	Used by ABE-EI	Persistence in CT Family Tree
First name (exact)	1850-1940	X	X	79.51
First name (NYSIIS)	1850-1940	X	X	82.11
Last name (exact)	1850-1940	X	X	72.90
Last name (NYSIIS)	1850-1940	X	X	86.75
Birthplace	1850-1940	X	X	95.14
Age	1850-1940	X	X	90.8
Middle Initial	1850-1940		X	84.59
Spouse present + first name	1850-1940		X	75.79
Father present + first name	1850-1940		X	77.63
Mother present + first name	1850-1940		X	74.63
County of residence*	1850-1940		X	76.11
Region of origin**	1850-1940		X	
Race	1850-1940			98.47
Immigration Year	1900-1930			53.9
State of residence	1850-1940			89.87
Parental nativity	1880-1940			96.05
Names of other relatives	1850-1940			
Names of neighbors	1850-1940			
Names of unrelated household members	1850-1940			
Age of first marriage	1930-1940*			
Duration of current marital status	1900-1910^			
Mother tongue	1910-1930^f			

Note: This table lists matching variables that can be used to link two census records and outlines both their availability in each census decade and persistence across adjacent decades. Persistence is calculated as the percent of individuals in the matched Census Tree Family Tree dataset for whom the variable is the same in both adjacent censuses (all matched years from 1900-1910 to 1930-1940 are pooled together). Note that year of immigration is unavailable for 1940, region of birth is excluded (since it is our synthetic variable), and variables that are likely unstable across time are not included.

f : foreign-born persons only.

^ : married individuals only.

* : ABE-EI uses county of residence as opposed to county distance, which measures the distance between counties of residence reported across different Census years.

** : ABE-EI uses region of origin only for matching Eastern European-born to 1910, which was missing Poland as an option for birthplace.

Table A2: Accordance rates with Family Tree hand-made links for matches made with and without the core 5 variables in the ABE-EI algorithm, 1900-1910

	Match also selected by ...									
	Basic linking variables		Parental Nativity		Race		State		Immig. Year	
	N	AR	N	AR	N	AR	N	AR	N	AR
At least one of the 5 ABE-EI variables select the match										
Exact Conservative	5,129,494	99.36	5,041,971	99.27	5,231,718	99.35	5,636,903	99.04	4,911,900	99.15
NYSIIS Conservative	5,036,695	99.43	5,580,068	99.25	5,325,668	99.41	6,065,837	99.01	5,026,828	99.05
Exact Standard	7,103,414	96.16	7,195,163	96.22	7,255,263	96.23	7,349,782	96.3	6,808,328	96.02
NYSIIS Standard	7,631,083	95.34	8,010,856	95.36	7,844,172	95.42	8,317,057	95.45	7,385,237	95.06
None of the ABE-EI core variables select the match										
Exact Conservative	1,248,753	68.62	1,200,597	70.3	1,251,307	69.69	834,352	62.2	1,106,886	65.51
NYSIIS Conservative	1,171,049	65.43	1,259,066	66.71	1,228,243	66.36	934,077	61.34	1,166,196	65.25
Exact Standard	2,206,445	46.7	2,116,482	48.29	2,179,138	48.23	1,454,392	44	2,012,817	47.77
NYSIIS Standard	2,562,611	46.29	2,529,075	47.93	2,550,601	47.69	1,795,980	42.99	2,343,208	47.21

Note: This table presents the match counts and accordance rates (with CT Family Tree) for matches made by the subset of Census variables not used by the main ABE-EI algorithm, conditional on whether one of the ABE-EI variables (used by the main algorithm) support this match. For the first panel, at least one ABE-EI variable agreed on the match; for the second, no ABE-EI variable participated in the match.

Table A3: Examples of potential matches rejected by ABE-EI

Individual	First Name	Last Name	Age	Birth-place	Middle initial	Mother's name	Father's name	County of residence	Mother foreign born	Father foreign born
Basic match, EI reject										
1900 source	Rupert	Fritz	20	MN		Mary	Andrew	MN 1450	yes	yes
1910 match	Rupert	Fritz	30	MN	J	Catherine		SD 990	yes	yes
Match made by basic + parental nativity only										
1900 source	John	Scholz	3	MN	J	Margaret	Joseph	MN 1450	no	yes
1910 match	John	Scholz	14	MN		Susanna	Allie	MN 530	no	yes

Note: This table shows two different 1900-1910 pairs: (1) a "mover" matched by ABE Basic but rejected by ABE-EI, and (2) a match made by a variable not used by the official ABE-EI algorithm - in this case, parental nativity - that is rejected by ABE-EI.

