The Creation of LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database Project

By Martha Bailey[1,2], Peter Z. Lin[1], A.R. Shariq Mohammed[3], Paul Mohnen[4], Jared Murray[5], Mengying Zhang[1], Alexa Prettyman[1]

[1] University of California-Los Angeles
[2] National Bureau of Economic Research
[3] Northeastern University
[4] University of Pennsylvania
[5] University of Texas, Austin

July 20, 2022

Abstract

This paper describes the creation of the Longitudinal, Intergenerational Family Electronic Micro-Database (LIFE-M), a new data resource linking millions of individuals and families living in the late 19th and 20th centuries and combining vital records and decennial censuses. This combination of records provides a life-course and intergenerational perspective on the evolution of health and economic outcomes. Vital records also enable the linkage of women, because they contain a crosswalk between women's birth (i.e., "maiden") and married names. Our overview of the methodology includes (1) a discussion of the data structure and linking process, (2) the coverage of vital records, (3) the characteristics of the data and resulting linked samples, including linking rates, representativeness, and weights; and (4) potential uses for the data.

Contact Information

Bailey: Department of Economics, University of California, Los Angeles, 315 Pertola Plaza, 90095. Email: marthabailey@ucla.edu. Website: https://sites.google.com/g.ucla.edu/marthajbailey

Acknowledgements

This project was generously supported by the National Science Foundation (SMA 1539228), the National Institute on Aging (R21 AG05691201), the University of Michigan Population Studies Center Small Grants (R24 HD041028), the Michigan Center for the Demography of Aging (MiCDA, P30 AG012846-21), the University of Michigan Associate Professor Fund, and the Michigan Institute on Research and Teaching in Economics (MITRE). We gratefully acknowledge the use the Population Studies Center's services and facilities at the University of Michigan (R24HD041028). The study team gratefully acknowledges the use of the services and facilities of the Population Studies Center at the UM (P2CHD041028) and the California Center for Population Research at UCLA (P2CHD041022). We are grateful to Dora Costa, Shari Eli, Adriana Lleras-Muney, Joseph Price, and the board members of the LIFE-M project, including Eytan Adar, George Alter, Hoyt Bleakley, Matias Cattaneo, William Collins, Katie Genadek, Maggie Levenstein, Bhash Mazumder, Evan Roberts, and Steven Ruggles for their helpful suggestions. We are also grateful to Garrett Anstreicher, Sarah Anderson, Meizi Li, Morgan Henderson, Alfia Karimova, Catherine Massey, and Annie Wentz for their excellent contributions to the LIFE-M project and assistance with this analysis.

I. Introduction

Is the U.S. still the "land of opportunity"? Is it still the "great melting pot"? How have investments in public health efforts affected wellbeing and longevity? How did the 20th century transformation of laws and institutions affect inequality in income, health, and welfare? How have changes in women's roles affected the structure of families and children's opportunities?

The Longitudinal Intergenerational Family Electronic Micro-dataset (LIFE-M) leverages the new resource of digitized *vital statistics* data to follow millions of individuals from birth to death, integrate health, demographic, and family outcomes with census data, and improve the linking of historical records. Vital statistics records include natality, marriage, and mortality certificates, which allow LIFE-M to add demographic and health information to socio-economic variables in the Census. Additional features of the Vital Statistics are that they (1) include many individuals who were not recorded in decennial censuses, because they died or emigrated and (2) include additional information to help link groups underrepresented in the Census. For instance, birth and marriage records contain women's birth ("maiden") *and* married names, allowing LIFE-M to follow women who change their surname at marriage from cradle to grave. Vital statistics also contain middle names, exact dates of birth, and multiple family members which increases the likelihood of linking individuals with common or misspelled names. Combining information from decennial censuses and vital statistics allows LIFE-M to include individual socio-economic, demographic, and health information across the life course for networks of multi-generational families.

This paper provides a detailed overview of the creation of the LIFE-M data for Ohio and North Carolina. We discuss (1) the data structure and linking process, including the creation of high-quality, hand-linked samples and the machine-learning algorithm; (2) the coverage of vital records; (3) the characteristics of the linked samples, including link rates, representativeness, and weights; and (4) potential uses for the data, including how to combine the LIFE-M data with other datasets, such as the 1880-1940 Decennial Censuses and geocoded data. We conclude with a discussion of how the LIFE-M data can be used to expand the research frontier in social science, population, and health.

II. Existing Data Resources and the Contribution of LIFE-M

Several types of available data permit analyses of twentieth century population dynamics, including linked historical data, linked contemporary data, and panel surveys. One example of linked historical data is the Early Indicators Project. Led by Dora Costa, these data provide an important longitudinal perspective on health and economic outcomes during the middle 19th century (Wimmer 2003). The data consist of 39,340 Union Army (UA) soldiers, approximately 6,200 of whom were "Colored Troops." These data measure the date of death and provide rich information on disability, health, use of medical care, and pension receipt for men reaching retirement age in the late 19th century. Through links to the 1850 and 1930 censuses, the UA data also include socio-demographic and economic variables. An important limitation of the UA data is that they consist of men only, most of whom were Northern born.

The Minnesota Population Center (MPC) has also led efforts to digitize and integrate historical census data. Thus far, they have created Integrated Public Use Microdata Series Linked Representative Samples (LRS) of the 1880 Census linked to the 1850-1930 Census one-percent samples (Ruggles et al. 2010).[1] These linked data record economic (e.g., occupation, literacy, labor-force participation, home ownership) and demographic (e.g., age, birthplace, race, marital status, number of children) outcomes for around 500,000 people. Though large in scale, important limitations of these data include that many women cannot be linked by name (because their last names change at marriage), longitudinal coverage is sparse (at most two points in time for any single person), and intergenerational coverage consists of at most two generations (primarily father-son pairs).

More recently, MPC released the Multigenerational Longitudinal Panel (MLP), which uses *supervised* learning to link millions of individuals between every pair of adjacent censuses from 1900 to 1940 (Helgertz et al. 2020). The resulting sample sizes range from around 18 million individuals linked

[1] This large, linked sample follows two earlier linking projects. Pullum and Guest created a national linked sample of two cohorts of men born in the 1880 and 1900 censuses (Guest 1987, N=4,014, linkage rate 39.4%). Ferrie (1996) linked a nationally representative sample of the 1850 census to the 1860 census (N=4,938, linkage rate 19.3%).

between the 1910-1920 and 1940 censuses and 26 million individuals linked between the 1930 and 1940 censuses. MLP's linking strategy is implemented in two steps. First, men are linked between adjacent censuses as individuals. In this step, MLP exploits rich training data and contextual information in the linking process (e.g., place of residence, co-resident individuals), in addition to names and basic demographics. This strategy increases match rates while reducing the likelihood of false matches, but the final linked sample overrepresents men who do not move with stable household members. In the second step, the procedure tries to link members living in the same household as men linked in the first step. This second step helps link women who are living with their spouses or daughters living with their fathers. Census data limitations, however, make it nearly impossible to link women who change households or their names at marriage.

Concurrent to the development of MLP, Abramitzky, Boustan, and Rashid (2020) released census links under the Census Linking Project (CLP), which also links millions of men between every pair of censuses from 1850 to 1940 (1890 Census is omitted as the microdata have been lost). Building on the linking approaches in Abramitzky, Boustan, and Eriksson (2012), Feigenbaum (2016), and Abramitzky, Mill, and Pérez (2018), CLP links data using *unsupervised* machine-linking based on name, race and time/place of birth information in the census. (Note that the difference between supervised and unsupervised machine-learning methods is that the latter do not use training data to control error rates or optimize performance). Due to census data limitations, CLP does not attempt to link women.

The CenSoc project at the University of California, Berkeley, also uses the Abramitzky, Boustan, and Eriksson (2012) unsupervised learning approach to link the 1940 Census to the Social Security Death Master File and the NUMIDENT file (Goldstein and Breen 2020). This project creates 15 million deaths that occurred between 1975 and 2005 for people over the age of 65 in order to analyze disparities in mortality across race, place, and nationality. Similar to Bailey, Mohammed, and Mohnen (2021), this project links women by following women's name changes in the Social Security records.

In addition to these large-scale data projects, smaller surveys offer a second type of longitudinal, intergenerational data. These include the Panel Survey of Income Dynamics (PSID), National Longitudinal

Surveys (NLS), and the Health and Retirement Study (HRS). The PSID and NLS longitudinal surveys began in the late 1960s and contain rich information on economic status, health, and well-being. The initial PSID sample consisted of 5,000 families and has grown to over 7,000 families as the study also follows descendants. The original NLS cohorts, which cover the periods 1966-1981 for young men, 1966-1990 for older men, 1967-2003 for mature women, and 1968-2003 for younger women, also began with initial sample sizes of around 5,000.[2] The HRS is a longitudinal survey that has followed Americans over age 50 from 1992 to the present. After beginning with the 1931-1941 birth cohorts (N=12,000), the HRS added 1924-1930 and 1942-1947 birth cohorts in 1998 (N=14,000). HRS data include health, disability, wealth, retirement, and financial literacy questions in addition to retrospective measures of early life and adult economic and outcomes. Note that the oldest individuals (born in 1924) need to survive to age 74 to make it into the survey, so HRS samples miss many individuals in these early cohorts dying at younger ages.

These surveys cover cohorts reaching adulthood about 100 years after the UA veterans—individuals born or reaching adulthood in the second half of the 20th century. This leaves a gap of roughly one century between the longitudinal coverage of the UA data and recent surveys. Moreover, these data have some common limitations: significant attrition (52.2% of the original PSID sample remained by 1989; around 5% of the NLS samples left per year; 15% attrition in the HRS as of 2004) and lack of geographic coverage, which constrain the representativeness of these data at the state and more local (county/town) levels.[3]

---

[2] The NLS has subsequently tracked supplemental samples. One covers ages 14-22 in 1979 (N=12,686) (and children for women in this survey) and another ages 12-16 in 1996 (N=9,000).

[3] A variety of independent administrative and restricted data sources offer a third type of longitudinal, intergenerational data. The National Longitudinal Mortality Study (NLMS) links the Current Population Surveys and other records to death certificates to examine the relationship of demographic and socio-economic characteristics with mortality rates. These large microdata samples (N> 340,000 deaths) generally link individuals 50 or older to demographic and socio-economic information in the CPS from about age 40. Researchers have also conducted labor-intensive hand-linkages across censuses (Ferrie 1996, Guest 1987, Long and Ferrie 2013, Collins and Wanamaker 2014, Bleakley and Ferrie 2017, Bleakley and Ferrie 2016, Bleakley and Ferrie 2013) or hand-linked samples of U.S. census data to other records, including ship manifests (Abramitzky, Boustan, and Eriksson 2014) and census records from other countries (Biavaschi and Elsner 2013). Many of these linked samples are the property of the researchers who collected or linked them and are not available for public use. Lack of access to these data and substantial barriers to creating such samples limit replication, new research using these data, and analyses of data quality.

In summary, available data limits the longitudinal and intergenerational analyses of 20th century Americans in several ways. Historical data have tended to focus on men, either soldiers (UA) or because they can be linked across historical sources (LRS, MLP, CLP). In addition, historical data have tended to include a handful of variables, either from one census and death records (CenSoc) or census pairs (LRS, MLP, CLP). More recent longitudinal surveys tend to have small samples (especially for minority populations) and limited temporal coverage. LIFE-M addresses these limitations by combining many of these features into a single database.

(1) Intergenerational linkages of families across four generations, including cohorts born from the 1860s to the 1940s

LIFE-M cohorts include the veterans of two World Wars and the children of the Great Depression. This group largely defines aging Americans in the last half of the 20th century. Adding to modern data that trace small numbers of families over time (PSID, NLSY, HRS) and historical data that trace men (and some women) (UA data or linked Census samples), LIFE-M permits the large-sample intergenerational analyses of entire families. Linked samples include a large set of mothers, daughters, and sisters, as they transition from their birth to marriage families.

(2) Large sample sizes of understudied populations, including women and underrepresented minorities:

LIFE-M follows large samples of women longitudinally and across generations and improves linkage rates for underrepresented men. Typically, historical linking projects link individuals based on surnames but cannot link women after marriage when they change their surnames. For instance, IPUMS MLP links only a handful of women who change their names at marriage, typically linking women when they remain in the same household after marriage. Vital birth and marriage records contain a crosswalk of information between women's birth and married surnames that facilitate the inclusion of women and girls. For example, 95% of birth certificates from Ohio (1,664,928/1,757,009) have a birth and married name.

Vital records also increase LIFE-M's sample sizes for other understudied groups, such as racial minorities and immigrants. Vital records contain high quality data on names, dates of birth, and birthplace

as well as parents' names and birth county in many instances.[4] In addition to containing more information, vital records are also collected with more accuracy than census records. Because the goal of the Census is to *count* individuals, enumerators may not enter full legal names, collect accurate information on age, or transcribe the correct birth state (Bailey, Cole, and Massey 2020).[5] The census goal of counting the population is met even if nicknames, partial names, or rounded ages are used. In contrast, the goal of vital records is to record the full legal name of the individual, date of the vital event (birth, marriage, or death), and place of the vital event—typically recorded by the individual themselves or a close family member or friend. Consequently, vital records are more likely to have full legal names (including middle name), exact date of birth, and a complete record of the place of the event. For cohorts born between 1900 and 1930, over 3.2 million birth records (64%) out of 5 million have middle names and 4.4 million (87%) have exact date of birth.

This additional, high-quality information crucially allows LIFE-M to link more individuals with shorter names, more common names, and misspelled names—characteristics associated with less education and being a racial/ethnic minority or immigrant. It also helps determine the correct match when many are possible—an especially large problem in contexts in historical data with limited information.[6] LIFE-M links more than 170,000 Black Americans and more than 368,000 foreign-born people.

