Title

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

Journal
Historical Methods: A Journal of Quantitative and Interdisciplinary History, 56(3)

ISSN
0161-5440

Authors
Bailey, Martha
Lin, Peter
Mohammed, A
et al.

Publication Date
2023

DOI
10.1080/01615440.2023.2239699

Peer reviewed

Published in final edited form as:
Hist Methods. 2023 ; 56(3): 138–159. doi:10.1080/01615440.2023.2239699.

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

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

Abstract

This paper describes the creation of the Longitudinal, Intergenerational Family Electronic Micro-Database (LIFE-M), a new data resource linking vital records and decennial censuses for millions of individuals and families living in the late 19th and 20th centuries in the United States. 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. We describe (1) the data sources, coverage, and linking sequence; (2) the process and supervised machine-learning methods to linking records longitudinally and across generations; and (3) the resulting linked samples, including linking rates, representativeness, and weights.

Keywords

vital records; historical record linkage; supervised machine-learning linking

I. Introduction

Some of the most important questions in social science, demography, and health relate to how individuals' lives and experiences change *over time*. For example, what are the determinants of social mobility and what forces have changed social mobility over time and for whom? How have health and aging today been shaped by a multitude of events, environmental factors, and policies occurring earlier in life? However, U.S. microdata spanning the late 19th and 20th centuries tend to be cross-sections—large sets of individuals *at one point in time*. Cross-sectional data limit the study of life-cycle and intergenerational processes, restricting research on many fundamental determinants of human well-being.

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

The Longitudinal, Intergenerational Family Electronic Micro-Database (LIFE-M) is a new historical, longitudinal and intergenerational panel for the United States, incorporating digitized vital records to follow millions of individuals from birth to death, integrating health, demographic, and family outcomes with census data, and improving the linking of historical records. Vital records include birth, marriage, and death records, which allow LIFE-M to add more detailed demographic and health information to socio-economic variables in censuses. Additional features of vital records 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 certain groups. For instance, birth and marriage records contain women's birth ("maiden") and married names, allowing LIFE-M to track the lives of women who change their surnames at marriage. Vital records 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 records allows LIFE-M to trace socio-economic, demographic, and health outcomes across the life course and intergenerationally for millions of individuals and large networks of three and four-generation families.

This paper provides a detailed overview of how we implemented automated linking at scale to create the LIFE-M database. We discuss (1) the data sources, data coverage, and linking process; (2) the creation of high-quality hand-links for training purposes; (3) supervised machine-learning algorithms to link records longitudinally and across generations; (4) the characteristics of the first data release, including link rates, representativeness, and weights; and (5) opportunities for research using the data.

II. Existing Data Resources and LIFE-M's Contributions to Data Infrastructure

Creating a longitudinal and intergenerational panel from U.S. historical data for life-course or intergenerational analyses comes with important challenges. Because there are no available numerical identifiers in most records, such as the census, individuals need to be linked across time using other information. However, digitized information for many historical records is limited in terms of the consistently available fields (in the census, name, age, and state of birth) and the accuracy of its reporting. Digitizing records also introduces additional measurement error into this limited information setting. Many records are hand-written, which is not always legible, and people transcribing the script make data entry errors. In addition, the available information changes over time. For example, women change their surnames at marriage. Several new data linking projects have begun to address these challenges and have created historical panel data to study 20th century population dynamics in the United States.[1] In this section, we review the content and construction

[1] Historical linking to create longitudinal data has been done extensively outside the United States. For the United Kingdom (U.K.), the Cambridge Group for the History of Population and Social Structure hosts several different datasets, primarily representing different areas and time periods for England. The most widely cited database is the family reconstitution data for 26 English parishes (Wrigley et al. 2018), which has been used to conduct individual-level studies of fertility, mortality and nuptiality. The Victorian Panel Study links vital and census records from 1851-1901 in Great Britain (Schürer 2007). Data from Sweden are linked longitudinally from 1650 to 1950 across full-count censuses, emigration records and death records (Wisselgren et al. 2014, Berger and Eriksson forthcoming). Available through the Swedpop project, these links have been used in multiple studies and include women as well as men (Dribe, Eriksson, and Scalone 2019). The Scanian Economic Demographic Database covers the entire population of five rural

of these datasets and conclude with a discussion of LIFE-M's unique contribution to data infrastructure.

A. Existing Longitudinal and Intergenerational Micro-Data Resources for the United States

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 mid-19th century (Wimmer 2003). The data consist of 39,340 Union Army (UA) soldiers, approximately 6,200 of whom were "Colored Troops," linked to rich information on disability, health, use of medical care, and pension receipt. Through links to the 1850 and 1930 Censuses, the UA data also include socio-demographic and economic variables. A limitation of the UA data is that they consist primarily of men, most of whom were Northern born.

The Minnesota Population Center (MPC) has also led efforts to digitize and integrate historical census data. The Integrated Public Use Microdata Series Linked Representative Samples (LRS) connects the 1880 Census to the 1850-1930 Censuses (Ruggles et al. 2010).[2] 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 the lack of women (who cannot be followed longitudinally because their last names change at marriage), sparse longitudinal coverage (typically two points in time for any single person), and intergenerational coverage consisting primarily of 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 1850 to 1940 (Helgertz et al. 2020). The resulting sample sizes range from around 6 million individuals linked between the 1850 and 1860 Censuses and 52 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. In the second step, MLP tries to link people living in the same household as the men linked in the first step. This second step helps link some women who were living with their spouses or daughters living with their fathers in both censuses. However, census data limitations make it nearly impossible to link women who change households or their names at marriage.

and semi-urban parishes and an industrializing town in southern Sweden between 1815-2017 (Dribe and Quaranta 2020, Bengtsson and Dribe 2021). For Canada, the BALSAC database covers the population from an area in Quebec from 1621 to 1965 (Vézina and Bournival 2020). The Canadian Peoples contains over 40 million linked census records, representing three generations from the mid-19th century to early 20th century (Foxcroft, Inwood, and Antonie 2022). Researchers have also linked records for the U.K. (Long and Ferrie 2013) and Norway (Modalsli 2017, 2021) as well as between Norway and the U.S. (Abramitzky, Boustan, and Eriksson 2013, Biavaschi and Elsner 2013). We focus our discussion on U.S. databases which are most related to LIFE-M.

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

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 (the 1890 Census is omitted because the population schedules were destroyed in a fire). Building on the linking approaches in Abramitzky, Boustan, and Eriksson (2012) and Abramitzky, Mill, and Pérez (2020), CLP relies on rule-based and unsupervised linking methods, which link records based on name, age, race and time/place of birth information. (The difference between supervised and unsupervised 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) rule-based linking methods to link the 1940 Census to the Social Security Death Master File and the Social Security Administration's Numerical Identification Files (Numident) (Goldstein et al. 2021). Similar to Bailey, Mohammed, and Mohnen (2022), this project links women by following women's name changes in the Social Security records.[3] Due to surname changes, CenSoc only links women to the Numident (limiting this sample to women dying between 1988 and 2005).

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), the 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.[4] 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. Individuals from the earliest cohort (born in 1924) needed to survive to age 74 to make it into the survey, so HRS samples miss many individuals 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;

[3] Mohammed and Mohnen (2023) also use a subset of the linked dataset used in Bailey, Mohammed, and Mohnen (2022) to study the impact of Rosenwald schools on labor market outcomes for both men and women.

[4] 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).

15% attrition in the HRS as of 2004) and limited geographic coverage, which constrain the representativeness of these data at the state and more local (county/town) levels.[5]

In summary, available data limit 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 most women cannot 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 and geographic coverage.

B. LIFE-M's Contribution to Data Infrastructure

The first version of the LIFE-M dataset was released through Inter-University Consortium for Political and Social Research (ICPSR) in 2022 along with extensive documentation, summary statistics, variable descriptions, and user guides (Bailey et al. 2022b). This release contains 15 million individuals born from 1841 to 1968 belonging to over 4 million families, including high-quality links for 6.9 million women and half a million underrepresented minorities.[6] LIFE-M makes five contributions to data infrastructure.