Table A4: ABE-EI matches divided by set of variables selecting the match, Exact Conservative 1900-1910

			Accordance rate with CT Family Tree Age in 1900	
Votes	Variables voting	% of sample	0-9	20+
1	Parent/Spouse	14.54	94.61	91.69
	County	25.28	89.35	95.38
	Middle Initial	4.86	74.26	80.73
2	Parent/Spouse + County	33.88	99.85	99.91
	Parent/Spouse + MI	3.97	99.72	99.71
	MI + County	6.18	99.29	99.72
3	All 3	10.74	99.98	99.98

Note: This table presents the accordance rates with the Census Tree Family Tree for ABE-EI (Exact Conservative, 1900-1910) broken down by rounds of voting variables (parent/spouse, county, and middle initial) and age group (0-9 and 20+). For those 0-9 years old, parent is considered a voting variable if either the mother or father's name votes; for those age 20+, spouse is considered instead of parent. The 'shares' columns represents the share of total matches in this sub-sample that were selected by that round of variables.

Table A5: Match rates for all census pairs

	1900-1910		1910-1920		1920-1930		1930-1940	
	N	Match rate	*N*	Match rate	*N*	Match rate	*N*	Match rate
ABE Exact Conservative								
Basic	6,357,902	16.4	8,397,981	17.64	10,624,900	20.23	12,563,664	20.23
Extra information	9,713,236	25.06	12,591,632	26.45	15,215,047	28.97	17,783,511	28.64
NYSIIS Conservative								
Basic	6,325,878	16.32	8,310,446	17.45	10,428,598	19.86	12,373,384	19.93
Extra information	12,097,709	31.21	15,312,887	32.16	18,494,345	35.21	22,014,773	35.46
No county	11,728,117	30.26	14,993,259	31.49	18,219,861	34.69	21,323,362	34.34
Exact Standard								
Basic	9,269,657	23.92	11,792,470	24.77	14,559,253	27.72	16,995,375	27.37
Extra information	10,185,582	26.28	12,865,519	27.02	15,515,569	29.54	18,315,441	29.5
No county	10,953,681	28.26	13,690,080	28.75	16,466,207	31.35	19,165,473	30.87
NYSIIS Standard								
Basic	10,159,055	26.21	12,910,364	27.12	15,940,169	30.35	18,804,068	30.29
Extra information	12,636,630	32.6	15,797,553	33.18	19,030,492	36.23	22,794,360	36.71
No county	13,220,429	34.11	16,406,157	34.46	19,753,949	37.61	23,315,322	37.55
Census Tree								
Family Tree	14,548,042	37.53	15,651,548	32.87	13,940,663	26.54	14,245,434	22.94
Machine Learning	14,075,305	36.31	17,878,518	37.55	20,649,807	39.32	27,675,758	44.57
All	24,976,021	64.44	30,164,025	63.35	35,888,058	68.33	42,665,479	68.72
MLP								
Round 1	14,764,543	38.09	18,158,485	38.14	22,265,243	42.39	26,154,302	42.12
All	17,097,065	44.11	20,894,327	43.88	25,238,517	48.05	29,528,077	47.56

Note: This table shows the number of links made between each pair of adjacent full count censuses (men only) from 1900 to 1940, as well as the match rate, which is the number of links as a share of all men in that year's census. ABE-EI uses first name, last name, place of birth (or region of birth for Eastern Europeans for 1900-1910 and 1910-1920), age, middle initial, mother's first name, father's first name, county of residence, and spouse's first name as matching variables.