(3) Integrates microdata combining socio-economic, demographic, and health information

LIFE-M integrates health, economic, family, and demographic data into a large-scale, longitudinal and intergenerational resource using a 5-digit alphanumeric identifier (lifemid) contained in the LIFE-M dataset. To facilitate linking LIFE-M to public census data, the LIFE-M identifier is posted on the IPUMS

[4] Appendix Table D.1 compares LIFE-M linking variables to Census linking variables for the interested reader.
[5] Age misreporting in the Census (e.g., there are a lot more 50- and 60-year-olds relative to 51- and 63-year-olds). Age misreporting is a particular problem among Black Americans (Elo and Preston 1994, Logan and Parmen 2011). Another example is lying about one's age to marry—a common practice through the 1970s to circumvent minimum age at marriage requirements (Blank, Charles, and Sallee 2009).
[6] Multiple matches have been so problematic that past work has eliminated common names entirely from samples to be linked (Ferrie 1996, Ruggles 2006).

MLP website (https://usa.ipums.org/usa/mlp/mlp.shtml), and census record ids (histid) are also contained within the LIFE-M data. Variables contained within the LIFE-M data include:

- birth family characteristics (e.g., birth order, sibling sex composition, age differences, twinning, number of siblings);
- parental and grandparent characteristics (e.g., age, race, occupation, literacy and education,[7] and birth state or country from the censuses);
- own economic and demographic outcomes (wages, employment, occupation, birth state or country, literacy and education[7]);
- health (date and place of death, longevity in days);
- marriage family characteristics (e.g., age at marriage, married name, spouse name and all characteristics above);
- own births (number of children, mortality of own infants and children, timing of births, sex composition, and twinning); and
- geographic location (town or address) at vital events and Census enumeration (lifetime mobility).

(4) Geographic detail allows for integration with other data and study of life-time migration

Information on geographic location across time presents a multitude of opportunities for researchers to combine LIFE-M with other data, including the availability of policies or programs (e.g., compulsory schooling laws, Great Depression era programs and policies; public health measures); environmental information (e.g., air, water, or soil pollution); or contextual or neighborhood factors (e.g., school quality, crime, public health). In addition, LIFE-M enables the study of location changes over time.

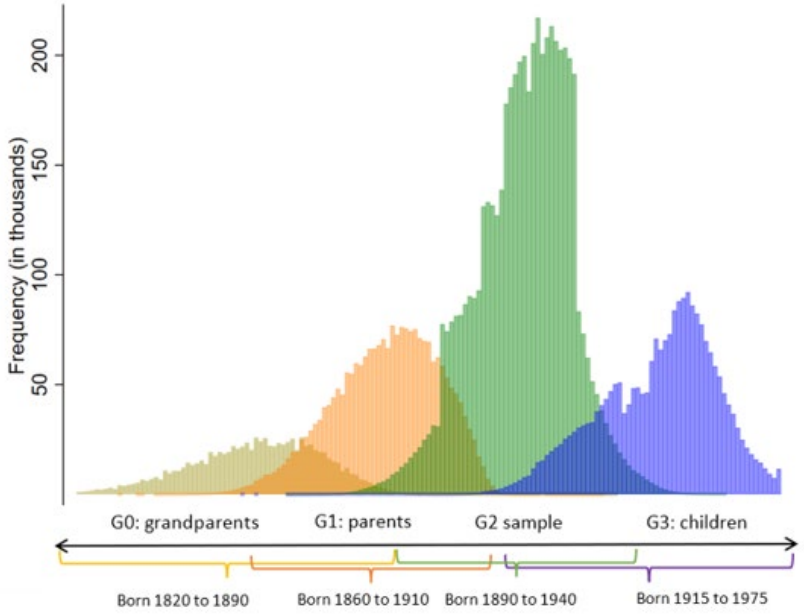
Detailed geography is stored in the LIFE-M location file, which contains individuals' county and state of residence at birth; when observed in the 1880, 1900, 1910, 1920, 1930, 1940 Censuses; at the time of marriage; and at death. Of the 543,965 individuals for whom we have non-missing state of birth, first marriage, and death, 45,164 (8.3%) have moved across state lines at some point over their lifetime. Among the 234,670 who have not moved across state lines with non-missing county information, 138,383 (59%) have lived in different counties over their lifetime.

[7] Completed education is first available in the 1940 Census; literacy is available in censuses prior to 1940.

III. LIFE-M's Generational Structure and Record Combination

The LIFE-M dataset was designed to represent the lives of individuals born in the first 30 years of the 20th century. As show in Figure 1, these cohorts and their siblings constitute "Generation 2" (or G2) in the larger LIFE-M generational structure. G1 are the parents of G2 and are born roughly from 1860 to 1910. G0 are the grandparents of G2 and are born from about 1820 to 1890. G3 are the children of G2 and were born from about 1915 to 1975.

Figure 1. LIFE-M Generation Structure



Notes: The project starts with G2 birth certificates to construct the intergenerational and longitudinal data. The figure shows the distribution of birth years for each generation. Generations have overlapping birth years so birth cohorts are approximations.

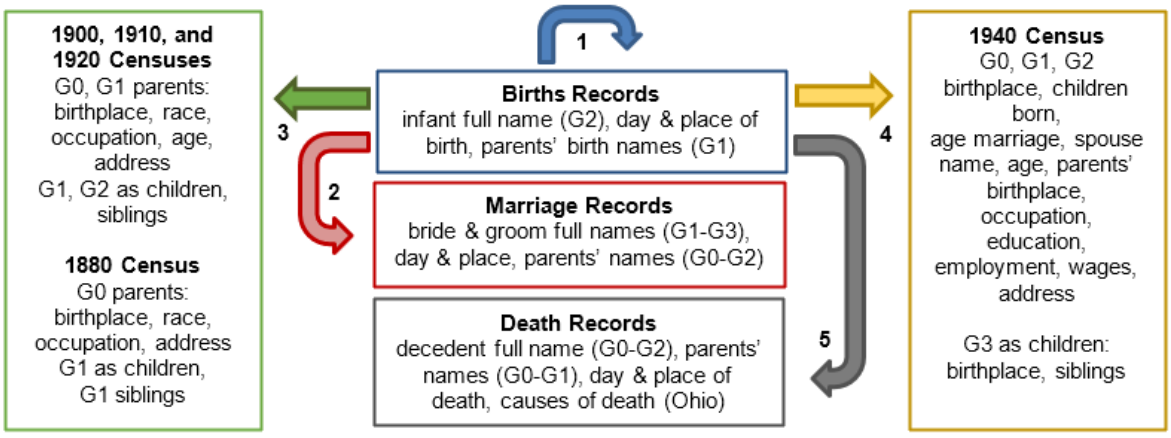
Newly digitized vital records and linking methods facilitated the development of this multigenerational, longitudinal database. On the data front, FamilySearch.org, a nonprofit genealogical website, has digitized millions of handwritten records and made them publicly available. At the time of this writing, these microdata include over 60 million births for 19 states, 70 million deaths for 41 states, and 55 million marriages for 47 states. Examining multiple sets of state records, we determined that only a handful of states had largely complete vital records for our period of interest and decided to link Ohio and North Carolina. On the linking front, rapid progress applying machine learning to historical data at the Census

Bureau and MPC as well as a partnership with academic innovators allowed the LIFE-M project to use cutting-edge methods to scale historical linking while maintaining accuracy.

Figure 2 provides an overview of LIFE-M's linking process. The first step of the process was to reconstitute birth and marriage families for the late 19th and early 20th century birth cohorts (G2) (arrow 1), using both their parents' full names (G1) and other information such as their parents' birthplaces. A second step linked G2 to their own children (G3). This step allows the reconstruction of two to three generations of interrelated families.

Next, marriage records were linked using the bride's and groom's names, exact dates of birth, and places of birth (when available) (arrow 2). Although 90% of women born in this period were married at some point (Bailey et al. 2014), marriage registration was highly incomplete which implies much lower match rates (Kennedy and Ruggles 2014). This step was also completed for G1 and G3.

Figure 2. Records Combined by LIFE-M



The third step links G2 to their grandparents (G0) using the 1880, 1900, 1910, and 1920 Censuses (arrow 3). Parents' (birth or married) names (G1) are linked to the 1900 census names, which provides key information on birthplace, age, and race. We link G1 to the 1880 census using their names only or names in addition to ages, birthplace, and race (obtained from the census links). This step connects G2 to G0 and is important because it allows for the addition of G1's early life family conditions, including G0

ancestry/heritage, economic circumstances such as occupation, race, and address. The fourth step links four generations (G0, G1, G2, G3) to the full-count 1940 census (arrow 4). This step uses full names (including birth and birth names of women), exact birth dates/age, and birthplace. The 1940 Census is the first census to include rich information on educational attainment, wages and salary, and many employment outcomes. This is only possible for some of G0 (many will have passed away before 1940), but most of G1, G2 as adults (in their marriage families), and G3 as children (in birth families).

The final step (arrow 5) linked all individuals (G0-G3) to death records, using full birth and/or married names, exact day of birth, and parents' names and place of birth when available in the collection. Almost all of G0-G2 had died in period covered by the death records, as did many (but not all) of G3.

IV. Supervised Machine-Learning for Linking Historical Data at Scale

Supervised machine-learning approaches are the gold standard for large-scale data linking, but machine-generated links are only as good as the data used to discipline these models. Consequently, the lack of ground truth for historical data has been a central challenge to implementing supervised methods. To remedy this deficit, the LIFE-M project created highly vetted hand-linked samples. While hand-linked data are not "ground truth" in the purest sense of the term, they were created to mimic this standard as closely as possible. This section describes the creation of hand-linked samples and how we trained machine models to link historical data at scale with low error rates.

A. Creation of Hand-Linked "Training" Data

To create the hand-linked data, we recruited students to hand-code links for different types of linking. We ensured data quality by developing a comprehensive semi-automated process.

Training Data Trainers

Before hand-coding data, prospective data trainers participated in a rigorous, multi-day orientation where they learned about the original records in script format, the process through which they were digitized, and idiosyncrasies of the records (e.g., age heaping in the census, common digitization errors of script, age misrepresentation on marriage records). Prospective trainers were mentored by more

experienced reviewers (either a long-term member of the LIFE-M project team or a more experienced data trainer). Orientation entailed over 30 hours of supervised practice reviewing and discussing how to link historical records, and trainers received high frequency feedback. To pass the final test and join the LIFE-M team, prospective trainers were required to reach a 0.95 correlation with a highly vetted and carefully chosen truth dataset, which was designed to test the trainers' knowledge and decision-making in a variety of linking contexts.

Distribution of Records to Data Trainers

An automated system managed the distribution of data to trainers. The automated system allowed trainers to log in from any computer, and the system automatically uploaded trained data in complete form to a central repository after completion. This system minimized concerns with data loss and transfers as well as the need for a team manager to distribute data around the clock for trainers on different class and work schedules. The system also collected metadata, including who trained the data and the time of start and completion for each batch. The system also randomized the distribution of data, which were delivered in small enough increments that trainers could finish within 15 to 20 minutes.

Displaying Sets of Candidates

LIFE-M displayed *multiple candidate record links* as shown in Table 1. Several factors motivated this decision. First, displaying multiple candidate links allows trainers to see the empirical frequency of name combinations and age combinations, which informs their certainty about a particular link. For instance, "Jason O'Sullivan, born in Ohio, age 35" may be a perfect match to another record, "Jason O'Sullivan, born in Ohio, age 35." However, as shown in Table 1, the number of close matches within the top-30 set that did not necessarily match exactly on name or age (±band, where band is typically 1-3 years) does not instill confidence that the exact, unique match is correct.[8] When trainers see three additional close

[8] This is due to name misspellings, incomplete names (e.g., nicknames, initials), transposed first and middle names, and other idiosyncrasies in historical records. The recording of age in the census tends to reflect "age heaping," the common practice of rounding ages to the nearest multiple of five (A'Hearn, Baten, and Crayen 2009, Hacker 2013).

links (especially, “Jason O’Sullivon, born in Ohio, age 35”), they tend to use this set-level information to reject the top candidate link. In this case, showing a set of candidate matches would tend to *increase* trainer uncertainty and reduce confidence in a link relative to showing only one candidate match at a time. Notably, even conservative versions of unsupervised methods such as the exact version of Abramitzky, Boustan, and Eriksson (2012) do not take account of the rich set-level features which allows trainers to see the uncertainty in links.

Table 1. Example of Sets Displayed to Trainer for Birth Certificate to Census Linking

A. Information Sets Decrease Trainer Certainty about Matches

	Record	Name	Age	Birthplace
4	Candidate	Jason O’Sullvan	33	Ohio
3	Candidate	Jason O’Sullivon	32	Ohio
2	Candidate	Jason O. Sullivan	33	Ohio
1	Candidate	Jason O’Sullivan	35	Ohio
	Primary	Jason O’Sullivan	35	Ohio

B. Information Sets Increase Trainer Certainty about a Match

	Record	Name	Age	Birthplace
7	Candidate	Susan H. Vielah	31	North Carolina
6	Candidate	Shelly H. Olive	35	North Carolina
5	Candidate	Sheilagh H. Oglvie	31	North Carolina
4	Candidate	Sally H. Givie	31	North Carolina
3	Candidate	Shelly O’Neill	33	North Carolina
2	Candidate	Sylvie H. Ogilbie	29	North Carolina
1	Candidate	Cecela F. Ogilvie	34	North Carolina
	Primary	Shelagh Harris Ogilvie	31	North Carolina

Another reason for displaying sets of information is that trainers can make links that less flexible methods might miss. For instance, “Shelagh Harris Ogilvie, born North Carolina, age 31” may have no matches based on exact name or age (±band), but the empirical frequency of similar candidates could make a less good match very strong. Candidate 5, “Sheilagh H. Oglvie, age 31,” shows a different spelling of the first and last name but the exact age. In addition, trainers learn from the set that there are no other Shelaghs + Olgivies with similar first names and within a similar age range. (Candidate #1 might be Shelagh’s

younger sister). Given that Shelagh/Sheilagh and Ogilvie/Oglvie may easily be transcription errors, the trainer feels very confident and chooses candidate #5. In this second case, displaying a set of candidate matches tends to *reduce* uncertainty about a particular match that differs in several dimensions.

In short, when there are several close matches (even if none are exact matches), methods using pairwise comparisons ignore information that can increase (Table 1A) or decrease uncertainty (Table 1B), whereas methods considering the set of possible candidates take it into account. The true link is not always an exact, unique match on name (or name similarity score) within an age (±band), nor is an exact, unique match necessarily a correct link when measurement error skews the number of close candidates.