Contribution 1: Large Samples of Women and Groups Underrepresented in Linked Censuses: Using vital records allows women to be linked in unprecedented numbers. Birth certificates contain information on the birth (or "maiden") names of each child's mother as well as the father's surname, allowing LIFE-M to link mothers to their own birth families and follow them in their married families. While less complete than birth certificates, marriage certificates supplement this information with the birth names of the bride.

Using vital records also increases LIFE-M's sample sizes for understudied groups, such as racial minorities and immigrants. Vital records contain high-quality data on names, dates of birth, and birthplaces as well as parents' names and birth counties in many instances. 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).[7] This objective 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

[5] 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 ages 50 and 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, 2015, 2022, Bleakley and Ferrie 2013, 2014, 2016). 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.

[6] LIFE-M links more than 170,000 Black Americans and more than 368,000 foreign-born people.

[7] Age misreporting is common in the Census (e.g., there are a lot more 50- and 60-year-olds relative to 51- and 63-year-olds) as well as on marriage certificates to circumvent minimum age requirements (Blank, Charles, and Sallee 2009). Age misreporting is more common for Black Americans (Elo and Preston 1994, Logan and Parman 2011).

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 example, for cohorts born between 1900 and 1930 in the LIFE-M data, 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 in vital data also 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.

In addition, rich variables in vital records permit LIFE-M to reduce multiple matches (even for common names) and identify the correct match.[8] First, vital records are official documents and contain full names (first, middle, and last). Middle names are especially helpful in distinguishing between similar names. Second, birth, marriage, and death records contain rich information on the exact date (not just age), county of birth, and parents' full names and birthplaces. These variables allow better linkages to censuses and other records so that researchers can study the long-run effects of parental characteristics and early childhood circumstances. Finally, death records allow researchers to understand why individuals do not have matches in census or marriage records.

Contribution 2: Four Generations of Intergenerational Coverage: LIFE-M is the first large-scale, longitudinal database for the United States to link networks of families in the 19th and 20th century—including large samples of both men and women—across four generations. LIFE-M cohorts span the historical eras of reconstruction and rapid industrialization, mass migration and urbanization, the expansion of public health and hospital infrastructure and policy, the Great Depression, and two World Wars. 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 large-sample intergenerational analyses of entire families. LIFE-M reconstitutes birth and marriage families[9] for the late 19th and early 20th century birth cohorts, which uniquely allows analyses of intergenerational relationships for men *and* women (mothers, sisters, and grandmothers) as well as how their outcomes relate to those of their ancestors, siblings, and offspring.

Figure 1 orients these generations in time by providing approximate years of birth for each "generation," as well as their empirical frequency. As we describe in Section III.C, LIFE-M draws its baseline sample from birth certificates (G2s, individuals born in the early 20th century) and then links them forward to their children (G3s) and backwards to their parents (G1s) and grandparents (G0s). Generations overlap because earlier-born G2s may have the same birth year as the parents (G1s) of some later-born G2s. The same logic applies to other generations as well.

[8] Multiple matches have been so problematic that past work has eliminated common names entirely from samples to be linked (Ferrie 1996, Ruggles 2006).

[9] We use the terms "birth family" and "marriage family" to distinguish between when someone is a child (birth family) and when they are married or a parent (married family).

Contribution 3: Expanded Longitudinal Coverage: LIFE-M tries to link all individuals' birth certificates to their marriage and death records and future censuses. Looking for all individuals born allows the database to capture many individuals who died prior to marriage or the census (especially infants and children) or who emigrated from the United States. By allowing researchers to merge in information about local characteristics (e.g., policies, environmental circumstances, the strength of the economy), information on county of birth also facilitates the study of the role of early-life local circumstances in determining socio-economic and health outcomes, such as longevity or cause of death. Finally, links to marriage records, censuses, and death records allow researchers to observe critical life transitions.

Contribution 4. Integrated Demographic, Family, Economic, and Health Information: LIFE-M combines socio-demographic, economic, and family-network information with longevity and cause of death. Isolated linked census samples or linked vital records contain only snapshots of these outcomes, but LIFE-M integrates health, economic, family, and demographic data into a single large-scale, longitudinal, and intergenerational database. Variables that are available or can be constructed/merged with 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,[10] and birth state or country from the censuses);
- own economic and demographic outcomes (e.g., wages, employment, occupation, birth state or country, literacy and education[10]);
- health (date and place of death, longevity in days, and cause of death for Ohio);
- marriage family characteristics (e.g., age at marriage, married name, spouse name and all characteristics above);
- own births (e.g., number of children, mortality of own infants and children, timing of births, sex composition, and twinning); and
- geographic location (county or state) at vital events and census enumeration.

These LIFE-M data can be linked to other databases in several ways. Individuals are traced using a 5-digit alphanumeric identifier (LIFEMID). To facilitate linking LIFE-M to the full-count census data, census record identifiers (HISTID) are contained within the LIFE-M data, and IPUMS provides a variable, LNKLIFEM (https://usa.ipums.org/usa-action/variables/LNKLIFEM), that indicates if a census record is in the LIFE-M database.

In addition, information on geographic location across time presents a multitude of opportunities 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

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

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 also provided in the LIFE-M data, which contains individuals' county and state of residence at the time of birth; in the 1880, 1900, 1910, 1920, 1940 Censuses; at the time of marriage; and at death. Of the 543,965 individuals with 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.

Contribution 5: Public Availability of Integrated Data with Large Sample Sizes: The release of LIFE-M through ICPSR ensures the longevity of the project and systematic version control. Access to these data is available for any researcher with access to ICPSR (which is free of charge). In addition, we released data documentation, a user guide (including instructions for loading and using the database in Stata and R), variable descriptions, and summary statistics. Analyzing health data in LIFE-M does not require access to restricted census data. Public availability speeds progress on important research questions.

III. IFE-M's Data Sources and Linking Sequence

The first release of LIFE-M includes links of multiple record sources, including two states' birth, death, and marriage records, and the 1880, 1900-1920, and 1940 Censuses.[11] We plan to incorporate more vital and non-vital records into LIFE-M in the future. In this section, we describe these data sources, including our choice of states and coverage, then conclude with an overview of the linking sequence.

A. Data Sources

LIFE-M uses newly digitized state vital records from FamilySearch.org, a nonprofit genealogical website which has digitized tens of millions of handwritten records and made them publicly available. Not all state vital records are complete, digitized, or publicly available. Differences in availability reflect state legal restrictions today, the timing of when states entered the vital registration area, as well as the work of organizations like FamilySearch and Ancestry to digitize these records. After examining quality and coverage for multiple sets of state records, we chose to begin with vital records in Ohio and North Carolina for several reasons. These two states are home to different demographic groups and were at different stages of economic development in the early 20th century. For instance, Ohio was more industrialized and attracted numerous European immigrants, while North Carolina was more rural and agricultural with a larger population of Black Americans. In addition, these states have near complete coverage of vital records for our periods of interest. Third, we wanted to maximize sample sizes given the fixed costs of processing and cleaning different state records. Ohio and North Carolina were the 4th and 11th most populated states in 1940 and represented approximately 8% of the U.S. population. One general limitation

[11] These refer to the full-count Censuses for the entire United States.

of relying on these states' vital records is that they omit vital events that occur outside of Ohio and North Carolina or in time periods not covered. Census records help overcome this limitation by providing additional family members observed outside of the states' vital records regardless of residence.