TABLE A6: ABE Basic and ABE-EI accordance rates with CT Family Tree hand-made links, and additional comparisons against CT machine learning links and MLP (1900-1930)

		Accuracy proxy	Additional comparisons	
	N	CT Family Tree	CT ML	MLP
ABE Basic				
Exact Conservative	4,582,564	97.15	98.04	98.39
NYSIIS Conservative	4,553,464	97.31	99	98.52
Exact Standard	6,722,308	92.97	93.86	94.86
NYSIIS Standard	7,433,921	92.33	94.55	93.51
ABE Extra Information				
Exact Conservative	5,272,292	98.31	97.7	98.47
NYSIIS Conservative	6,373,319	97.83	97.55	97.84
Exact Standard	5,628,365	97.72	96.77	98.01
NYSIIS Standard	6,986,558	96.87	96.21	96.95

Note: This table presents the accordance rate of ABE Basic and ABE-EI links with the Census Trees's Family Tree links (CT Family Tree), the Census Tree's Machine Learning links (CT ML), and MLP links. The Census Tree links are obtained from censustree.org/data. The Family Tree links, generated by individual users who manually connect relatives to census records, are considered to be highly accurate. For each version of ABE, the number of matches made by that method is reported ('N'). The accordance rate is the percent of links made by both methods that are identical (i.e., conditional on both methods attempting to match the same individual from 1900 to 1910, the percent of the time that the two methods pick the same individual from 1910 as the match).

Table A7: ABE Basic and ABE-EI accordance rates with CT Family Tree hand-made links (all years)

	1900-1910	1910-1920	1920-1930	1930-1940
ABE Basic				
Exact Conservative	96.07	96.98	97.25	97.29
NYSIIS Conservative	96.36	97.11	97.22	97
Exact Standard	90.63	91.7	92.88	94.14
NYSIIS Standard	89.95	90.53	91.54	92.8
ABE Extra Information				
Exact Conservative	98.25	98.7	98.86	98.94
NYSIIS Conservative	98.09	98.52	98.66	98.63
Exact Standard	97.69	98.28	98.54	98.69
NYSIIS Standard	97.31	97.90	98.17	98.23

Note: This table shows the match counts and accordance rates for ABE Basic and ABE-EI with the Census Tree Family Tree for all adjacent decades from 1900-1910 to 1930-1940.

Table A8 (Panel A): Coefficients from regressions comparing matched samples to population in 1900, non-normalized

	White	Black	US-born	Married	Literate	Urban	Farm	Group quarters	Resides in state of birth	Parent present	Has children	Occscore	Age (decile)
Exact Conservative													
Basic	0.0735	-0.0665	0.0741	-0.0328	0.0616	-0.0324	0.0454	-0.0209	0.0416	0.1138	-0.0188	0.1992	-3.9679
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0064)	(0.0082)
Extra Information	0.0738	-0.0651	0.0741	0.0060	0.0673	-0.0228	0.0431	-0.0295	0.0624	0.1175	0.0164	0.5341	-4.0249
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0054)	(0.0070)
NYSIIS Conservative													
Basic	0.0642	-0.0574	0.0673	-0.0407	0.0505	-0.0475	0.0555	-0.0202	0.0438	0.1175	-0.0261	-0.0993	-4.2105
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0065)	(0.0082)
Extra Information	0.0790	-0.0697	0.0675	0.0020	0.0666	-0.0300	0.0526	-0.0321	0.0699	0.1322	0.0141	0.3597	-4.6817
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0051)	(0.0065)
Exact Standard													
Basic	0.0629	-0.0554	0.0715	-0.0310	0.0569	-0.0271	0.0390	-0.0192	0.0424	0.1060	-0.0182	0.1837	-3.9504
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0055)	(0.0071)
Extra Information	0.0729	-0.0640	0.0734	-0.0022	0.0672	-0.0148	0.0390	-0.0287	0.0670	0.1211	0.0093	0.5431	-4.2548
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0054)	(0.0069)
NYSIIS Standard													
Basic	0.0576	-0.0508	0.0597	-0.0412	0.0471	-0.0358	0.0450	-0.0185	0.0461	0.1122	-0.0275	-0.0520	-4.3405
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0053)	(0.0069)
Extra Information	0.0766	-0.0674	0.0632	-0.0037	0.0657	-0.0159	0.0433	-0.0310	0.0767	0.1316	0.0092	0.4261	-4.7831
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0050)	(0.0065)

Table A8 (Panel B): Coefficients from regressions comparing matched samples to population in 1900, non-normalized