Hand-Linking Pipeline and Final Linking Decisions

As the examples in Table 1 make clear, making links can be challenging, and even experienced data trainers make errors. To minimize errors in the hand-linked data, each case was independently evaluated by two different, randomly assigned data trainers. LIFE-M data trainers were instructed to reject a link if they were not completely certain it was correct. If the two initial trainers reached an agreement, we coded their choice as the final decision (either link or no link). If the two trainers disagreed, the case was sent to three, randomly assigned trainers for independent re-examination (where the random assignment was constrained to exclude the original two trainers). Then if the three additional trainers reached agreement, we take their agreement as the final decision and treat the disagreement by one of the first-round trainers as measurement error. If the three additional trainers could not reach an agreement (decisions are split 2-3 for or against a link), we code "no link" as the final decision. That is, these cases are those where even experienced trainers cannot agree, which is enough uncertainty to reject the link. In practice, trainers reached an agreement for most cases. As shown in Table 2, when linking G2 birth certificates to the 1940 Census, trainers agreed 93.3% of the time for Ohio and 91.9% for North Carolina. These figures range between 95.7% and 99.8% for other links in Ohio, and 92.3% and 99.1% for other links North Carolina.

Table 2. LIFE-M Training Data by Trainer Decisions

Links	Linkable Cases	Clear	Measurement error	Difficult	Linked but overturned
			Panel A. Ohio		
G2-Marriage	42,617	42,196	167	88	166
		99.01%	*.39%*	*.21%*	*.39%*
G2-Death	45,325	43,135	1,481	649	60
		95.17%	*3.27%*	*1.43%*	*.13%*
G2-Census1940	36,007	30,933	2,166	2,406	502
		85.91%	*6.02%*	*6.68%*	*1.39%*
G1-Marriage	31,661	28,939	1,090	968	664
		91.4%	*3.44%*	*3.06%*	*2.1%*
G1-Death	23,645	21,783	954	678	230
		92.13%	*4.03%*	*2.87%*	*.97%*
G1-Census1940	31,815	28,136	1,314	1,157	1,208
		88.44%	*4.13%*	*3.64%*	*3.8%*
G1-Census1920	31,669	28,016	1,467	934	1,252
		88.47%	*4.63%*	*2.95%*	*3.95%*
G1-Census1910	31,605	28,238	1,381	976	1,010
		89.35%	*4.37%*	*3.09%*	*3.2%*
G1-Census1900	23,946	20,573	1,163	871	1,339
		85.91%	*4.86%*	*3.64%*	*5.59%*
G1-Census1880	8,704	7,750	363	376	215
		89.04%	*4.17%*	*4.32%*	*2.47%*
			Panel B. North Carolina		
G2-Marriage	62,587	56,224	3,173	2,948	242
		89.83%	*5.07%*	*4.71%*	*.39%*
G2-Death	68,265	66,155	1,099	585	426
		96.91%	*1.61%*	*.86%*	*.62%*
G2-Census1940	54,304	44,754	3,905	4,422	1,223
		82.41%	*7.19%*	*8.14%*	*2.25%*
G1-Marriage	44,246	37,056	2,805	1,718	2,667
		83.75%	*6.34%*	*3.88%*	*6.03%*
G1-Death	28,898	26,091	1,089	814	904
		90.29%	*3.77%*	*2.82%*	*3.13%*
G1-Census1940	44,428	35,127	3,548	3,427	2,326
		79.07%	*7.99%*	*7.71%*	*5.24%*
G1-Census1920	44,440	38,469	2,459	2,296	1,216
		86.56%	*5.53%*	*5.17%*	*2.74%*
G1-Census1910	41,436	35,956	2,285	2,252	943
		86.77%	*5.51%*	*5.43%*	*2.28%*
G1-Census1900	27,870	22,814	2,054	1,319	1,683
		81.86%	*7.37%*	*4.73%*	*6.04%*
G1-Census1880	5,504	4,582	487	229	206
		83.25%	*8.85%*	*4.16%*	*3.74%*

Notes: Linkable cases are all cases presented to trainers. Among them, clear cases are those for which the two trainers make an agreement (either linkage or non-linkage) at the first round. Measurement error cases are those for which the first two trainers disagree at the first round, and the additional three trainers reached an agreement in the review session. Ambiguous cases are those for which the first two trainers disagree and the additional three trainers failed to reach an agreement with absolute majority. Lastly, the linked but overturned cases are those linked by trainers but finally overturned due to conflicts.

Quality Control in Hand-Linking

To maintain the quality of the training data and minimize costs, the system provided data trainers with weekly feedback on their accuracy.[9] Accuracy was determined through the distribution of "audit batches" through the data distribution system. Audit batches were carefully selected to represent the records being trained and appeared identical to others, so that trainers could not distinguish between the audit batch and their training work. The system then computed the errors made on audit batches and provided feedback to the trainers at the weekly meeting. In addition, errors on the audit batches that occurred for many trainers were used in weekly discussions. Weekly meetings featured discussions of historical-contextual factors affecting the quality of the records (e.g., men and women misreporting age on marriage certificates but not their day of birth), common linking mistakes, and disagreements over how to code difficult cases. The resulting hand-linked data are not error free, but thoughtful discussion, multiple independent reviews, and audit batches minimize false links and also allow quantification of trainer confidence in different links.

In a final step to ensure quality, the Family History and Technology Lab at Brigham Young University (BYU) performed an independent quality check of the LIFE-M links. BYU compared 1,043 LIFE-M links to those already on the FamilySearch.org "Tree." (FamilySearch.org tree links are created by genealogists and users of FamilySearch.org, who are independent of the LIFE-M process.) For 1,043 birth certificates linked to the 1940 Census by LIFE-M and FamilySearch.org users, the LIFE-M links agreed with FamilySearch.org users 96.7%. Under the assumption that the FamilySearch.org Tree is always correct, this implies a LIFE-M error rate of 3.3%. The true error rate in LIFE-M is likely lower given that some observations on the Tree are incorrect.

Table 3 summarizes the hand-linked resulting from this systematic and careful process. Hand-links to the 1940 Census include 25,727 men (10,977 in Ohio, 14,750 in North Carolina) for a match rate of 42

[9] The project also tracked and provided trainers with feedback on their speed, which was determined using the metadata from the distribution system, which time-stamped upload and downloads of each batch. Tracking trainer speed helped minimize training costs due to inattention. Increasing accuracy also minimized training costs by reducing the number of records sent for discrepancy review.

to 45%, depending on the state. While these rates are low by the standards of modern administrative data, they are considerably higher than studies using supervised or unsupervised learning *without adjusting for high rates of linking error*. In addition, these link rates do not account for mortality or emigration—doing so would raise these link rates further. In addition, LIFE-M hand links to 14,071 women (6,557 in Ohio, 7,514 in North Carolina) to the 1940 Census for a match rate of 23 to 30%, depending on the state. LIFE-M hand links include 34,491 links to death certificates, 21,632 links to marriage certificates, and 46,598 links to G3 children in the birth certificate data and in the 1940 census. Links rates vary from 15 to 51 percent depending on the state, record, and sex.

Table 3: Link Rates [Number of Hand Links] for G2 in the LIFE-M Data

	Siblings (1)	1940 Census (2)	Death (3)	Marriage (4)	Children (5)
			Panel A: Ohio		
All	87.3 [31,299/35,857]	35.4 [17,534/49,519]	29.9 [14,945/49,940]	21.9 [10,941/49,940]	47.7 [23,803/49,940]
Males	87.7 [13,260/15,126]	45 [10,977/24,377]	37.8 [9,295/24,570]	21.9 [5,390/24,570]	50.6 [12,434/24,570]
Females	87.8 [11,901/13,562]	29.9 [6,557/21,931]	24.4 [5,394/22,104]	25.1 [5,551/22,104]	51.4 [11,369/22,104]
			Panel B: North Carolina		
All	83.5 [28,831/34,548]	31.9 [22,264/69,832]	27.8 [19,546/70,219]	15.2 [10,691/70,219]	32.5 [22,795/70,219]
Males	82.8 [13,730/16,588]	42.3 [14,750/34,903]	35.8 [12,574/35,110]	14.6 [5,114/35,110]	34.6 [12,133/35,110]
Females	83.8 [12,731/15,186]	23.2 [7,514/32,328]	21.5 [6,972/32,484]	17.2 [5,577/32,484]	32.8 [10,662/32,484]

Notes: Link rates are calculated as number of links divided by the number of linkable people in the hand-linked sample. Linkable people for G2-siblings link are the G2 who were in the hand-linked sample (for Ohio, a random sample of people born between 1909 and 1920; for North Carolina, a random sample of G2 born between 1915 and 1919) and have non-missing parents' names (father's first name, last name, and mother's first name). Linkable people for G2-1940 Census link are the G2 who were in the hand-linked sample, born by 1940, and have non-missing names (both first and last names). Linkable people for G2-death, G2-marriage, and G2-child links are the G2 who are in the hand-linked sample and have non-missing names. Links to children include all G2s with at least one child (G3) found from either birth records or the 1940 Census. The number of links and linkable people for males and the females may not sum up to the total number, because some G2 have unknown sex (not displayed for brevity). Note that these link rates to not account for mortality and emigration. Doing so would raise link rates. All counts in this table and subsequent tables/figures come from version 1 of the public LIFE-M data released February 2022. Link rates and counts are updated as the data are modified. The most recent data description can be found in documentation at https://life-m.org/.

Appendix Tables D.2-D.4 report the number of hand links and match rates for G0, G1, and G3. Overall, there are more links, and the match rate is higher for G1. Conversely, there are fewer links for G0 and G3, which is expected because (1) more G0 will have died or emigrated before the 1940 Census and the mortality and marriage records begin; (2) more G3 will be born after the 1940 Census, or their records are only available for years beyond those available in digitized collections.

B. Scaling Hand-Linked Data using a Machine Linking Algorithm

Because hand-linking millions of records is cost and time prohibitive, we use our highly vetted hand-linked data to test and select an automated approach. Bailey et al. (2020) report the results of this testing process using three different datasets: (1) the hand links from the Longitudinal Intergenerational Family Electronic Micro-database project (LIFE-M) (created as described previously), (2) the Early Indicators Data (a high-quality hand-linked dataset of men in the Union Army by Dora Costa and team), and (3) a synthetic dataset where the objective truth is known. To be as fair as possible to automated linking methods and provide an independent metric of LIFE-M's hand-linking performance, Bailey et al. (2020) determined linking errors using the following "police line-up" process:

(1) If a record was hand-linked and the output of an automated method, we coded the link as correct.

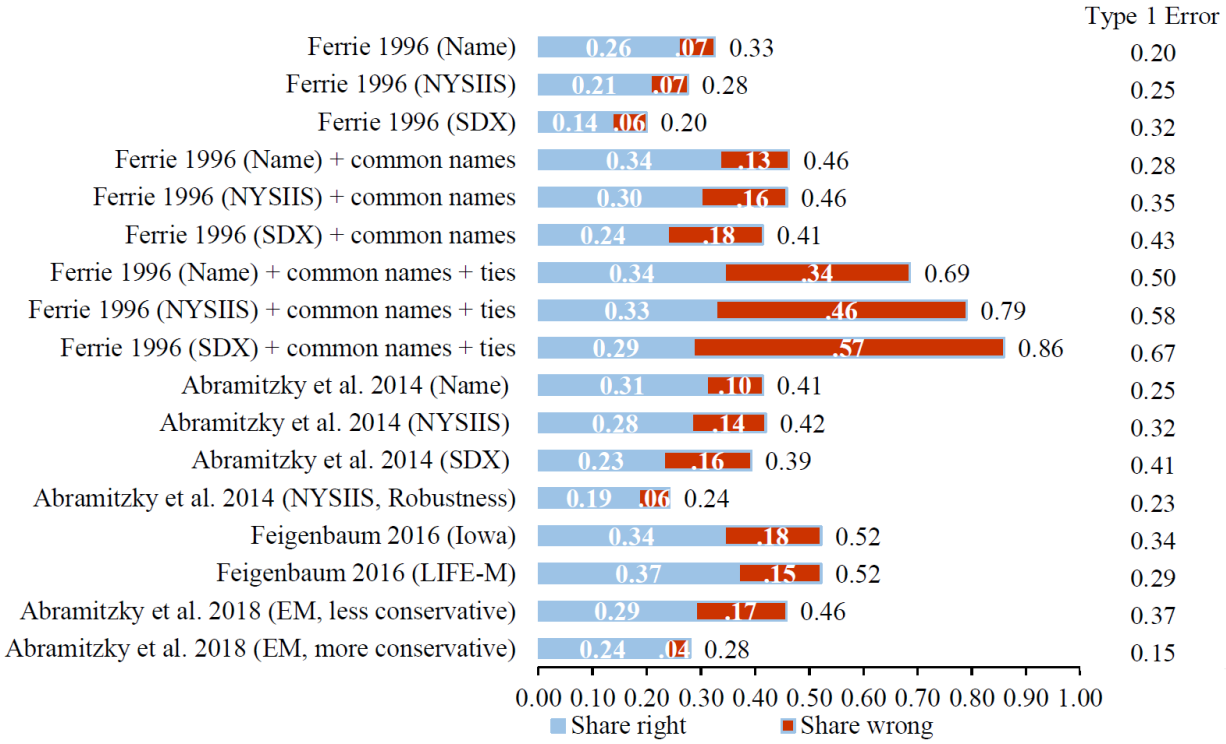
(2) If the algorithm link differed from the hand-link, the record was re-reviewed by two additional reviewers. The set of candidate links included the LIFE-M hand-link (if one was made), the link made by the automated method (if one was made), and a machine-generated set of close matches. Reviewers did not know which link was chosen by which method and were asked to determine the correct link. If these two reviewers agreed, their decision is coded as the truth.

(3) If the two independent reviewers disagreed, the set was sent to an additional three reviewers. If the three trainers agreed, their decision is coded as the truth. If the three trainers disagreed resulting in a 2-3 split among the five trainers, we code "no link."

This process gives the links from the hand-match and the automated method an equal shot at being chosen (or not chosen) to avoid preferential treatment.