B. Data Coverage

To characterize the completeness of vital records used in LIFE-M, we compare them to published tabulations of births and deaths after Ohio and North Carolina enter the Federal Registration Area. In the years before vital tabulations are available, we estimate birth counts based on population counts in the first decennial census following a cohort's birth (e.g., the birth count for the 1901 cohort is the number of 9-year-olds in the 1910 Census, for the 1902 cohort it is the number of 8-year-olds in the 1910 Census, and so on). Figure 2 plots the total count of birth and death records in LIFE-M and the vital records over time. The overlap between the LIFE-M (in blue) and vital records (in black) confirms the completeness of the birth certificate microdata from around 1910-1950 and the completeness of the death certificate microdata from around 1918-1990 (except in the late 1950s for Ohio where a few years of death records are not yet available). In Ohio, the coverage of birth records hovered above published vital statistics beginning in the early 1900s, and the coverage of death records followed published vital statistics very closely beginning in 1915. In North Carolina, the coverage of birth records reached completeness around 1915 and gradually declined after 1950. The death records in North Carolina have been close to complete since 1915. Consequently, LIFE-M birth and death records capture nearly the universe of births and deaths for a large share of the 20th century. U.S. marriage records are known to be incomplete, so we omit these plots for brevity (Kennedy and Ruggles 2014, Ruggles 2016).

C. Linking Sequence

Figure 3 describes the records combined by LIFE-M and the information obtained from each source. The first step in the linking process reconstitutes birth families of the late 19th and early 20th century using birth records, which provide infants' and parents' names as well as the dates and places of birth. To reconstitute birth families, we used the universe of birth records in Ohio and North Carolina from 1900-1929 as baseline samples. Then, we linked these records to all their siblings using parents' full names listed on the birth records. All individuals born 1900-1929 and their siblings form G2.

The second step is linking G2 birth records to death, marriage, and the 1940 Census records (restricted to people born in North Carolina and Ohio). In addition, we link G2s to the universe of birth records to find their own children (G3s). This step differs from the sibling linking in the first step, because we link G2s to the mother or father on G3 birth records. We supplement these links with information on G2s' children from the linked 1940 Census, because many G2s were married and resided with their children. The 1940 Census also provides information such as educational attainment, wages, and employment outcomes for most G2s and their spouses in adulthood.

The third step is to link G3s (birth records or 1940 Census records) to their own death and marriage records. Information on many G3s in childhood comes from the 1940 Census.

The fourth step is to link G1s, listed as parents on G2s' birth certificates, to their own death and marriage records as well as to the 1880, 1900, 1910, 1920, and 1940 Censuses. Census links provide key information on birthplace, age, and race for G1s, which allows us to link them to other sources. We obtain their parents' (G0s) information from all of these linked records, whenever available. Earlier census records allow for the addition of G1s' birth family conditions, including ancestry/heritage and their race, location, and parents' occupations.

The final step is to link G0s to their death and 1940 Census records. We were not able to link G0s to marriage records due to the lack of records for these cohorts. However, we can still construct marriage families for G0s when spouse information is present in the censuses or through the G1 links.

IV. IFE-M's Methodology for Linking Historical Data at Scale

LIFE-M generated these links in a series of steps: data cleaning; the creation of candidate links; the creation of high-quality, hand-linked training samples; and finally, the development of customized, supervised learning models to automate linking at scale.

A. Data Cleaning

The records were cleaned and standardized before linking. The purpose of this data cleaning is to increase true match rates by identifying and correcting potential spelling and digitization errors before attempting to link records. For example, we cleaned common name abbreviations (e.g., "Wm." to "William"), standardized names (e.g., "Le Roy" to "LeRoy"), removed suffixes and prefixes (e.g., "Colonel", "General", "Major", etc.), and corrected the location of errant information (e.g., mothers' information appearing in the field for the father). Then, we parsed the single name string into substrings for first, middle, married surname, and birth name, which only differed from the surname for mothers. We also cleaned and standardized dates and geographic codes when possible. Appendix A provides more details about the data cleaning process.

B. Generating Sets of Candidate Links

We next generated sets of candidate links. To minimize computational burden, we "blocked" on certain characteristics, such as the first letter of the last name, place of birth, and year of birth. Blocking requires all candidate links to agree with the primary record in the blocked characteristic. For example, blocking on the first letter of the last name requires that the candidate links share the first letter of the last name with the primary links. Then, we ranked candidate links within the block according to their name similarity with the primary record and the candidates with the highest similarity scores. The blocking criteria and string similarity scores varied across the types of linking. For example, records were linked as individuals (e.g., G2-to-1940 linking), as couples (e.g., G1-to-marriage linking), and as families (e.g., G1-to-1920 linking). Appendix B provides further details regarding how we generated sets of candidate links, and Appendix Table A.1 summarizes our blocking and ranking variables, from which it is clear whether records are linked as individuals, couples, or families.

C. Creation of Hand-Links as Training Data

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 machine models. Consequently, the lack of ground truth for historical data has been a central challenge to using these techniques. 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. To create the hand-linked data, we recruited over 50 data trainers to review and match people across records. We ensured data quality by developing a comprehensive training program and a semi-automated data distribution system.

Data Trainers and Process Management—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 trainers. Orientation entailed over 30 hours of practice reviewing and linking historical records, where trainers received detailed feedback about their linking decisions. To join the LIFE-M hand-linking team, prospective trainers’ decisions had to match a highly vetted and carefully chosen truth dataset 95% of the time. The process was designed to develop the trainers’ knowledge and decision-making in a variety of linking contexts.

We also developed a semi-automated system to manage the hand-linking process. The system automated the distribution of data to individual trainers, monitored trainers’ speed and accuracy, and used a streamlined interface to minimize distractions. The system allowed trainers to log in from any computer, randomized the distribution of batches (which were delivered in small enough increments that trainers could finish within 15 to 20 minutes), and automatically uploaded the trained data to a central repository after completion. This system minimized concerns with data transfers and loss and eliminated the need for a team member 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.

To maintain quality and minimize costs, the system provided data trainers with weekly feedback on their accuracy.[12] Accuracy was determined through the distribution of “audit batches” as part of the batch distribution system. Audit batches were carefully selected to represent the records being trained and appeared identical to other batches, so that trainers could not distinguish between the audit batch and their training work. The system then

[12] The project also tracked and provided trainers with feedback on their speed, which was determined using the metadata collected from time-stamped uploads and downloads of each batch from the distribution system. 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.

computed the errors made on audit batches and provided feedback to the trainers at the weekly meeting. In addition, commonly occurring errors on the audit batches were the topics of weekly, in-person trainer discussions. Weekly meetings also featured discussions of historical or contextual factors affecting the quality of the records, common linking mistakes, and disagreements over how to code difficult cases. The resulting hand-linked data are not error free, but thoughtful discussion and audit batches helped maintain a high-quality linking process.

Displaying Sets of Candidates—LIFE-M displayed sets of candidate links to trainers. Table 1 shows two examples of what trainers observed when making a decision. Displaying multiple candidate links allows trainers to infer the frequency of name 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 set of candidate links does not instill confidence that the exact, unique match is correct.[13] When trainers see three additional close links (especially Candidate 3, "Jason O'Sullivon, born in Ohio, age 35"), they tend to use this set-level information to reject the top candidate link–even if it is an exact match–knowing that a handwritten "a" in the last name may easily be mistaken for an "o" during the digitization process. In this case, showing a set of candidate matches can *decrease* trainer certainty and reduce confidence (relative to showing only one candidate match at a time)—even when the link is an exact unique match.

Another reason for displaying sets of information is that trainers can make links that automated 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 low number of similar candidates tends to increase trainer confidence in links with slightly lower match scores. 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 within a similar age range. (Candidate #1 might be Shelagh's older sister). Given that Shelagh/Sheilagh and Ogilvie/Oglvie may easily be transcription errors, a trainer may feel more confident and choose candidate #5 as the link. In this case, displaying a set of candidate matches can help *increase* certainty about a match that differs from the primary in several dimensions.

In short, methods comparing two records at a time ignore information that can decrease (Table 1A) or increase certainty (Table 1B) about a match being correct, whereas methods considering the set of possible candidates take this additional information into account. The true link is not always an exact, unique match on name (or phonetic name) within an age band, nor is an exact, unique match necessarily a correct link when measurement error skews the number of close candidates.