	White	Black	US-born	Married	Literate	Urban	Farm	Group quarters	Resides in state of birth	Parent present	Has children	Occscore	Age (decile)
Census Tree													
Family Tree	0.1046	-0.0939	0.1108	0.0625	0.0643	-0.1302	0.1403	-0.0458	0.0601	0.1606	0.1206	-0.7194	-5.0746
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0048)	(0.0063)
Machine Learning	0.0754	-0.0663	0.0596	-0.0028	0.0645	-0.0220	0.0473	-0.0297	0.0681	0.1091	0.0130	0.2063	-4.0190
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0048)	(0.0063)
All	0.1097	-0.0941	0.0827	0.0295	0.0751	-0.0545	0.0893	-0.0509	0.0865	0.1689	0.0642	-0.0145	-6.7914
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0046)	(0.0063)
MLP													
Round 1	0.0944	-0.0840	0.0724	0.0138	0.0699	-0.0364	0.0644	-0.0401	0.0924	0.1524	0.0386	0.2590	-5.5522
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0048)	(0.0062)
All	0.0987	-0.0872	0.0696	0.0393	0.0673	-0.0433	0.0747	-0.0452	0.0890	0.1492	0.0827	0.2503	-5.7425
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0047)	(0.0061)
Geographic Alts.													
State, no county	0.0714	-0.0631	0.0661	-0.0070	0.0648	-0.0216	0.0346	-0.0222	0.0331	0.0977	0.0034	0.5126	-3.9631
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0053)	(0.0069)
No county	0.0828	-0.0744	0.0896	0.0207	0.0746	-0.0206	0.0325	-0.0313	0.0568	0.1447	0.0219	1.1498	-5.4287
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0063)	(0.0078)

Note: The balance table for year 1900 (Panels A and B) presents the non-normalized point estimates and standard errors for a variety of characteristics and matching methods. For each method and variable, the entry represents the coefficient and standard error for a dummy variable for presence in the linked sample, compared to the full count Census of that year. The same statistics are represented visually in Figure 3. For the geographic alternatives, versions of ABE-EI were created that: a) include state instead of county as a matching variable, and b) include neither state nor county. In both cases, the results refer to the exact conservative version of the algorithms.

Table A9: Share of ABE-EI links selected by each matching variable, by mover/stayer status and age, 1900-1910

	0-9		10-19		20+	
	Move	Stay	Move	Stay	Move	Stay
Matching variable						
Middle initial	28.36	21.19	60.16	24.45	50.88	26.03
County	0	76.60	0	82.56	0	85.26
Mother's name	67.23	59.39	37.8	37.64	2.75	4.69
Father's name	66.87	60.19	34.01	34.79	1.7	3.1
Spouse's name	0	0	.44	.24	61.6	46.59

Note: This table presents the percent of the time that each ABE extra information variable 'votes' on a match, broken down by age group and by whether their state of residence in 1910 is the same as their state of residence in 1900.

Table A10: Examples of 'mover' matches made or rejected by ABE-EI

Individual	First Name	Last Name	Age	Birth-place	Middle initial	Spouse's name	Mother's name	Father's name	County of residence
ABE-EI match									
1900 source	George	Joynson	10	MN	J		Emma	Joseph	MN 1050
1910 match	George	Joynson	20	MN	J		Emma	Joseph	OK 710
ABE-EI rejection									
1900 source	Frank	Hanley	5	MN			Frances	Henry	IA 650
1910 match	Frank	Hanley	15	MN			Harriet	James	WA 570

Note: This table shows two different 1900-1910 pairs: (1) a 'mover' matched by ABE-EI (the state of residence changes between the two Census years), and (2) a 'mover' rejected by ABE-EI.