Figure 3 summarizes the findings using the LIFE-M hand-links of boys' birth certificates to the 1940 Census. The length of each bar represents the match rate, defined as the share of the baseline sample of boys who were matched to the 1940 Census. Ferrie's (1996) method matched 28% of the baseline sample, and Abramitzky, Boustan, and Eriksson (2014) achieved a higher link rate of around 40%, because the method does not impose Ferrie's (1996) uncommon name restriction. Feigenbaum's (2016) regression-based method matches 52% of the baseline sample, both when using coefficients from his dataset (Iowa) and coefficients from a random sample of the LIFE-M links. Abramitzky, Mill, and Pérez (2018) implementation of Fellegi and Sunter (1969) links 46% of the sample when using less conservative cutoffs and 28% of the sample with more conservative cutoffs.

Figure 3. Match Rates and False Links for Commonly Used Linking Methods



Notes: The bars show the performance of different algorithms for LIFE-M boys linked to the 1940 Census. The length of the bar represents the total share of the records linked. The share of the linked records that is incorrect is displayed in red, and the share of the linked records that are correct is displayed in blue. The Type I error rate (share incorrect/link rate) is displayed in the right column. See Bailey et al. (2020) for more details. A Stata ado-file, "autolink.ado," which we posted at the repository at the Interuniversity Consortium in Political and Social Science Research, can assist other researchers with replicating these analyses (Bailey and Cole 2019).

The column on the right in Figure 3 shows that the estimated Type I, or matching, error (i.e., the share of matches that are incorrect, or share incurred divided by the link rate). The share of false links for automated methods is much higher than clerical review. The lowest Type I error rate occurs in the more conservative version of Abramitzky, Mill, and Pérez (2018) at 15%. Ferrie's (1996) method of selecting uncommon names achieves the second lowest Type I error rate at 25%. Abramitzky, Boustan, and Eriksson (2014) refinement of Ferrie (1996) increases match rates to 40%, but only half of the added links appear to be correct, and the Type I error rate increases to 32%. Feigenbaum's (2016) regression-based machine learning model produces a Type I error rate of 34% when using the Iowa coefficients, and the Type I error rate decreases to 29% when estimated using hand linked data. Finally, Abramitzky, Mill, and Pérez (2018) less conservative cut-off results in the highest error rate at 37%.

Notably, all of the methods in Figure 3 except Feigenbaum (2016) are unsupervised methods. As we later discuss, an important limitation of unsupervised linking algorithms is that they produce samples with high rates of false matches in historical data. Moreover, linking errors have sizable effects on inference (Bailey et al. 2020, Anbinder et al. 2021, Ghosh, Hwang, and Squires 2021).

Based on the insights from this evaluation, the LIFE-M project followed MPC's approach and implemented a supervised learning method to link historical records. A key advantage of using training data to discipline machine models is that researchers can *control* the trade-off between making more links and making fewer false links. The resulting method is designed for linking in limited information settings and uses careful feature engineering to improve performance.

Model Architecture

We developed two model architectures based on the number of links per primary. Link types that allow only one match per primary record are called one-to-one links (e.g., birth certificate to census), and those that allow multiple matches per primary are called one-to-many links (e.g., birth certificates to sibling birth certificates, birth certificate to marriage certificate). We use a methodology developed in conjunction with Jared Murray to model the first case, and random forest to model the second (Breiman 2001).

The one-to-one linking model is developed in two stages. The first stage is called the "any-match" model, and the second stage is called "which-match" model. Any-match model estimates the likelihood that a link exists *within the set of records for a case*, while the which-match model is used to predict the *pairwise* probability for each primary and potential within the set. We multiply the probability from the first and second stage to obtain the final probability that a link exists between each primary and candidate pair. The multiplication is motivated by the law of total probability conditioning on whether the set contains the link or not. We use a random forest to estimate the any-match model, and a custom log-linear model to estimate the which-match model (Breiman, 2001). The custom log-linear model ensures the probability of all primary-candidate pairs within the set sums up to 1, so that the final output yields at most one unique link per set. One-to-many linking models use a flexible random forest. The model takes pairwise information for the primary and candidate links and then predicts the probability of each being a match.

Feature Generation

The quality of the machine-models also depends on the selection of features, or input attributes that describe the characteristics of the data. The aim of the any-match model is to predict probability that the set contains *one and only* match for the primary record in consideration. Models typically include a set of 20 to 80 features ("variables") that help the model distinguish a link from a no-link or ambiguous case. Examples of the set-level features include the commonality of the primary record name, the number of candidates having exact or close matches on the first and last names, and the difference of the name similarity scores between the first and the second closest candidate (Feigenbaum 2016). We also find that including features relating to the top-five or top-ten matches further aids the model in learning the human decision-making process.

The aim of the which-match model is to predict *the pairwise* probability for each primary and candidate within the set. The goal is for the model to output a high probability if the name and other information between the primary and candidate match well, and a low probability if not. This model typically includes five to ten pairwise features. Examples of highly predictive features include Jaro-Winkler

distance between the primary and candidate names and the age difference between the primary and candidates.

We also found it beneficial to performance to customize model features to different records and states, because the information used for linking and its quality differs. The decision-making process also varies due to state-specific demographics and data availability. For example, Ohio has more foreign-born names, whereas North Carolina has more shorter and common names.

Model Training and Cross Validation

To test model performance, we divided the hand-linked data into two equal parts. One part is the training sample, while the second part is the test sample and is reserved for validation of the final model performance. We use the training sample to estimate the model using cross-validation and then apply the model on the test sample to report performance.

In the training sample, we follow the common practice of using ten-fold cross validation (Hastie, Tibshirani, & Friedman 2009). This works by first shuffling the training data and then randomly splitting into ten subsets of the same size ("folds"). For each subset, we select the rest of the nine subsets to estimate the model and use the single hold-out subsample to predict and estimate the error rate. After this process is repeated for each subsample, we average the error rates obtained from ten subsets to get the final performance metrics. The main advantages of cross-validation are (1) it uses all the available data to estimate the model, (2) avoids overfitting to a specific subset of the data, and (3) has more accurate performance metrics.

After the final probability of a match is estimated for every primary-candidate pair, we generate machine links using a simple threshold rule: pairs with a probability exceeding some probability thresholds are classified as links. A key decision is how to set this threshold. Each threshold is associated with a "recall rate," the share of hand links that the model is able to reproduce, and a "precision rate," the share of model links that are correct according to the hand links. Raising the threshold tends to increase the precision rate at the cost of a lower recall rate, while lowering the threshold has the opposite effect. There is no optimal threshold, so this choice ultimately depends on the modeler's preferences. LIFE-M chooses the threshold

associated with a 97% precision rate, or equivalently, a 3% Type-I error rate. The trained model and probability thresholds are then used to classify links for the full sample of records. The threshold varies model by model, and but it is typically between 0.70 to 0.95. Model performance is evaluated using the test sample with the cutoff threshold determined by the training sample.

Figure 4 presents precision-recall frontiers for methods commonly used in historical linking for male birth certificates linked to the 1940 Census. The x-axis is the share of true links found by the algorithm divided by number of hand-links links, and precision is 1− the share of links made by the algorithm that are incorrect (or 1−the Type I error rate in Figure 3). For low rates of recall, unsupervised methods deliver high precision. However, the precision of these methods falls precipitously as recall increases. The trade-off is less dramatic for supervised methods, such as Feigenbaum and LIFE-M, which can deliver precision rates over 90% for recall rates exceeding 70% for North Carolina and 80% for Ohio. Our testing reveals that differences between the performance of LIFE-M's and Feigenbaum's methods reflects the extensive set of features in the former (not the difference between using a random forest and regression).

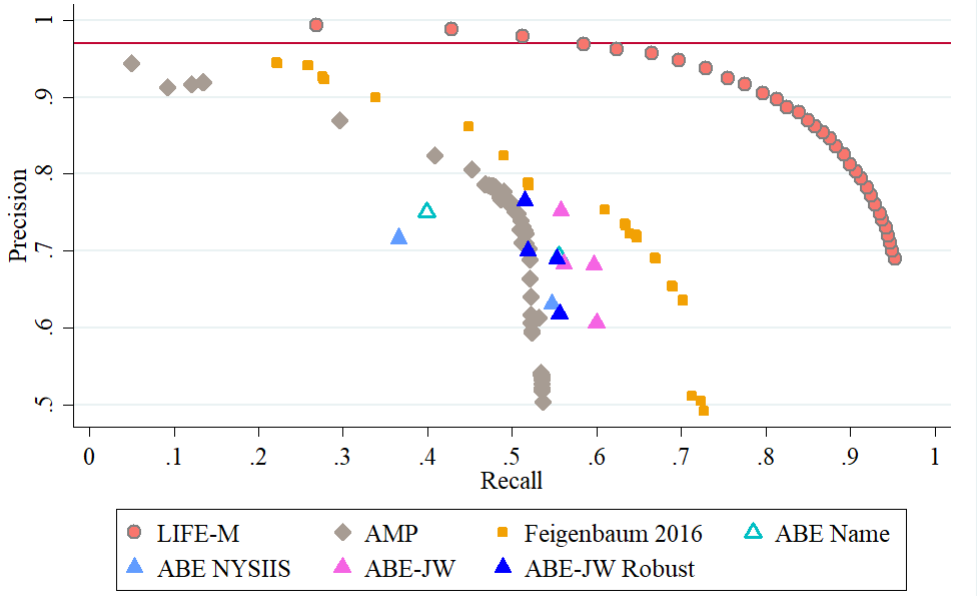
V. Linked Samples

Ohio and North Carolina contain over 3.9 million linkable G2 individuals.[10] Depending on the state, 41 to 44% of these linkable observations are women. Without supervised learning methods, connecting these many individuals to their birth and marriage families, death certificates, and census data would be nearly impossible. However, supervised learning methods described in section IV allow us to combine these records with high precision at scale. The next sections describe link rates, representativeness, and the creation of weights.

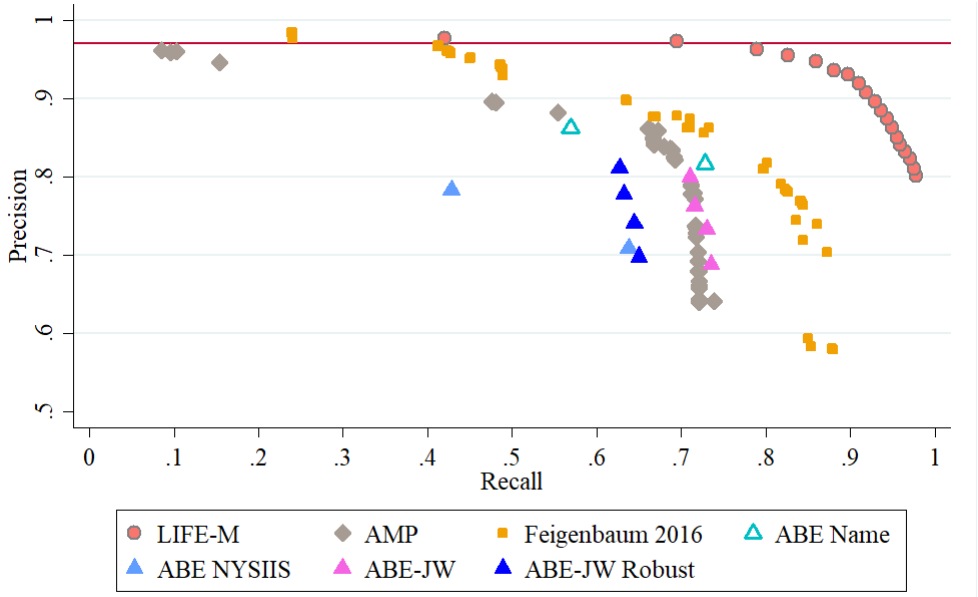
[10] Linkability is determined by the completeness of name and birth year (for certain links). The condition of being linkable people for different links are described in the notes of Table 4.

Figure 4. LIFE-M Methods Expand the Precision-Recall Frontier in Historical Linking

A. North Carolina



B. Ohio



A. Link Rates and Counts

Table 4 provides the link rates for G2 in the full data. The link rates are calculated separately for each state and represent the proportion of observations that can be linked with 97% precision to their siblings (column 1), the 1940 Census (column 2), their death and marriage records (columns 3 & 4), and their G3 children (column 5).[11] Because the model is only able to replicate some of trainers decisions with a 3% error rate, link rates are lower in the full sample. Depending on the state and sex, 67 to 74% of G2 are linked to their siblings, 12 to 28% are linked to the 1940 Census, 13 to 28% are linked to their death records, 7 to 22% are linked to their marriage records, and 18 to 32% are linked to their children in G3. Overall, the link rates are higher in Ohio than North Carolina. This is due to Ohio having more complete records than North Carolina. In addition, link rates are higher for men than women, except for the marriage records. In general men are easier to link because their last name does not change across records. Approximately, the same number of men and women are linked to their marriage records, so the difference in the rate comes from women being more difficult to link overall.

As in the linking process described in section III, G2 can be connected to grandparents (G0), parents (G1), and children (G3). The Appendix Tables D.5-D.7 provides the total number of links and link rates for G1 in the full data. There are over 3.4 million and 815 thousand linkable G1 in Ohio and North Carolina, respectively. The counts of men and women and match rates across sex are similar, largely because G1s are often linked as couples. Match rates to census records range from 14 to 30% in Ohio and 11 to 21% in North Carolina. The highest match rate is for the 1940 Census. The match rates to marriage records range from 25 to 34%, depending on the state, and the match rates to death records range from 11 to 18%. Approximately, 28 to 39% of observations could be linked to their parents (G0).

[11] Linking with 97% precision, means the error rate is only 3%. For the 1940 Census and death records, we can also link with higher error rates of 5 and 10%. The advantage of a higher error rate is more links, thus larger samples. However, the samples only increase in size by, at most a few hundred thousand. The linking rates at 90 and 95% precision are available upon request.