[13] 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).

Multiple Reviews and Final Linking Decisions—As the examples in Table 1 demonstrate, making correct links can be challenging, and even experienced data trainers make errors. To minimize such errors in the hand-linked data, each primary record was independently evaluated by two different, randomly assigned data trainers. If the two initial trainers reached an agreement (decisions are unanimous, 2-0), we coded their choice, either link or no link, as the final decision. If the two trainers disagreed, the case was sent to three new, randomly assigned trainers for independent re-examination. If the three additional trainers reached an agreement (decisions are split, 4-1), 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 did not agree (decisions are split, 2-3), we code “no link” as the final decision. These 2-3 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. Columns 2 and 3 of Table 2 show the share of records where trainers agreed. In Ohio, agreements occurred around 92% of the time and 91% in North Carolina. Ambiguous cases (2-3 split decisions) ranged from 0.21% to 6.68% in Ohio and from 0.86% to 8.14% in North Carolina. Lastly, some links made by trainers were later overturned due to “conflicts” that arose when comparing different links to each other, either within the same type of link (e.g., cases in which the same record in the target dataset is linked to two different records in the origin dataset) or across types of links (e.g., cases in which a birth record is linked to a person in the 1940 Census but is also linked to a death record prior to 1940). The incidence of overturned links ranges from 0.13% to 5.59% in Ohio and from 0.39% to 6.04% in North Carolina.

Hand-Linked Samples and Quality Evaluation—Table 3 summarizes the hand-linked G2 samples resulting from this systematic and careful process. Hand links to the 1940 Census include 25,727 men (10,977 from Ohio, 14,750 from North Carolina) for a match rate of 42 to 45%, depending on the state. While these rates are low by the standards of modern administrative data, they are considerably higher than existing studies using supervised or unsupervised learning—and especially high given the very low rate of linking error. These link rates do not account for mortality or emigration—doing so would raise these link rates further. In addition, LIFE-M hand links 14,071 women (6,557 from Ohio, 7,514 from North Carolina) to the 1940 Census for a match rate of 23 to 30%, depending on the state. LIFE-M hand links also include 34,491 links to death certificates, 21,632 links to marriage certificates, and 46,598 links to G3 children. Links rates vary from 15 to 51% depending on the state, record, and sex. Appendix Tables A.2-A.4 report the number of hand links and match rates for G0s, G1s, and G3s for the interested reader.

To evaluate the quality of hand links, we asked the Family History and Technology Lab at Brigham Young University (BYU) to perform an independent quality check of the LIFE-M hand links. BYU compared a random sample of 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% of the time. 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's hand-linked data could be lower, if some observations on the Tree are incorrect.

D. Scaling Hand-Linked Data using Supervised Machine Linking Algorithms

Because hand-linking millions of records is cost and time prohibitive, we rely on supervised learning approaches and our highly vetted hand-linked data to automate record linkage at scale. We first test and compare the performance of several commonly used automated record linking methods, and then describe our development of supervised models to shift the linking frontier.

The Performance of Different Linking Algorithms—A number of studies use rule-based algorithms and unsupervised methods to link data. A common feature of these algorithms is that they tend to rely on exact name matches (or use phonetic conversions of names such as NYSIIS and SDX codes) and use deterministic rules on names and ages. Bailey et al. (2020) tested the performance of these different algorithms and reported the results using three different test datasets: (1) the hand links from the LIFE-M project (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 to 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, the link is coded as correct.
2. If the algorithm link differed from the hand-link, the record was re-reviewed by two additional trainers. The trainers saw a set of candidate links, which included the LIFE-M hand-link ,if one was made, the link made by the automated method, and a machine-generated set of close matches. The trainers did not know which link was chosen by which method and were asked to determine the correct link from the set, if any. If these two reviewers agreed, their decision was coded as the truth.
3. If the two independent trainers disagreed in (2), the same set was sent to an additional three trainers. If the three trainers agreed, their decision, either link or no link, was coded as the truth . If the three trainers disagreed resulting in a 2-3 split among the five trainers, we coded those cases as the record having 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 4 summarizes the performance of several matching algorithms using the LIFE-M hand-links of boys' birth certificates to the 1940 Census (1, above), but the results are very similar using the other datasets as well. 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 (1996) matched 28% of the baseline sample, and Abramitzky, Boustan, and Eriksson (2014) achieved a higher link rate of around 40%, because the method did not impose Ferrie's (1996) uncommon name restriction. Feigenbaum's (2016) regression-based method matched 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' (2020) method adapts the Expectation-Maximization (EM) algorithm of Fellegi and Sunter (1969) to estimate matching probability and linked 46% of the sample when using less conservative cutoffs and 28% of the sample with more conservative cutoffs.

The column on the right in Figure 4 shows the Type I rate (the share of matches that are incorrect divided by the link rate). The share of incorrect links for automated methods was much higher than for clerical review. The lowest Type I error rate occurred in the more conservative version of Abramitzky, Mill, and Pérez (2020) at 15%. Ferrie's (1996) method of selecting uncommon names achieves the second lowest Type I error rate at 25%. Abramitzky, Boustan, and Eriksson's (2014) refinement of Ferrie (1996) increased match rates to 40%, but only half of the added links appeared to be correct, and the Type I error rate increased to 32%. Feigenbaum's (2016) regression-based machine learning model produced a Type I error rate of 34% when using the Iowa coefficients, and the Type I error rate decreased to 29% when estimated using hand linked data. Finally, the less conservative version of Abramitzky, Mill, and Pérez (2020) resulted in the highest Type I error rate at 37%.

An important limitation of rule-based and unsupervised linking algorithms (all but Feigenbaum in Figure 4) is that they produce samples with high rates of false matches in historical data and also miss many true links. Moreover, linking errors have sizable effects on inference (Bailey et al. 2020, Anbinder et al. 2021, Ghosh, Hwang, and Squires forthcoming).

Based on the insights from this evaluation, the LIFE-M project followed Feigenbaum (2016) and MPC's approaches and developed supervised learning methods. 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 incorrect links. The LIFE-M models explicitly (1) set the Type I error rate at 3%, while maximizing the number of total links made; (2) use random forest models to capture non-linearities in linking decisions; and (3) deploy careful feature engineering to model ambiguities. The resulting methods are optimized for linking in limited information settings and use cross-validation to assess out-of-sample performance. The following sections describe our approach in more detail and provide comparisons with other linking methods.

Model Architecture—We developed two model architectures based on the nature of the linking problem. Link types that allow only one match per primary record are called one-to-one links (e.g., linking birth certificates to census records), and those that allow multiple matches per primary are called one-to-many links (e.g., linking birth certificates to sibling birth certificates or birth certificates to marriage certificates).

The one-to-one linking model has two stages. The first stage is called the "any-match" model, and the second stage is called "which-match" model. The any-match model estimates the probability that a link exists *within the set of records*, while the which-match model predicts the *pairwise* match probability for each primary and potential within the set, conditional on the existence of a link within the set of potentials. We multiply the probabilities from the first and second stage to obtain the final match probability for every primary-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 (Breiman 2001) and a log-linear model to estimate the which-match model. The custom log-linear model ensures the probabilities of all primary-candidate pairs within the set sum up to one, so that the final output yields at most one unique link per set.

One-to-many linking models use a random forest. The model takes pairwise information for the primary and candidate links and then predicts the probability of each candidate being a match. Appendix C provides more details on these models.

Feature Generation—The quality of the machine-models also depends on the selection of features. A “feature” is a measurable property or characteristic in machine learning and is similar to an explanatory variable in a regression analysis. To distinguish machine learning from regression analyses, we follow the literature and use “feature” instead of “variable.” Our models typically include 20 to 80 features that help distinguish a link from a non-link or ambiguous case. Examples of pair-level features include Jaro-Winkler distance between the primary and candidate names and the age difference between the primary and candidates. Examples of 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 closest and the second closest candidate (Feigenbaum 2016). We also find that including features relating to the top-five or top-ten matches in the set further aids the model in learning about the human decision-making process.