Table A11: ABE-EI accordance rates with Family Tree hand-made links by mover/stayer status and age, 1900-1910

	0-9				10-19				20+			
	Move		Stay		Move		Stay		Move		Stay	
	N	AR	N	AR	N	AR	N	AR	N	AR	N	AR
ABE Basic												
Exact Conservative	227,057	93.54	1,695,269	96.87	235,328	89.48	1,128,335	97.40	549,304	78.47	2,522,609	97.94
NYSIIS Conservative	237,744	93.88	1,725,773	97.23	243,929	90.08	1,111,784	97.72	561,445	79	2,445,203	98.09
Exact Standard	324,630	86.55	2,347,439	93.05	388,807	77.04	1,607,325	93.61	1,101,596	57.47	3,499,860	94.68
NYSIIS Standard	375,537	85.48	2,616,266	92.81	454,754	75.31	1,739,010	93.27	1,287,000	54.67	3,686,488	94.22
ABE-EI												
Exact Conservative	262,228	97.73	2,728,849	98.35	145,626	95.37	1,754,981	98.44	413,907	91.36	4,407,645	98.82
NYSIIS Conservative	338,515	97.45	3,457,817	98.36	183,554	94.54	2,175,243	98.33	531,192	89.73	5,411,388	98.66
Exact Standard	277,991	97.2	2,837,275	97.97	162,707	94.26	1,865,798	97.95	478,442	88.29	4,563,369	98.39
NYSIIS Standard	350,562	96.67	3,550,108	97.87	201,967	92.78	2,301,513	97.63	617,449	85.48	5,615,031	98.04

Note: This table shows the match counts and accordance rates for ABE Basic and ABE-EI, decomposed by age group and whether an individual's state of residence is the same or different from 1900 to 1910.

Table A12: ABE-EI geographic variants comparisons on match rate, accordance and representativeness, 1900-1910

	N	Match Rate	Acc. Rate	Coeff. on =1 if matched (relative to pop) (DV = lives in state of birth)
With county	9,713,236	25.06	98.25	0.0624 (0.0002)
Without county	7,225,573	18.64	98.35	0.0568 (0.0002)
With state, no county	10,408,650	26.85	95.37	0.0331 (0.0002)

Note: This table presents the match counts, match rates, and accordance rates with the CT Family Tree for modified versions of ABE-EI where either 'all' variables (including the unused Census variables) vote or the 'immutable' variables (ABE Basic matching variables, middle initial, parental nativity, race, year of immigration) vote. In both cases, unanimity is enforced during the voting. Mover representativeness is taken from the representativeness table (namely, share living in their state of birth), where the magnitude above (below) zero represents how much the group is over-represented (under-represented). Standard errors are in parentheses.

TABLE A13: ABE Basic and ABE-EI match rate for subgroups, and Accuracy rates with CT Family Tree hand-made links across censuses

	Full Sample									Irish Born								
	1860-70			1900-10			1930-40			1860-70			1900-10			1930-40		
	N	Match rate	Acc. Rate	*N*	Match rate	Acc. Rate	*N*	Match rate	Acc. Rate	*N*	Match rate	Acc. Rate	*N*	Match rate	Acc. Rate	*N*	Match rate	Acc. Rate
ABE Basic																		
NYSIIS Conservative	1,857,605	13.18	94.36	6,325,878	16.32	96.36	12,373,384	19.93	97	34,178	4.14	77.75	36,442	4.85	91.39	27,630	7.91	94.81
NYSIIS Standard	3,144,162	22.31	86.40	10,159,055	26.21	89.95	18,804,068	30.29	92.8	89,926	10.88	59.69	93,148	12.39	72.23	59,852	17.13	82.38
ABE Extra Information																		
NYSIIS Conservative	3,493,949	24.79	96.71	12,097,709	31.21	98.09	22,014,773	35.46	98.63	112,007	13.56	89.04	137,851	18.34	95.44	81,175	23.23	97.13
NYSIIS Standard	3,705,629	26.3	95.76	12,636,630	32.6	97.31	22,794,360	36.71	98.23	144,238	17.46	82.91	165,283	21.99	90.67	92,033	26.34	94.79

Note: This table presents the match rate of Basic and ABE-EI links on the Irish population, as well as their accuracy rate (Acc. rate) with the Census Trees's Family Tree links across time. The Census Tree links are obtained from censustree.org/data. The Family Tree links, generated by individual users who manually connect relatives to census records, are considered to be highly accurate. For each version of ABE, the number of matches made by that method is reported ('N'). The accordance rate is the percent of links made by both methods that are identical (i.e., conditional on both methods attempting to match the same individual from 1860 to 1870 (1900 to 1910, 1930 to 1940), the percent of the time that the two methods pick the same individual from 1870 (or 1910, 1940, respectively) as the match).