Table 4: Link Rates (Total Links/Linkable Records) for G2 in the LIFE-M Data

	Siblings	1940 Census	Death	Marriage	Children
	(1)	(2)	(3)	(4)	(5)
			Panel A: Ohio		
All	73.7 (2,388,882/3,239,296)	24.1 (819,715/3,408,048)	22.5 (778,745/3,458,533)	19.4 (670,329/3,458,533)	28.1 (972,179/3,458,533)
Males	74.3 (1,040,556/1,401,213)	28 (461,135/1,645,821)	27.8 (464,533/1,670,413)	20.1 (335,190/1,670,413)	31.5 (525,876/1,670,413)
Females	74.4 (948,696/1,275,803)	23.9 (358,580/1,498,495)	19.3 (293,726/1,519,088)	22.1 (335,139/1,519,088)	29.4 (446,303/1,519,088)
			Panel B: North Carolina		
All	67.1 (484,046/721,668)	13.5 (173,408/1,289,239)	15.7 (204,870/1,305,141)	7.4 (96,564/1,305,141)	17.9 (233,196/1,305,141)
Males	67.8 (221,808/327,190)	18.6 (108,655/585,879)	22.9 (136,206/593,678)	8.2 (48,694/593,678)	22 (130,868/593,678)
Females	68.5 (204,672/298,960)	12.1 (64,753/533,541)	12.7 (68,663/540,697)	8.9 (47,870/540,697)	18.9 (102,328/540,697)

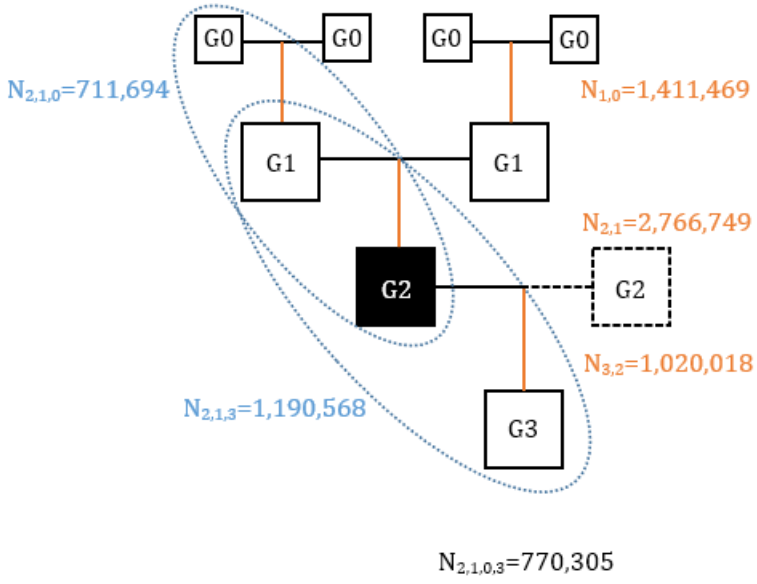
Notes: Link rates are calculated as number of links divided by the number of linkable people. Linkable people for column 1 are G2s who were born from 1900-1929 and have non-missing parents' names (father's first name, last name, and mother's first name). Linkable people for column 2 are G2s who were born by 1940 and have non-missing first and last names. Linkable people for columns 3-5 are G2s who have non-missing names. The number of links and linkable people for the males and females may not sum up to the total, because G2 have unknown sex (unknown text category not displayed for brevity).

There are over 2.6 million G0 in full data (2.2 million in Ohio, and 400 thousand in North Carolina). G0 is linked to censuses and death records at full scale (except for the G0-1940 Census, which was only implemented by hand linking). Link rates of G0-Census links range from 5 to 19%, with very few variations across sex. Finally, there are over 2.3 million G3 in full data (1.9 million in Ohio and 460 thousand in North Carolina). Link rates of G3-1940 Census links range between 17 and 20%, depending on state and sex. Link rates of G3-death and G3-marriage links are significantly lower (between 3 and 11%, depending on state and sex group), because the death and marriage records used for linking become increasingly incomplete after mid or late 20th century, when most G3 reach marriage age or old age.

Figure 5 provides an overview of the number of adjacent links (parents to children) and non-adjacent links (grandparents to grandchildren) we can make to determine the number of families in LIFE-

M. There are over 4 million unique adjacent links and more than 2 million unique non-adjacent links. Over 770,000 G2 can be linked to at least one grandparent and a child.

Figure 5. LIFE-M Partial Four-Generation Family Tree



Notes: The dark shaded G2 box provides a reference point for most of the family tree definitions, and the dotted G2 box shows the presence of a spouse. However, not all spouses in G2 are present, nor do they need to be to determine the partial family tree. The orange links graphically show 2-generation families and the numbers ($N_{3,2}$, $N_{2,1}$, $N_{1,0}$) indicate the number of unique 2-generation families from the child's perspective. Families with multiple children are only counted once. The blue circles show how 3-generation families can be constructed. The blue numbers ($N_{2,1,3}$, $N_{2,1,0}$) identify the number of 3-generation families, relative to G2. $N_{2,1,3}$ identifies the number of G2 that are both a parent and child. $N_{2,1,0}$ identifies the number of G2 with a unique grandparent. Siblings in G2 have the same grandparents and are only counted once. Lastly, the entire graphic shows a partial 4-generation family. The black number ($N_{2,1,0,3}$) shows how many G2 have both at least one grandparent and at least one child. Approximately, 73% of G2 are in Ohio, and the remaining 27% are in North Carolina.

B. Representativeness of Linked Sample

We benchmark the representativeness of the G2 and G1 linked samples against population data. Table 5 reports the F-statistics of Wald tests for joint significance of a set of covariates on the linkage status (Detailed regression results can be found in Appendix Tables C.1-C.2). For all types of linking, the joint significance tests reject the hypothesis that the linked sample is a random sample of the population.

Table 5 Representativeness of Linked G2 and G1 by Link Types

	G2-1940 Census	G2-Death	G2-Marriage	G1-1940 Census	G1-1920 Census	G1-1910 Census	G1-1900 Census	G1-1800 Census	G1-Death	G1-Marriage
All	19507.9	14471.4	19623.9	258727.6	149041.9	57725.5	49010.9	8802.8	52393.9	224693.6
p-value	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*
Male	11936.7	5975.3	10918.0	112248.5	79551.0	51781.4	33963.9	5988.3	35361.9	114303.0
p-value	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*
Female	9498.7	4414.7	11057.0	116201.9	84318.8	53665.0	27964.8	3782.0	26570.2	114759.5
p-value	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*

Notes: We report F-statistics of heteroskedasticity-robust Wald-test from each separate regression of a binary dependent variable which equals to one for linked record in LIFE-M data. Corresponding p-values are reported in italics and parenthenesses. The covariates for G2 links include the birth day of G2 children, number of children in G1-G2 family, name length of G2 children, name length of G1 father, name length of G1 mother, share of birth records with misspelled father's name, share of birth records with misspelled mother's name, and a dummy variable for Ohio links.

C. Creation of Weights for Linked Data

To address the imbalance in observed characteristics, we use inverse propensity-score weighting as described in Bailey, Cole, and Massey (2020). The approach assigns a larger weight for the people under-represented relative to the target population, and a lower weight for over-represented subgroups.

Taking G2-1940 Census links as an example, Table 6 compares the mean of some demographic and socioeconomic variables for the linked G2 (before and after weighting) with a target population of the 1940 Census. Notably, the unweighted linked sample (column 1) is statistically different from the 1940 Census (column 4). After applying weights, however, the mean differences of observed characteristics between the linked G2 sample and the population in census are both significantly smaller (e.g., 0.06 difference in age) and typically insignificant.

Table 6. Characteristics of the Linked Sample, with and without Weights

	Linked G2, Unweighted (1)	Linked G2, Weighted (2)	1940 Census (3)	Difference (1)-(3)	Difference (2)-(3)
		Panel A: Ohio			
Age in 1940	20.009	19.510	19.454	0.555***	0.057***
Male	0.561	0.504	0.505	0.056***	-0.001
White	0.983	0.970	0.970	0.013***	0.000
Never Married	0.692	0.659	0.660	0.032***	-0.001
Urban Status	0.615	0.641	0.642	-0.028***	-0.001
Farm Status	0.192	0.165	0.162	0.030***	0.003***
Migration: 5-year	0.036	0.049	0.048	-0.012***	0.001***
High School Graduate	0.546	0.503	0.501	0.045***	0.003***
Occupational Income Score	23.500	23.704	23.703	-0.203***	0.002
		Panel B: North Carolina			
Age in 1940	20.572	19.615	19.670	0.902***	-0.055
Male	0.613	0.509	0.510	0.104***	-0.000
White	0.819	0.723	0.716	0.103***	0.007***
Never Married	0.669	0.632	0.628	0.041***	0.003
Urban Status	0.284	0.286	0.287	-0.003*	-0.001
Farm Status	0.444	0.452	0.450	-0.007***	0.002
Migration: 5-year	0.045	0.045	0.045	0.001	0.001
High School Graduate	0.407	0.336	0.334	0.073***	0.002
Occupational Income Score	19.662	18.620	18.578	1.085***	0.042

Notes: Column (1) reports the unweighted means of characteristics of G2s linked to the 1940 Census. Column (2) reports the weighted mean of these characteristics for G2s linked to the 1940 Census. Column (3) reports the means of characteristics for the 1940 Census population for those who were born in Ohio/North Carolina between 1900 and 1940. Migration status from birth to 1940 is defined by comparing the residence state in the 1940 Census and reported birthplace. Five-year migration status is directly reported in the 1940 Census for people born by 1935. Column (4) and (5) reports difference in means between indicated columns.

The resulting weights create global adjustments to individuals to enhance the representativeness of the total sample or subgroups with respect to the linkable population. Using a similar approach, researchers could generate inverse propensity-score weights to make the linked sample balanced relative to the universe of birth records available for linking or other reference populations.

VI. Opportunities for Research

The LIFE-M have been released for public use and the most updated version can be downloaded from https://life-m.org (data are also hosted at OpenICPSR[12]. LIFE-M's millions of linked records provide a new historical data resource for health and social scientists and reduce barriers for path-breaking and high-impact research on fertility, mobility, health, and aging—both across the life course and across generations. LIFE-M also facilitates more research on families, by connecting a large set of women—mothers, daughters, and sisters—and allowing researchers to follow them across time alongside fathers, sons, and brothers.

On-going research connects LIFE-M to new health data. New work has digitized cause of death for some Ohio individuals (Bailey et al. 2022-03-15), which allows researchers to study the correlates of aging and mortality, including the relationships between a multitude of early-life and intergenerational factors contributing to longevity *and* the cause of death. LIFE-M can also be linked to other historical data sources, such IPUMS-MLP (which includes lifemid on its website), and World War II enlistment records. Finally, the geographic file follows the location of individuals from birth to death, including census records and marriage, which allow researchers to connect the millions of individuals to previously unstudied policy interventions based on the county and period when they occurred.

VII. References

A'Hearn, Brian, Jörg Baten, and Dorothee Crayen. 2009. "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital." *The Journal of Economic History* 69 (3):783-808. doi: 10.1017/S0022050709001120.

[12] https://www.openicpsr.org/openicpsr/project/155186

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2012. "Web Appendix: Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration."

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122 (3):467-506.

Abramitzky, Ran, Leah Boustan, and Myera Rashid. 2020. Census Linking Project: Version 1.0.

Abramitzky, Ran, Roy Mill, and Santiago Pérez. 2018. "Linking Individuals Across Historical Sources: a Fully Automated Approach." *National Bureau of Economic Research Working Paper Series* No. 24324. doi: 10.3386/w24324.

Anbinder, Tyler, Dylan Connor, Cormac Ó Gráda, and Simone Wegge. 2021. "The Problem of False Positives in Automated Census Linking: Evidence from Nineteenth-Century New York's Irish Immigrants." *CAGE working paper no. 568 2021*.

Bailey, Martha, Karen Clay, Price Fishback, Michael Haines, Shawn Kantor, Edson Severnini, and Anna Wentz. 2016. U.S. County-Level Natality and Mortality Data, 1915-2007. Ann Arbot, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor].

Bailey, Martha, Connor Cole, and Catherine G. Massey. 2020. "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850-1930 IPUMS Linked Representative Historical Samples." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (2):80-93. doi: 10.1080/01615440.2019.1630343.

Bailey, Martha J., and Connor Cole. 2019. "Autolink.ado." accessed 2019-06-13. http://doi.org/10.3886/E110164V1.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine G. Massey. 2020. "How Well Do Automated Linking Methods Perform? Evidence from US Historical Data." *Journal of Economic Literature* 58 (4):997-1044. doi: http://www-personal.umich.edu/~baileymj/Bailey_Cole_Henderson_Massey.html.

Bailey, Martha J., Susan H. Leonard, Joe Price, Evan Roberts, Logan Spector, and Mengying Zhang. 2022-03-15. LIFE-M Ohio Causes of Death. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Bailey, Martha J., Abdul Raheem Shariq Mohammed, and Paul Mohnen. 2021. "Measuring Intergenerational Mobility in the U.S. in the Early 20th Century." *Working paper*.

Biavaschi, Costanza, and Benjamin Elsner. 2013. "Let's be Selective about Migrant Self-Selection." *IZA Discussion Paper 7865*.

Blank, Rebecca M., Kerwin Kofi Charles, and James M. Sallee. 2009. "A Cautionary Tale about the Use of Administrative Data: Evidence from Age of Marriage Laws." *American Economic Journal: Applied Economics* 1 (2):128-149.

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

Bleakley, Hoyt , and Joseph P. Ferrie. 2013. "Up from Poverty? The 1832 Cherokee Land Lottery and the Long-run Distribution of Wealth." *NBER Working Paper 19175*.

Bleakley, Hoyt, and Joseph Ferrie. 2017. "Land Opening on the Georgia Frontier and the Coase Theorem in the Short- and Long- Run." http://www-personal.umich.edu/~hoytb/Bleakley_Ferrie_Farmsize.pdf.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45:5-32. doi: https://doi.org/10.1023/A:1010933404324.

Collins, William J., and Marianne H. Wanamaker. 2014. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics* 6 (1):220-252.

Elo, Irma T, and Samuel H. Preston. 1994. "Estimating African-American mortality from inaccurate data." *Demography* 31 (3):427-458.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking." http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf?m=1423080976. Accessed March 28, 2016.

Fellegi, Ivan P., and Alan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328):1183-1210. doi: 10.1080/01621459.1969.10501049.

Ferrie, Joseph P. 1996. "A New Sample of Males Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules." *Historical Methods* 29 (4):141-156.

Ghosh, Arkadev, Sam Il Myoung Hwang, and Munir Squires. 2021. "Links and legibility: Making sense of historical U.S. Census automated linking methods." *Working paper*.