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

Model Training and Cross-Validation—To assess model performance, we divided the hand-linked data into two equal parts: a training sample and a test sample. We use the training sample to train the model and the test sample to evaluate the out-of-sample performance of the model.

In the training sample, we follow the common practice of using ten-fold cross-validation (Hastie, Tibshirani, and Friedman 2009). This method shuffles the training data and then randomly creates ten, equal-sized subsets (“folds”). For each subset, we select nine subsets to train 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) yields more accurate performance metrics.

Once the probability of a match is determined for every primary-candidate pair using ten-fold cross-validation, we generate machine links using a simple threshold rule: pairs with a probability exceeding a given threshold 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. Altering the threshold to increase precision tends to lower recall and vice versa. There is no universally correct choice, but LIFE-M chose a 97% precision rate, or equivalently, a 3% Type-I error rate. The precision-recall frontier is a function of the features the model uses to learn (reproduce) the decisions of trainers. We maximize the recall rate at a 97% precision rate by iterating over the model features. The threshold varies from model to model, but it is typically between 0.70 to 0.95. Once we are satisfied with the performance of the model, the trained model and probability threshold are used to classify links in the full sample of records. The out-of-sample performance of the model is evaluated using the test sample with the cutoff threshold determined by the training sample.

Figure 5 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 recall rate, and the y-axis is the precision rate. 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 (2016) and LIFE-M, which can deliver precision rates over 90% for recall rates exceeding 70% for North Carolina and 80% for Ohio. Our test reveals that differences between the performance of LIFE-M's and Feigenbaum's methods primarily reflect the extensive set of features in LIFE-M's models.

V. Machine-Linked Samples

Ohio and North Carolina contain over 3.9 million linkable G2 individuals.[14] Depending on the state, 41 to 44% of these linkable observations are women. Supervised learning methods allow us to combine these records with high precision at scale. The next sections describe link rates, representativeness, and the creation of weights using the full sample linked by machine.

A. Link Rates and Sample Sizes

Table 4 provides the link rates for G2s in the full data. The link rates are calculated separately for each state and represent the proportion of linkable 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 and 4), and their G3 children (column 5).[15] Because the model is only able to replicate some of the trainers' decisions with a 3% error rate, link rates are lower in the full sample than in the training data. Depending on the state and sex, 67 to 74% of G2s are linked to their siblings, 12 to 28% are linked to the 1940 Census, 13

[14] "Linkability" is determined by the completeness of name and birth year and is described in the notes of Table 4.

[15] 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.

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 and that names in Ohio tend to be longer than in North Carolina. In addition, link rates are slightly higher for men than women, except for the marriage records.

Appendix Table A.5 presents the total number of links and link rates for G1s in the full data. There are over 3.4 million and 815,000 linkable G1s 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 (G0s).

There are over 2.6 million G0s in the full data (2.2 million in Ohio, and 400 thousand in North Carolina). G0s are 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 G3s 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 the mid or late 20th century, when most G3s reach marriage age or old age. (See Appendix Tables A.6-A.7 for the link rates for G0s and G3s).

Figure 6 provides an overview of the number of adjacent links (parents to children) and non-adjacent links (grandparents to grandchildren) in LIFE-M. There are over 4 million unique adjacent links and more than 2 million unique non-adjacent links. Over 770,000 G2s can be linked to at least one grandparent (through their parent) and a child.

B. Representativeness of Linked Samples

We evaluate the representativeness of the G2 and G1 linked samples by comparing them to the unlinked G2s and G1s in the LIFE-M data. Our test is implemented as a linear regression in which the binary dependent variable is equal to one if the observation is in the linked sample and zero otherwise. The covariates in these regressions include various characteristics of G2s and G1s that are available for all observations, regardless of whether they are linked or unlinked. Table 5 shows the full regression results for G2 linking (regression results for G1 linking are provided in Appendix Table A.8), and Table 6 summarizes our tests of whether the set of covariates jointly predicts whether the observation is linked. For all types of linking, the set of covariates significantly predicts whether an observation is linked. Said another way, the data reject that the linked sample is a random sample of the population, indicating that the linked sample is not representative.

C. Creation of Weights for Linked Data

To address the imbalance in observed characteristics in the linked sample and population, we use inverse propensity-score reweighting (IPW) as described in Bailey, Cole, and Massey (2020). This approach assigns a larger weight to people with characteristics that are under-represented relative to the reference population and a lower weight to people with over-represented characteristics.

The IPW technique and implementation are both simple and powerful. As an example, consider the G2-to-1940 Census links which a researcher would like to reweight to resemble the 1940 population for the same birth state and birth year cohorts. Table 7 compares the means for some demographic and socio-economic variables for the linked G2 sample with the reference population from the 1940 Census. Notably, the unweighted linked sample (column 1) is statistically different from the 1940 Census (column 4). Our generated IPW weights adjust the importance of different individuals to enhance the representativeness of the total sample or subgroups with respect to certain covariates. After applying weights, the mean differences of observed characteristics between the linked G2s and the reference population are both much smaller in magnitude (e.g., 0.402 difference in age versus −0.005) and typically statistically insignificant.

There are no “correct” weights for all analyses—the weight for a particular analysis depends upon the reference population. Using a similar approach, researchers may create IPW for their own analysis to make the linked sample more balanced relative to their reference population of interest.

VI. Opportunities for Research

LIFE-M provides data on more than 15 million individuals born from 1841 to 1968 and belonging to over 4 million families, including high-quality links for 7 million women and half a million underrepresented minorities.[16] These data and corresponding documentation, including variable descriptions and a user guide, have been released for public use, and the most updated version of the data can be downloaded from openICPSR (https://doi.org/10.3886/E155186) and the project website at https://life-m.org (Bailey et al. 2022b).

Ongoing research connects LIFE-M to new health data. Recent work has digitized the cause of death for about 200,000 Ohio individuals (Bailey et al. 2023). These records have been released at ICPSR (http://doi.org/10.3886/E149841) (Bailey et al. 2022a). Using a merge key, researchers can connect these data to the LIFE-M infrastructure and 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. In addition, LIFE-M can be connected to the full-count historical census data as well as IPUMS-LRP and MLP using HISTID. Finally, the LIFE-M geographic file follows the location of individuals from birth to death, including census records and marriage, which allows researchers to connect the millions of individuals to previously unstudied policy interventions based on the county and period when they occurred.

[16] LIFE-M links more than 170,000 Black Americans and more than 368,000 foreign-born people.

The LIFE-M data open many new possibilities for research. For example, LIFE-M uniquely allows a more comprehensive analysis of the role of grandmothers, mothers, and daughters in determining their offsprings' and siblings' outcomes, as well as analyses of how women's own experiences were shaped by their ancestors, birth and marriage families, communities, and life experiences. Another new opportunity for research is the analysis of the interrelationship of health and socio-economic outcomes such as employment, occupation, education, and wage earnings, allowing the construction of mortality gradients over time and across different places in the United States. Yet another example is the unique ability to link individual life outcomes such as mortality to early life exposures and experiences during periods of rapid industrialization, urbanization, economic collapse, and war.