Goldstein, J.R. , and C. Breen. 2020. "Title." Working Paper.

Guest, Avery M. 1987. "Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century." *Historical Methods* 20 (2):63-77.

Hacker, J. David. 2013. "New Estimates of Census Coverage in the United States, 1850-1930." *Social Science History* 37 (1):71-101.

Hastie, Trevor, Robert J. Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer-Verlag.

Helgertz, Jonas, Steven Ruggles, John Robert Warren, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Joseph P. Price, Evan Roberts, and Matthew Sobek. 2020. PUMS Multigenerational Longitudinal Panel: Version 1.0 [dataset]. edited by Minnesota Population Center. Minneapolis, MN: IPUMS.

Kennedy, Sheela, and Steven Ruggles. 2014. "Breaking Up Is Hard to Count: The Rise of Divorce in the United States, 1980-2010." *Demography* 51 (2):587-598.

Logan, Trevon, and John M. Parmen. 2011. "Race, Socioeconomic Status, and Mortality in the 20th Century: Evidence from the Carolinas." *University of Michigan Population Studies Center Working Paper* SC Research Report No. 11-739.

Long, Jason, and Joseph Ferrie. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States since 1850." *American Economic Review* 103 (4):1109-1137.

Ruggles, Steven. 2006. "Linking Historical Censuses: A New Approach." *History and Computing* 14 (1-2):213-224.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2010. Integrated Public Use Microdata Series (Version 5.0) [Machine-readable database]. Minneapolis: University of Minnesota.

Wimmer, Larry T. 2003. "Reflections on the Early Indicators Project: A Partial History." In *Health and Labor Force Participation over the Life Cycle: Evidence from the Past*, edited by Dora L. Costa, 1-11. Chicago, IL: University of Chicago Press.

Appendix A. Completeness of Vital Records

To characterize the completeness of vital records used in LIFE-M, Figure A.1 compares the statistics of LIFE-M vital records to aggregated counts reported by published tabulations of births and deaths data. It is noteworthy that comparison to vital tabulations is only possible *after* 1909, when Ohio entered the Federal Registration Area. Therefore, in the years before vital tabulations were available, we estimate birth counts based on the age distribution in most recent decennial censuses from the 1880, 1900, and 1910 Census. We evaluate this method's validity by comparing the estimated birth counts (in the same approach) to the actual vital tabulations between 1909 and 1940.[13]

Figure A.1 plots over-time ratio of total birth counts of LIFE-M vital records to counts of vital tabulations (or estimated counts from census before 1909). We also evaluate completeness of death records by comparing the total counts of LIFE-M death records to death tabulations. Given that some vital records can be duplicates, we calculate coverage ratios either before (the read solid line in Figure A.1) or after (the blue dash line in Figure A.1) dropping the duplicated records. In Ohio, the coverage of birth records hovered around 100% beginning in the early 1900s, and the coverage of death records were very close to 100% beginning in 1915, when the death tabulations were available. The coverage of birth records reached completeness around 1915 and gradually decline after 1950. The death records in North Carolina remained very close to completeness since 1915. This suggests that LIFE-M birth and death records capture a sample very close to the universe of births and deaths for most of the 20th century.

Appendix B. Data Cleaning Procedure

The ideal in record linkage is an exact, unique match. In reality, errors in data from a variety of sources and common record features frequently yield inexact matches or non-unique matches. The LIFE-M linkage process begins with a comprehensive record cleaning to minimize errors in the vital records and

[13] Figure A.2 in Appendix plots the estimated birth counts from age distribution in Censuses, as well as the vital tabulated birth counts. For overlapping years (1909-1940), they were very close to each other. However, we are also aware that this can be a result of continuous decline in infant and children mortality rates in 20th century.

census data. Cleaning aims to increase true match rates by identifying and correcting potential problems before linkage occurs. For example, we clean name abbreviations and diminutives (“Wm.” to “William”), standardize names (“Le Roy" to “LeRoy”), remove suffixes and prefixes (“Colonel”, “General”, “Major”, etc.), transcribe parent names when father’s name is reported in mother’s column and vice-versa, and finally, parse names into new name substring fields (first, middle, last, mother’s birth name etc.). These data processing steps allow us to increase true match rates by improving comparability in the datasets to be linked.

In addition to names, the cleaning also involves parsing dates and assigning geographic codes when possible. Years of an event (year of birth, year of marriage, year of death) are often very simple to clean, day and month of an event are cleaned only when this information is available and can be confidently classified. For geographic cleaning, we only clean counties for Ohio and North Carolina, when available. Examples include county of birth, marriage, and death. Broadly, county cleaning follows three steps:

(1) Clean county strings. Examples of this cleaning include removing un-informative words such as Ohio (because counties are known to belong to Ohio), USA, Out of State etc. This cleaning also includes trimming empty spaces and setting words to the proper case.

(2) Search for exact matches on county names. If more than one county name is found within a string, then we keep the county name that appears later in the string.

(3) Inspect misspelled county names. In order to maximize the gains, we restrict our inspection to missed county names that occur more than a certain threshold (e.g., 100, 50). Examples of cases cleaned in this stage include Ashtabula misspelled as Ashtubula or Auglaize misspelled as Augalize.

Once strings are assigned to one of the county names within a state, we use the ICPSR coding scheme to assign county codes. Unlike counties, cleaned places relate to states within the United States and foreign countries. Examples include the place of birth of kids, father, mother, and spouse. The steps followed in county cleaning are similar to county cleaning except that we use IPUMS BPL coding scheme to assign geographic codes to places.

Appendix C. Generating Sets of Potential Candidates

As an illustrative example, consider the problem of linking a boy from his birth certificate in Ohio to the 1940 Census. To generate a set of potential candidates based on name similarity scores (which is computationally intensive), we must first reduce the dimensionality of the problem by narrowing the set of potential candidates, which we achieve "blocking" on certain characteristics following the literature (e.g., Ferrie 1996, Abramitzky Boustan Eriksson 2012). In addition to reducing the computational burden, another advantage of blocking is to impose certain conditions that we want links to satisfy, and thereby simplify trainers' decision problem. The downside of blocking is that the true link might not satisfy the conditions we impose due to misspecified information.

Therefore, our approach is to block on a few key characteristics that are sufficient to make the problem tractable. In the example above, we restrict the set of potential candidates in the 1940 Census to individuals who were born in Ohio, have a last name initial matching the last name initial on the birth certificate, and specify an age within a +/- 3 year window around the age implied by the year of birth on the birth certificate. We could have further restricted the set of potentials to men, but it is possible for sex to be misspecified on birth certificates or in the census, so we chose not to.

Once we have restricted the set of potentials to those that satisfy the conditions above, we compute a string similarity score between the primary under consideration and *every* potential in the block. Returning to the example above, we compute a bi-gram score normalized between 0 and 1 based on two concatenated strings: "[last name] [first name]" and "[first name] [middle name]." We then rank the potentials, drop candidates whose score is too low, and select the top 30 remaining candidates (there can be fewer if the block contains fewer than 30 candidates, or if fewer than 30 candidates satisfy the minimum score requirement). This process is repeated for all primaries we want to link. Note that the set of blocking criteria and name similarity scores naturally varies across linkages depending on the available information (see Appendix Table D.1 for details).

Appendix D. Machine-Linking Algorithm

The LIFE-M team used the hand-linked data to train a machine model to make similar linking decisions as the human trainers. To explain the mechanics of the model, some notation is helpful. Consider a primary record i with k_i potential candidate matches. For each candidate j for primary i (where $1 \leq j \leq k_i$), let Y_{ij} be an indicator equal to 1 if j is a match for primary i and 0 otherwise. Let R_i be an indicator equal to 1 if $Y_{ij} = 1$ for any candidate j and 0 otherwise. The input for the model consists of the training data (i.e., vector of R's for all primaries and vector of Y's for each primary i), as well as two types of features chosen by the modeler. Pair-level features, which we denote by X_{ij}, are covariates that are functions of primary/candidate pairs (e.g., name-similarity scores, age difference discrete variable). Set-level features, which we denote by Z_i, are covariates that are functions of the entire primary/candidate *set*. Examples of set-level features include average/maximum string distance scores, the difference between the highest and second-highest score, and indicators for a unique potential matching the primary along various dimensions.

The goal of the model is to estimate the probability that j is a match for i, which we denote by π_{ij}, using the provided features. Our approach to modeling π_{ij} is to decompose it into two components:

$$\begin{aligned}\pi_{ij} &\equiv \Pr(Y_{ij} = 1 \mid X_{ij} = x_{ij},\ Z_i = z_i) \\ &= \Pr(Y_{ij} = 1,\ R_i = 1 \mid X_{ij} = x_{ij},\ Z_i = z_i) \\ &= \Pr(Y_{ij} = 1 \mid R_i = 1, X_{ij} = x_{ij}) \Pr(R_i = 1 \mid Z_i = z_i). \qquad (1)\end{aligned}$$

The first equality follows from the definition of Y and R: if j is a match for primary i, then there must exist a match in the set of potential candidates for primary i (i.e., $R_i = 1$). The second equality follows from the law of total probability. The first term on the right-hand side of equation (1) denotes the probability that any record in the candidate set is the match, conditional on the set containing a match, which we call the "which-match" probabilities. The second term denotes the probability that the set of candidates for primary i contains a match, which we call the "any-match" probabilities.

The any-match probabilities can be estimated using standard regression/classification methods, such as multinomial logistic regression. However, in our application we found that simple parametric methods were outperformed by more flexible nonparametric methods. Since the trainer's decision to declare a match is a complex function of different features of the primary-candidate set, we expect this will be true in general. In practice, we opted for random forests, which we estimate using the vector of R's and the vector of Z's. The output of the random forest are probabilities that correspond to the fraction of decision trees in the random forest that classified primary *i* as having a match.

Modeling the which-match probabilities requires a custom model because these probabilities must sum to one by definition. We impose this one-to-one matching constraint by using a log-linear model

$$\Pr\left(Y_{ij} = 1 \middle| R_i = 1, X_{ij} = x_{ij}\right) = \frac{\exp\left(x_{ij}^{'}\beta\right)}{\sum_{h=1}^{k_i} \exp\left(x_{ih}^{'}\beta\right)}, \tag{2}$$

where primary *i* has k_i candidate matches, and β is a vector of regression coefficients. The β parameters are identified provided the *x* vector does not include an intercept, and that each covariate has nonzero variance over the candidate set for at least one primary. The latter condition is satisfied by any useful covariate in the which-match model – covariates that are constant within every primary-candidate set belong in the any-match model. The coefficients in the "which-match" model are directly interpretable:

$$\log\left[\frac{\Pr\left(Y_{ij} = 1 \middle| X_{ij} = x_{ij}, R_i = 1\right)}{\Pr\left(Y_{ih} = 1 \middle| X_{ih} = x_{ih}, R_i = 1\right)}\right] = \left(x_{ij} - x_{ih}\right)'\beta \tag{3}$$

That is, the log odds of a match to record *j* over record *h* – given that there is a match to be found – is linear in the difference of their distance or similarity measures. This also provides some intuition for why our method outperforms pairwise models, as we infer a set of coefficients that are specifically designed to discriminate between candidate matches.

We fit the which-match model using penalized maximum likelihood with a ridge penalty on β, to stabilize the estimation of coefficients corresponding to rare and low-variance covariates. The model is estimated using the vector of Y's and X's, on the subset of primaries that had a match ($R_i = 1$). The likelihood function for the which-match model is simply

$$L(\beta; Y) = \prod_{i:R_i=1} \prod_{j=1}^{k_i} \left(\frac{\exp(x'_{ij}\beta)}{\sum_{h=1}^{k_i} \exp(x'_{ih}\beta)} \right)^{y_{ij}}, \quad (4)$$

which yields a tractable maximization problem.

In practice, the any-match and which-match probabilities are estimated using 10-fold cross-validation (Hastie, Tibshirani, and Friedman 2009).[14] We first estimate these probabilities using a fraction of the data (referred to as training sample) while reserving rest of the data as validation set for final evaluation (referred to as test sample). The resulting probabilities are multiplied to yield the estimated unconditional match probability $\hat{\pi}_{ij}$ for every primary-candidate pair. Machine links can then be generated using a simple threshold rule: pairs with a probability exceeding $\bar{\pi}$ are called links. A key decision is how to set this threshold. Each threshold is associated with a "recall rate," the share of hand links that the model is able to reproduce, and a "precision rate," the share of model links that are correct according to the hand links. Raising the threshold tends to increase the precision rate at the cost of a lower recall rate, while lowering the threshold has the opposite effect. There is no optimal threshold, this choice ultimately depends on the modeler's preferences. At LIFE-M we typically chose the threshold associated with a 97% precision rate. Note that match probabilities derived from the model can also be used in weighting or imputation (Abowd et al., 2019). Lastly, with the trained model and probability threshold in hand, linking can be scaled to the full sample of records.

Note that the framework above can accommodate the richness of the LIFE-M training data and ambiguous cases where different trainers might reasonably reach different decisions. In the context of the model, we can simply allow R_i to take three values instead of two, with the additional category capturing sets in which there was some disagreement among the trainers (as opposed to all trainers agreeing there was no link or all selecting the same link). The any-match model then generates three probabilities instead

[14] 10-fold cross-validation involves splitting the data into 10 equal-sized "folds," and using 9 folds to generate predicted probabilities for the remaining fold. This process is repeated 10 times, each time varying the fold under consideration, such that the resulting probabilities are not fitted on the data itself.

of two, which requires an additional "calibration" step to map them back into the canonical setup with two probabilities.[15] In practice, we have found that this additional layer has little effect on the model performance.

While our machine learning algorithm yields at most one link for each primary, it is possible for the same potential in the target dataset to separately be linked to multiple primaries which is a violation of one-to-one linking. This can occur for potentials with very common names and follows from the fact that each primary is linked in isolation, as in most algorithms.[16] We follow standard procedure and conservatively discard all conflicting links in such cases (unless one link is of significantly higher quality than all the others, as measured by the model-based match probability).