In addition, LIFE-M provides useful methodological insights for other data linking projects. LIFE-M has designed a semi-automated process for creating high-quality training data at scale. The availability of high-quality training data has highlighted the importance of identifying and minimizing high rates of incorrect links in historical linking. In addition, the project has shown that researchers need not accept high rates of linking errors when using automated methods. Using supervised machine learning and thoughtful feature engineering can shift the precision-recall frontier outward, making more correct links and reducing linking errors at the same time. We are optimistic that future research will be able to push this frontier even further. Another contribution of the LIFE-M project is applying the well-known method of inverse propensity-score reweighting from labor economics to linked historical records. This method allows researchers to customize sample weights to balance observed characteristics in the linked sample with a reference population. This technique is helpful in assessing and mitigating issues of non-representativeness in linked samples. As historical linking grows in popularity and availability, identifying problems with existing linking methods and developing data-driven solutions are critical for improving data quality and using these data to answer important research questions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This project was generously supported by the National Science Foundation (SMA1539228), the National Institute on Aging (R21AG05691201), the University of Michigan Population Studies Center Small Grants (R24HD041028), the Michigan Center for the Demography of Aging (MiCDA, P30 AG012846-21), the University of Michigan Associate Professor Fund, and the Michigan Institute on Research and Teaching in Economics (MITRE). We gratefully acknowledge the use of the Population Studies Center's services and facilities at the University of Michigan (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 the 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 project.

VII. References

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

Abowd John M. 2017. "Large-scale Data Linkage from Multiple Sources: Methodology and Research Challenges." NBER Summer Institute Methods Lecture.

Abramitzky Ran, Boustan Leah, and Eriksson Katherine. 2013. "Have the Poor Always been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration." Journal of Development Economics 102:2–14. [PubMed: 26609192]

Abramitzky Ran, Boustan Leah, and Eriksson Katherine. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." Journal of Political Economy 122 (3):467–506. [PubMed: 26609186]

Abramitzky Ran, Boustan Leah Platt, and Eriksson Katherine. 2012. "Europe's Tired, Poor, and Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." American Economic Review 102 (5):1832–1856. [PubMed: 26594052]

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

Abramitzky Ran, Mill Roy, and Santiago Pérez. 2020. "Linking Individuals Across Historical Sources: a Fully Automated Approach." Historical Methods: A Journal of Quantitative and Interdisciplinary History 53 (2):94–111.

Anbinder Tyler, Connor Dylan, 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.

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

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

Bailey Martha J., Cole Connor, Henderson Morgan, and Massey Catherine G.. 2020. "How Well Do Automated Linking Methods Perform? Evidence from US Historical Data." Journal of Economic Literature 58 (4):997–1044. [PubMed: 34294947]

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

Bailey Martha J., Leonard Susan H., Price Joseph, Roberts Evan, Spector Logan, and Zhang Mengying. 2023. "Breathing new life into death certificates: Extracting handwritten cause of death in the LIFE-M project." Explorations in Economic History 87:101474. doi: 10.1016/j.eeh.2022.101474. [PubMed: 36778518]

Bailey Martha J., Lin Peter Z., Shariq Mohammed AR, Paul Mohnen, Jared Murray, Mengying Zhang, and Alexa Prettyman. 2022b. LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2022-12-21. 10.3886/E155186

Bailey Martha J., Shariq Mohammed AR, and Paul Mohnen. 2022. "U.S. Educational Mobility in the Early Twentieth Century." Working paper.

Bengtsson Tommy, and Dribe Martin. 2021. "The Long Road to Health and Prosperity, Southern Sweden, 1765-2015. Research Contributions From the Scanian Economic-Demographic Database (SEDD)." Historical Life Course Studies 11:74–96.

Berger Thor and Eriksson Björn. forthcoming. “Social Mobility in Sweden Before the Welfare State.” Journal of Economic History.

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

Blank Rebecca M., Kerwin Kofi Charles, and Sallee James M.. 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 Ferrie Joseph. 2013. "Up from Poverty? The 1832 Cherokee Land Lottery and the Long-run Distribution of Wealth." NBER Working Paper 19175.

Bleakley Hoyt and Ferrie Joseph. 2014. "Land Openings on the Georgia Frontier and the Coase Theorem in the Short- and Long- Run." Working paper.

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

Breiman L. 2001. "Random Forests." Machine Learning 45:5–32. doi: 10.1023/A:1010933404324.

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

Collins William J., and Wanamaker Marianne H.. 2015. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants." Journal of Economic History 75 (4):947–992.

Collins William J., and Wanamaker Marianne H.. 2022. "African American Intergenerational Economic Mobility since 1880." American Economic Journal: Applied Economics 14 (3):84–117.

Dribe Martin, Eriksson Björn, and Scalone Francesco. 2019. "Migration, marriage and social mobility: Women in Sweden 1880–1900." Explorations in Economic History 71:93–111.

Dribe Martin, and Quaranta Luciana. 2020. "The Scanian Economic-Demographic Database (SEDD)." Historical Life Course Studies 9:158–172.

Elo Irma T, and Samuel H. Preston. 1994. "Estimating African-American mortality from inaccurate data." Demography 31 (3):427–458. [PubMed: 7828765]

FamilySearch.org. https://www.familysearch.org/en/about/ (accessed 2015).

Feigenbaum James J. 2016. "A Machine Learning Approach to Census Record Linking." Working paper.

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

Ferrie Joseph P. 1996. "A New Sample of Males Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules." Historical Methods: A Journal of Quantitative and Interdisciplinary History 29 (4):141–156.

Foxcroft Jeremy, Inwood Kris, and Antonie Luiza. 2022. "Linking Eight Decades of Canadian Census Collections." International Journal of Population Data Science 7 (3):2076.

Ghosh Arkadev, Sam Myoung Hwang Il, and Munir Squires. forthcoming. "Links and legibility: Making sense of historical U.S. Census automated linking methods." Journal of Business and Economic Statistics.

Goldstein JR, Alexander M, Breen C, Miranda A González F Menares M Osborne M. Snyder, and Yildirim U. 2021. Censoc Project. In CenSoc Mortality File: Version 2.0. Berkeley: University of California.

Guest Avery M. 1987. "Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century." Historical Methods: A Journal of Quantitative and Interdisciplinary History 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, Tibshirani Robert J., and Friedman Jerome. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer-Verlag.

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

Kennedy Sheela, and Ruggles Steven. 2014. "Breaking Up Is Hard to Count: The Rise of Divorce in the United States, 1980-2010." Demography 51 (2):587–598. [PubMed: 24399141]

Logan Trevon, and Parman John M.. 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 Ferrie Joseph. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States since 1850." American Economic Review 103 (4):1109–1137.

Modalsli Jørgen. 2017. "Intergenerational Mobility in Norway, 1865-2011." The Scandinavian Journal of Economics 119 (1):34–71. doi: 10.1111/sjoe.12196.

Modalsli Jørgen. 2021. "Multigenerational persistence: Evidence from 146 years of administrative data." Journal of Human Resources:1018–9825.

Mohammed AR Shariq, and Paul Mohnen. 2023. “Black Economic Progress in the Jim Crow South: Evidence from Rosenwald Schools.” Working paper.

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

Ruggles Steven. 2016. "Marriage, Family Systems, and Economic Opportunity in the USA Since 1850." In Gender and Couple Relationships, edited by McHale SM et al. Switzerland: Springer International Publishing.

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

Ruggles Steven, Fitch Catherine A., Goeken Ronald, Hacker J. David, Nelson Matt A., Evan Roberts, Megan Schouweiler, and Matthew Sobek. 2021. IPUMS Ancestry Full Count Data: Version 3.0 [dataset]. Minneapolis, MN: IPUMS.

Schürer Kevin. 2007. "Focus: Creating a Nationally Representative Individual and Household Sample for Great Britain, 1851 to 1901—The Victorian Panel Study (VPS)." Historical Social Research/ Historische Sozialforschung:211–331.

Vézina Hélène, and Bournival Jean-Sébastien. 2020. "An overview of the BALSAC population database: past developments, current state and future prospects." Historical Life Course Studies 9:114–129.

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 Costa Dora L., 1–11. Chicago, IL: University of Chicago Press.

Wisselgren Maria J., Edvinsson Sören, Berggren Mats, and Larsson Maria. 2014. “Testing Methods of Record Linkage on Swedish Censuses.” Historical Methods: A Journal of Quantitative and Interdisciplinary History 47 (3): 138–151.

Wrigley EA, Davies RS, Oeppen JE, and Schofield RS. 2018. 26 English parish family reconstitutions. Colchester, Essex: UK Data Archive

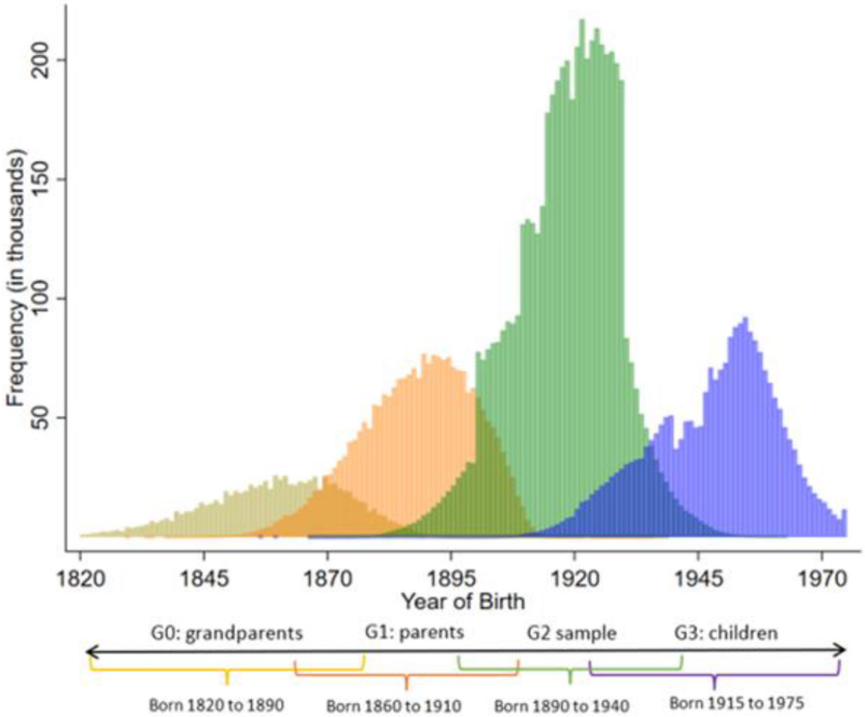

Figure 1. LIFE-M's Approximate Generational Structure

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

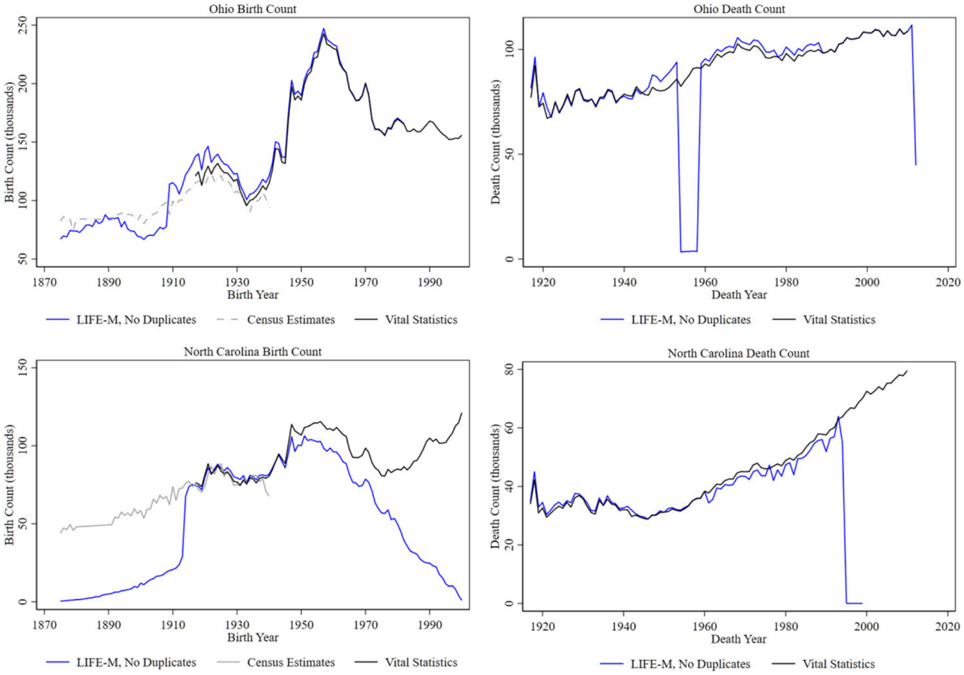

Figure 2. Counts of Microdata Births and Deaths and Vital Statistics and Census Tabulations

Notes: Figures show the counts of birth and death microdata records used by LIFE-M (blue line), birth and death tabulations from vital statistics (black line), and the birth count estimates from the 1880, 1900, 1910, and 1940 Censuses (light gray line). The published vital statistics were digitized and published by Bailey et al. (2016). The population estimates are based on the most recent census prior to a cohort's birth. For instance, we impute the birth counts of the 1900 cohort with the 0-age population in the 1900 Census.

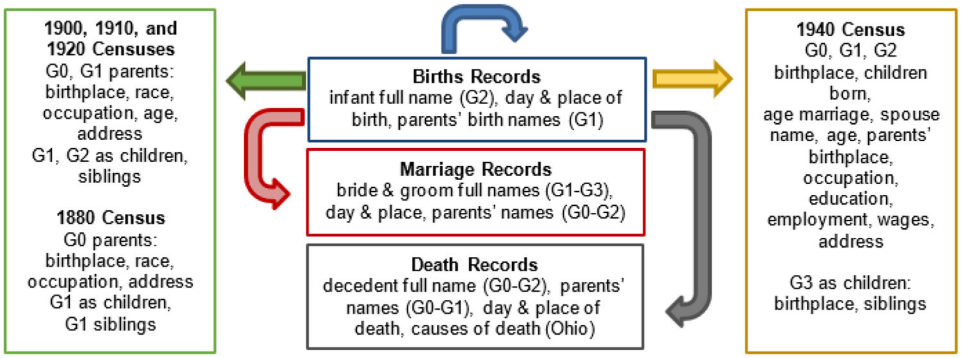
1900, 1910, and 1920 Censuses
G0, G1 parents: birthplace, race, occupation, age, address
G1, G2 as children, siblings
1880 Census
G0 parents: birthplace, race, occupation, address
G1 as children, G1 siblings
Births Records
infant full name (G2), day & place of birth, parents' birth names (G1)
Marriage Records
bride & groom full names (G1-G3), day & place, parents' names (G0-G2)
Death Records
decedent full name (G0-G2), parents' names (G0-G1), day & place of death, causes of death (Ohio)
1940 Census
G0, G1, G2
birthplace, children born,
age marriage, spouse name, age, parents' birthplace, occupation, education, employment, wages, address
G3 as children: birthplace, siblings

Figure 3. Records Combined by LIFE-M

Notes: LIFE-M combines birth, marriage, and death records with 1880, 1900, 1910, 1920, and 1940 Censuses for multiple generations. Vital records provide full legal names (including middle name) and exact date and place of vital events, while census records provide rich social economic information and additional family members not covered in state vital records. All vital records are obtained from FamilySearch.org and full count census records are obtained from Ruggles et al. (2021).