Lastly, some of the LIFE-M linkages are instances of one-to-*many* linking. The three instances of one-to-many linking are: (1) linking siblings together in birth records, (2) linking individuals in birth records to their spouses in marriage records, and (3) linking individuals to their own children in birth records. Naturally, one can have more than one sibling, spouse, or child. For these linkages, we analogously generate sets of potential candidates for primary, but we instead consider each primary-potential pair as a separate linking decision (even though trainers technically saw the entire set of potential candidates and made simultaneous linking decisions). We use a simple random forest model based on pair-level features to generate match probabilities. Similarly, these probabilities are generated using 10-fold cross-validation and links are determined using the threshold associated with a 97% precision rate.

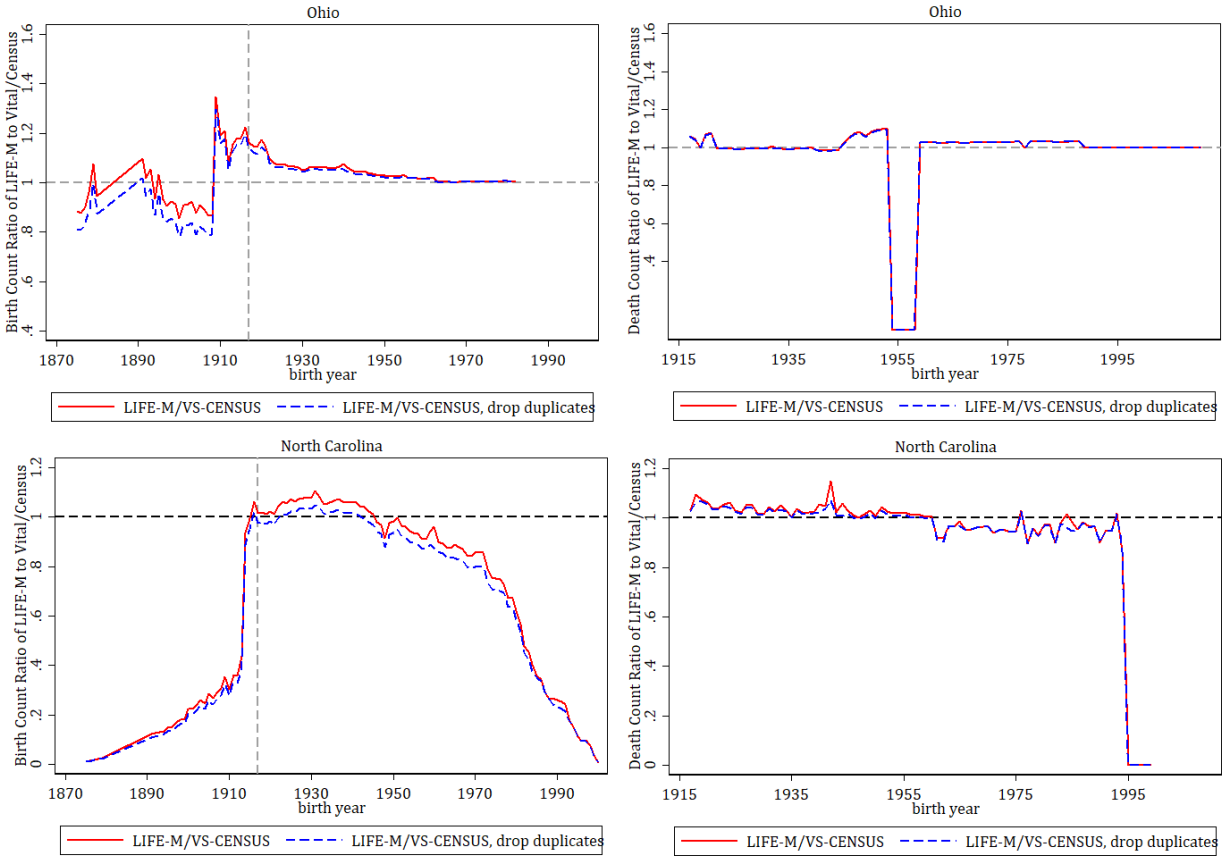
The match rates for hand-links and machine links for version 1 of the data are presented in Appendix Tables D.2-D.7.

---

[15] This can be accomplished using a logistic regression where the dependent variable is an indicator for the set containing a link and the explanatory variables are the three estimated probabilities from the any-match model. The calibrated probability that there exists a link in the set is just the predicted value according to the estimated probabilities.

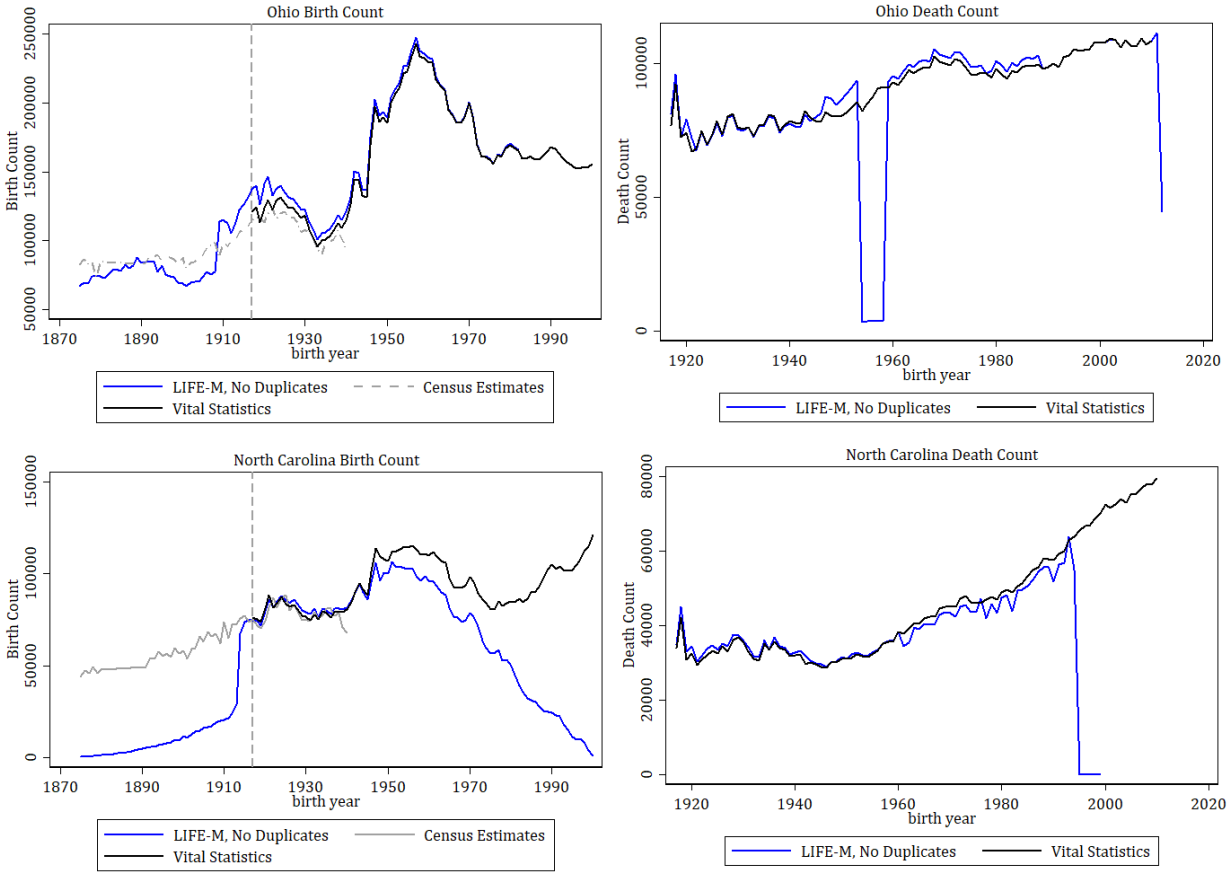
[16] The alternative, linking primaries sequentially and removing potentials that are already linked from the set of potential candidates for subsequent primaries, is computationally prohibitive. Moreover, such a procedure would understate the amount of ambiguity for primaries considered later in the sequence, likely increasing the incidence of false matches.

Appendix Figure A.1 Completeness of LIFE-M Birth and Death Records



Notes: Figures show the ratio of birth and death records in LIFE-M data to the Vital Statistics Reports for the original records (solid red line) as well as with counts after dropping potential duplicates (blue dashed line). The Vital Statistics comes from U.S. County-Level Natality and Mortality Data [ICPSR 36603] (Bailey et al. 2016). For years before 1915, we estimate the birth counts based on age and birthplace information from 1880, 1900, 1910, and 1920 Decennial Census. See appendix figure A.1 for time series of birth and death counts**.**

Appendix Figure A.2 Birth and Death Counts from LIFE-M and Vital Statistics



Notes: Figures show the counts of birth and death records in LIFE-M data (blue line), Vital Statistics Reports (dark grey line), and the 1880-1940 Censuses (light grey line). See also Figure A.1 notes.

Appendix Table C.1. Representativeness of Linked G2 Records in LIFE-M Data

	(1) G2-1940 Census	(2) G2-Death	(3) G2-Marriage
Day of Birth	-0.000246***	-0.0000265***	0.00000970***
	(0.00000190)	(0.00000188)	(0.00000172)
Number of Siblings	0.00758***	0.00918***	0.00677***
	(0.0000953)	(0.0000961)	(0.0000879)
Length of Name	0.0105***	0.00857***	0.00715***
	(0.0000457)	(0.0000459)	(0.0000408)
Length of Father's Name	0.00185***	0.00177***	0.000299***
	(0.0000548)	(0.0000540)	(0.0000482)
Length of Mother's Name	-0.000401***	0.00187***	-0.00000436
	(0.0000442)	(0.0000445)	(0.0000378)
Share of Birth Records with Misspelled Father's Name	-0.00424***	0.00756***	-0.0133***
	(0.00109)	(0.00110)	(0.00101)
Share of Birth Records with Misspelled Mother's Name	0.0525***	0.0342***	0.0585***
	(0.000971)	(0.000970)	(0.000909)
Link in Ohio	0.0862***	0.00381***	0.105***
	(0.000459)	(0.000507)	(0.000395)
Female	-0.0409***	-0.0870***	0.0175***
	(0.000394)	(0.000394)	(0.000362)
Constant	0.00273***	0.0458***	-0.0630***
	(0.000957)	(0.000944)	(0.000824)
Observations	4,300,877	4,300,877	4,300,877
R^2	0.034	0.027	0.029
Wald F Statistics	19507.9	14471.4	19623.9

Notes: Dependent variables are a dummy variable equal to one for G2 birth records linked to specified sources. Regression sample include all G2 birth records in LIFE-M data. Results of regression by gender are available upon request. Reported F-statistics is for a heteroskedasticity-robust Wald-test of joint significance of covariates. Robust standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table C.2. Representativeness of Linked G1 Records in LIFE-M Data

	(1) G1-1940 Census	(2) G1-1920 Census	(3) G1-1910 Census	(4) G1-1900 census	(5) G1-1880 census	(6) G1-Death	(7) G1-Marriage
Number of Children	0.0642***	0.0498***	0.0346***	0.0273***	0.00443***	0.0299***	0.0458***
	(0.000145)	(0.000145)	(0.000131)	(0.000118)	(0.0000553)	(0.000123)	(0.000149)
Length of Name	0.00885***	0.00368***	0.000527***	0.0000916***	-0.000384***	0.00301***	0.00581***
	(0.0000276)	(0.0000258)	(0.0000231)	(0.0000195)	(0.00000956)	(0.0000215)	(0.0000286)
Length of Spouse's Name	0.00871***	0.00340***	0.000718***	0.000940***	-0.000162***	0.00414***	0.00579***
	(0.0000276)	(0.0000258)	(0.0000233)	(0.0000201)	(0.0000103)	(0.0000220)	(0.0000286)
Share of Children's Birth Records with Misspelled Name	0.00294**	0.144***	0.180***	0.103***	0.0328***	0.0944***	0.175***
	(0.00131)	(0.00130)	(0.00122)	(0.00105)	(0.000538)	(0.00108)	(0.00138)
Share of Children's Birth Records with Misspelled Spouse's Name	-0.00769***	0.140***	0.162***	0.0930***	0.0430***	0.0851***	0.176***
	(0.00130)	(0.00129)	(0.00122)	(0.00106)	(0.000558)	(0.00109)	(0.00138)
Link in Ohio	0.0711***	0.0799***	0.0680***	0.0431***	0.00795***	-0.0334***	0.109***
	(0.000386)	(0.000368)	(0.000324)	(0.000274)	(0.000141)	(0.000335)	(0.000411)
Female	0.00354***	0.00584***	-0.00158***	-0.0200***	-0.00971***	-0.0297***	0.000133
	(0.000364)	(0.000342)	(0.000297)	(0.000253)	(0.000118)	(0.000269)	(0.000382)
Constant	-0.164***	-0.0800***	-0.0278***	-0.0156***	0.00793***	-0.0283***	-0.0864***
	(0.000368)	(0.000352)	(0.000311)	(0.000263)	(0.000116)	(0.000274)	(0.000384)
Observations	5528046	5528046	5528046	5528046	5528046	5528046	5528046
R^2	0.207	0.159	0.081	0.071	0.016	0.078	0.182
F	258727.6	149041.9	57725.5	49010.9	8802.8	52393.9	224693.6

Notes: Dependent variables are a dummy variable equal to one for a G1 linked to specified sources. Regression sample include all G1 parents showing up in G2 birth records in LIFE-M machine-linked full data. Results of regression by gender are available upon request. Reported F-statistics is for a heteroskedasticity-robust Wald-test of joint significance of covariates. Robust standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table D.1: Linking Methods Summary by Blocking and Linking Variables

	G0		G1							G2							G3		
Source	G1-1900, G1-1880, G1-Death, G1-Marriage, all as parents		G2 Birth Certificates as parents							G2 Birth Certificates as kids							G3 Birth Certificates, G2-1940, both as kids		
	G0-1940	G0-Death	G1-1940	G1-1920	G1-1910	G1-1900	G1-1880	G1-Death	G1-Marriage	G2-Sibling	G2-1940	G2-Death		G2-Marriage	G2-G3Birth		G3-Death		G3-Marriage
	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH/NC	OH	NC	OH/NC	OH	NC	OH	NC	OH/NC
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
A. Blocking variables																			
last name initial	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
year of birth (YOB)		✓	✓			✓	✓	✓		✓	✓	✓	✓	✓			✓	✓	✓
birth place (BPL)		✓	✓	✓	✓	✓	✓	✓											
mom last name initial												✓					✓		
B. Linking variables																			
date of birth (DOB)												✓	✓				✓	✓	
YOB		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BPL			✓	✓	✓	✓	✓	✓							✓				
first and middle name	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
last name (maiden)						✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓
last name (married)	✓	✓	✓	✓	✓	✓	✓				✓	✓							
dad first and middle name										✓		✓	✓	✓			✓	✓	✓
dad last name										✓		✓	✓	✓			✓	✓	✓
mom first and middle name										✓		✓	✓	✓			✓	✓	✓
mom last name (maiden)										✓		✓	✓	✓			✓	✓	✓
spouse first and middle name	✓	✓	✓	✓	✓	✓		✓	✓						✓	✓			
spouse last name (maiden)	✓	✓						✓	✓	✓	✓	✓			✓	✓			
kid YOB			✓	✓	✓														
kid BPL			✓	✓	✓														
kid first and middle name			✓	✓	✓														
C. Information obtained from other links		YOB and BPL from G0-1940.				YOB and BPL from G1-1910,1920, 1940, G1-Marriage.	YOB and BPL from G1-1900, 1910, 1920, 1940, G1-Marriage.	YOB and BPL from G1-1900, 1910, 1920, 1940 and G1-Marriage.			Spouse last names from G2-Marriage, G2-Death, used as married last name for female.	Spouse last names from G2-Marriage, G2-G3Birth, used as married last name for female				Spouse names from G2-1940, G2-Death, G2-Marriage.			
D. Special Linking Restrictions		Only G0 with nonmissing YOB.				Only G1 with nonmissing YOB.	Only G1 with nonmissing YOB.	Only G1 with nonmissing YOB.											Only G3 with mom last name (maiden) available.

Notes: Linking variables refer to any variables we used to create the set of candidate matches for the hand-linking or machine learning process. Blocking variables refer to variables used for restricting candidates considered during linking. Blocking and linking variables differ across states and linking types due to record availability and completeness. All primaries need to satisfy link specific name completeness condition, but we imposed additional restrictions in some cases as noted in panel D. Often, one or more linking methods can be applied, and we use the union of information used for each link type. The default window for year of birth (YOB) in blocking is +/- 3, except for G2-sibling linking which had a +/- 34 year window.

Appendix Table D2: Number of Hand Links (Link Rates) for G1 in the LIFE-M Data

	1940 Census	1920 Census	1910 Census	1900 Census	1880 Census	Death	Marriage	G0 Parents
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Ohio				
All	37.6 13,938/37,069	51.8 19,196/37,069	31.2 11,575/37,069	35.8 9,953/27,806	16.8 1,656/9,856	21.6 6,246/28,961	41.3 15,304/37,069	53 19,632/37,069
Male	36.7 6,755/18,429	51.5 9,491/18,429	30.9 5,692/18,429	36.6 5,135/14,029	20.1 1,192/5,928	25.4 3,665/14,405	41 7,563/18,429	53.8 9,920/18,429
Female	38.5 7,183/18,640	52.1 9,705/18,640	31.6 5,883/18,640	35 4,818/13,777	11.8 464/3,928	17.7 2,581/14,556	41.5 7,741/18,640	52.1 9,712/18,640
				Panel B: North Carolina				
All	29.7 14,189/47,806	44.4 21,217/47,806	17.9 8,546/47,806	31 10,755/34,739	14.7 1,451/9,880	38 13,667/35,938	39.9 19,052/47,806	49.9 23,866/47,806
Male	29.2 6,969/23,846	43.8 10,450/23,846	17.3 4,119/23,846	32.1 5,602/17,472	17.8 1,142/6,427	37.5 6,705/17,870	39.7 9,469/23,846	49.2 11,733/23,846
Female	30.1 7,220/23,960	44.9 10,767/23,960	18.5 4,427/23,960	29.8 5,153/17,267	8.9 309/3,453	38.5 6,962/18,068	40 9,583/23,960	50.6 12,133/23,960

Notes: Link rates are calculated as number of links divided by the number of linkable people in the G1 hand-link sample. Linkable people for column 1 are G2s who were born in the core years (1900-1929) and have non-missing parents' names (father's first name, last name, and mother's first name). Linkable people for column 2 are G2s who were born by 1940 and have non-missing names (both first and last names). Linkable people for column 3-5 are G2s who have non-missing names. The number of links and linkable people for the males and females may not sum up to the total number because of some G2 with unknown sex (not displayed for brevity).

Appendix Table D3: Number of Hand Links (Link Rates) for G3 in the LIFE-M Data

	1940 Census	Death	Marriage
	(1)	(2)	(3)
	Panel A: Ohio		
All	18	13.4	4.3
	9,039/50,225	6,532/48,568	2,106/48,568
Male	20.9	16.6	4.6
	4,809/23,058	3,820/23,036	1,069/23,036
Female	20.7	11.2	5.1
	4,230/20,483	2,296/20,467	1,037/20,467
	Panel B: North Carolina		
All	23.2	7	3.5
	10,251/44,217	2,915/41,595	1,450/41,595
Male	25.2	9.5	3.8
	5,349/21,246	1,923/20,312	781/20,312
Female	24.3	5.1	3.5
	4,902/20,157	992/19,283	669/19,283

Notes: Link rates are calculated as number of links divided by the number of linkable people. The linkable people for G3-death and G3-marriage are the G3 who have non-missing first and last names. The link rates of G3-1940 are calculated as the share of all G3 that are linked to a 1940 Census record. (We did not link G3 to the 1940 Census directly, the links come from associated G2-1940 Census links). The link rates of G3-death and G3-marriage are lower than other generations because the death and marriage records used for linking is becomes less complete over time (Ohio marriage records were significantly incomplete after 1950s and death records were significantly incomplete after 2005. North Carolina marriage records were significantly incomplete after 1960s and death records were incomplete after 1995). The number of links and linkable people for the male and the female may not sum to the reported total number because some G3 have unknown sex and are not reported (for space saving).

Appendix Table D4: Number of Hand Links (Conditional Link Rates) for G0 in the LIFE-M Data

	1940 Census (1)	1920 Census (2)	1910 Census (3)	1900 Census (4)	1880 Census (5)	Death (6)
			Panel A: Ohio			
All	19.2 3,576/18,661	10.8 4,056/37,639	14.9 5,613/37,639	34.6 13,013/37,639	9.6 3,607/37,639	10 1,863/18,661
Male	18.6 1,941/10,433	10.4 1,937/18,703	14.8 2,764/18,703	34.2 6,396/18,703	10.1 1,880/18,703	9.7 1,012/10,433
Female	19.9 1,635/8,228	11.2 2,119/18,936	15 2,849/18,936	34.9 6,617/18,936	9.1 1,727/18,936	10.3 851/8,228
			Panel B: North Carolina			
All	9.3 1,674/18,006	7.4 3,358/45,114	10.1 4,544/45,114	32.9 14,835/45,114	6.4 2,907/45,114	6.3 1,143/18,006
Male	9.7 1,038/10,754	6.6 1,495/22,576	9.7 2,188/22,576	31.9 7,194/22,576	6.5 1,478/22,576	6.7 721/10,754
Female	8.8 636/7,252	8.3 1,863/22,538	10.5 2,356/22,538	33.9 7,641/22,538	6.3 1,429/22,538	5.8 422/7,252

Notes: Link rates are calculated as number of links divided by the number of linkable people for G0 in hand-linked sample. The hand-linked sample of G0 include grandparents of hand-linked G2 sample. The linkable people for G0-1940 Census are the G0 who were born by 1940 and have non-missing first and last names. The link rates for G0-1920 Census, G0-1910 Census, G0-1900 Census, and G0-1880 Census are calculated as the share of all G0 that are linked to each census (we don't link G0 to these censuses directly and the links come from associated G1 census links). The linkable people for G0-death are the G0 who have non-missing names and known birth year (birth year comes from the linked census records). G0-death link rates are lower than other generations because the death records used for linking are not complete before certain years (death records in both states became complete after 1910s).

Appendix Table D5: Number of All Links (Link Rates) for G1 in the LIFE-M Data

	1940 Census (1)	1920 Census (2)	1910 Census (3)	1900 Census (4)	1880 Census (5)	Death (6)	Marriage (7)	G0 Parents (8)
				Panel A: Ohio				
All	30 1,031,202/3,441,255	25.3 870,727/3,441,255	18.2 626,022/3,441,255	24.7 415,269/1,679,414	15.1 80,679/535,997	21.3 413,719/1,940,603	33.5 1,153,043/3,441,255	38 1,308,471/3,441,255
Male	29.7 511,361/1,720,381	24.9 429,105/1,720,381	18.2 312,998/1,720,381	25.9 228,140/879,302	16 50,669/316,145	24.9 242,489/972,088	33.5 576,471/1,720,381	39 670,318/1,720,381
Female	30.2 519,841/1,720,874	25.7 441,622/1,720,874	18.2 313,024/1,720,874	23.4 187,129/800,112	13.7 30,010/219,852	17.7 171,230/968,515	33.5 576,572/1,720,874	37.1 638,153/1,720,874
				Panel B: North Carolina				
All	20.6 167,597/815,515	16.9 137,988/815,515	11.5 94,075/815,515	19.7 62,780/319,411	15.1 15,723/104,132	36.7 141,292/384,801	23.5 191,472/815,515	27.9 227,241/815,515
Male	20.3 82,914/407,753	16.7 67,947/407,753	11.8 47,930/407,753	20.9 35,328/169,052	16.6 10,666/64,447	36.4 70,220/192,998	23.5 95,734/407,753	28 114,310/407,753
Female	20.8 84,683/407,762	17.2 70,041/407,762	11.3 46,145/407,762	18.3 27,452/150,359	12.7 5,057/39,685	37.1 71,072/191,803	23.5 95,738/407,762	27.7 112,931/407,762

Notes: Link rates are calculated as number of links divided by the number of linkable people. The linkable people for G1-1940 Census, G1-1920 Census, and G1-1910 Census are the G1 who belong to a linkable couple (a linkable couple must have non-missing husband's first name, last name and wife's first name). The linkable people for G1-1900 and G1-1880 are the G1 who belong to a linkable couple and were born by 1900 and 1880, respectively. The linkable people for G1-death are the G1s who belong to a linkable couple and have known birth year (birth year comes from G1 Census links). The linkable people for G1-marriage are the G1 who belong to a linkable couple; For G1-parents, the link rates are calculated as the share of G1 (who belong to a linkable couple) that have at least one parent in the dataset. G1's parents (G0) come from G1 Census links, death link, or marriage link.

Appendix Table D6: Number of All Links (Link Rates) for G3 in the LIFE-M Data

	1940 Census	Death	Marriage
	(1)	(2)	(3)
	Panel A: Ohio		
All	17.4 337,396/1,936,221	9.5 180,906/1,895,576	5.5 103,323/1,895,576
Male	19.4 173,868/898,022	11.6 104,288/897,399	5.2 47,068/897,399
Female	20.3 163,528/803,773	8.3 66,711/803,468	7 56,255/803,468
	Panel B: North Carolina		
All	16.7 76,927/460,540	5.9 25,610/437,592	3.3 14,450/437,592
Male	18 39,548/220,312	8.1 17,100/212,031	3.1 6,537/212,031
Female	17.8 37,379/209,924	4.2 8,510/202,408	3.9 7,913/202,408

Notes: Link rates are calculated as number of links divided by the number of linkable people. The linkable people for G3-death and G3-marriage are the G3 who have non-missing first and last names. The link rates of G3-1940 are calculated as the share of all G3 that are linked to a 1940 Census record. (We did not link G3 to the 1940 Census directly, the links come from associated G2-1940 Census links). The link rates of G3-death and G3-marriage are lower than other generations because the death and marriage records used for linking is decreasingly complete after certain years (Ohio marriage records were significantly incomplete after 1950s and death records were significantly incomplete after 2005. North Carolina marriage records were significantly incomplete after 1960s and death records were incomplete after 1995). The number of links and linkable people for the male and the female may not sum to the reported total number because some G3 have unknown sex and are not reported (for brevity).

Appendix Table D7: Number of All Links (Link Rates) for G0 in the LIFE-M Data

	1940 Census (1)	1920 Census (2)	1910 Census (3)	1900 Census (4)	1880 Census (5)	Death (6)
			Panel A: Ohio			
All	19.4 3,614/18,661	8.1 180,531/2,222,523	15.6 345,857/2,222,523	18.9 420,600/2,222,523	7.1 157,171/2,222,523	31.2 206,222/661,924
Male	18.9 1,968/10,433	7.7 84,542/1,093,190	15.4 167,912/1,093,190	18.9 206,107/1,093,190	7.3 79,954/1,093,190	30.9 119,657/387,617
Female	20 1,646/8,228	8.5 95,989/1,129,333	15.8 177,945/1,129,333	19 214,493/1,129,333	6.8 77,217/1,129,333	31.6 86,565/274,307
			Panel B: North Carolina			
All	9.8 1,764/18,006	6.1 24,423/398,597	12.3 48,944/398,597	16.9 67,173/398,597	7.1 28,448/398,597	22 22,939/104,485
Male	10.1 1,084/10,754	5.5 10,895/199,629	11.7 23,309/199,629	16.2 32,413/199,629	7.1 14,184/199,629	21.7 13,841/63,896
Female	9.4 680/7,252	6.8 13,528/198,968	12.9 25,635/198,968	17.5 34,760/198,968	7.2 14,264/198,968	22.4 9,098/40,589

Notes: Link rates are calculated as number of links divided by the number of linkable people. The linkable people for G0-1940 Census are the G0 who were born by 1940 and have non-missing first and last names. All the G0-1940 Census links are made by trainers. The link rates for G0-1920 Census, G0-1910 Census, G0-1900 Census, and G0-1880 Census are calculated as the share of all G0 that are linked to the corresponding census (we don't link G0 to these censuses directly and the links comes from associated G1 census links). The linkable people for G0-death are the G0 who have non-missing names and known birth year (birth year comes from the linked census records). G0-death link rates are lower than other generations because the death records used for linking are not complete before certain years (death records in both states became complete after 1910s).