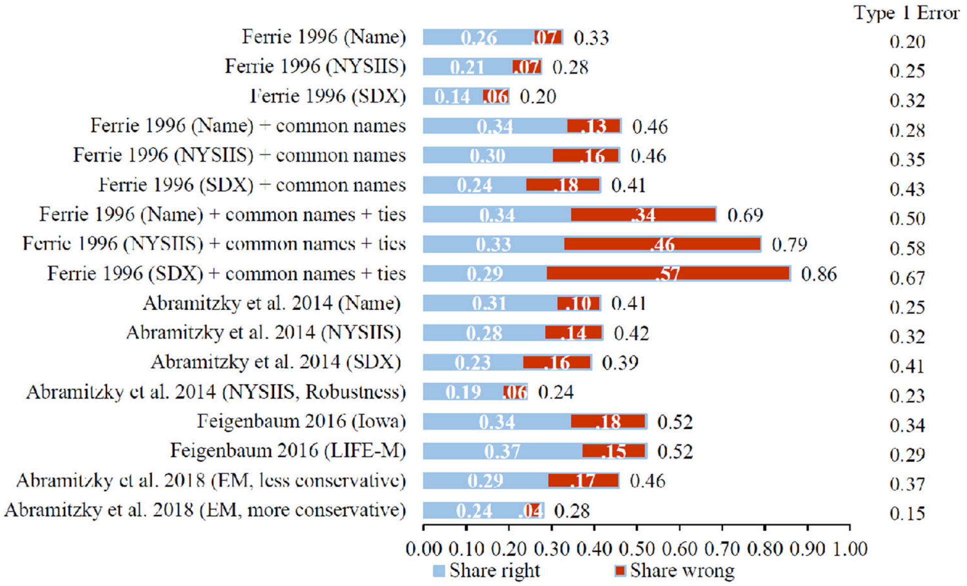

Figure 4. Link Rates and False Links for Commonly Used Linking Methods

Notes: The bars show the performance of different algorithms for LIFE-M boys linked to the 1940 Census based on our hand-linked data. The length of the bar represents the total share of the records linked. The share of incorrect linked records is displayed in red, and the share of correct linked records 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 Inter-University Consortium in Political and Social Science Research, can assist other researchers with replicating these analyses (Bailey and Cole 2019).

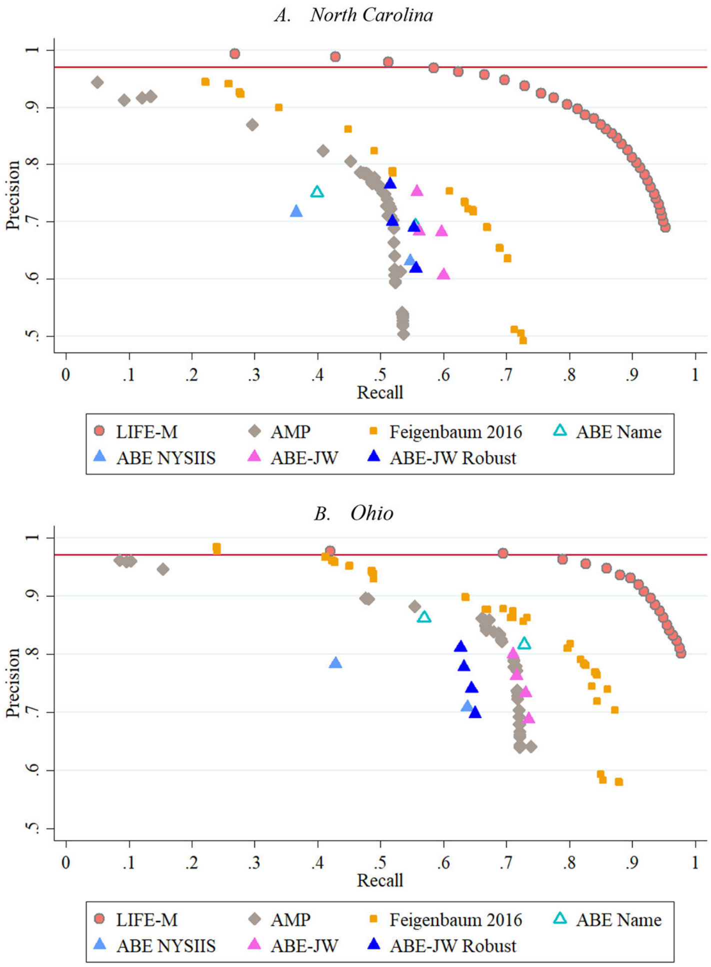

Figure 5. LIFE-M Methods Expand the Precision-Recall Frontier in Historical Linking

Notes: Figures present the precision-recall frontiers for methods commonly used in historical linking. The data are male birth certificates linked to the 1940 Census.

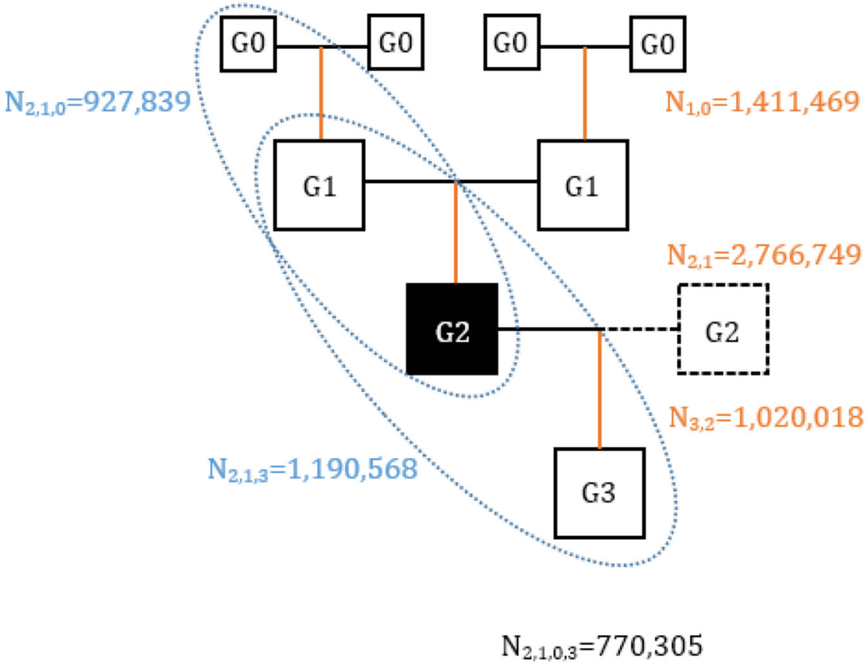

Figure 6. LIFE-M Partial Four-Generation Family Tree

Notes: The dark shaded G2 box provides a reference point for most of the family tree definitions, and the dashed G2 box shows the presence of a spouse. Not all spouses for G2s are present, nor do they need to be to determine the partial family tree. The orange links 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 dashed 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 G2s. $N_{2,1,3}$ identifies the number of G2s that have both a parent and child. $N_{2,1,0}$ identifies the number of G2s with a unique grandparent. Siblings of G2s have the same grandparents and are only counted once. The entire graphic shows a partial 4-generation family. $N_{2,1,0,3}$ identifies how many G2s have at least one grandparent and at least one child.

Table 1.

Example of Sets Displayed to Trainer for Birth Certificate to Census Linking

A. Information Sets May Decrease Trainer Certainty about a Match

	Record	Name	First/Last Name Commonality Scores	Age	Birthplace
4	Candidate	Jason O’Sullvan	(0.58/0.32)	33	Ohio
3	Candidate	Jason O’Sullivon	(0.58/0.32)	35	Ohio
2	Candidate	Jason O. Sullivan	(0.58/0.84)	33	Ohio
1	Candidate	Jason O’Sullivan	(0.58/0.84)	35	Ohio
	Primary	Jason O’Sullivan	(0.58/0.84)	35	Ohio

B. Information Sets May Increase Trainer Certainty about a Match

	Record	Name	First/Last Name Commonality Scores	Age	Birthplace
7	Candidate	Susan H. Ovie	(0.76/0.078)	31	North Carolina
6	Candidate	Shelly H. Olive	(0.53/0.58)	35	North Carolina
5	Candidate	Sheilagh H. Oglvie	(0.091/0.098)	31	North Carolina
4	Candidate	Sally H. Ogivie	(0.76/0.050)	31	North Carolina
3	Candidate	Shelly O’Neill	(0.53/0.71)	33	North Carolina
2	Candidate	Sylvie H. Ogilbie	(0.40/0.28)	29	North Carolina
1	Candidate	Cecela F. Ogilvie	(0.35/0.53)	34	North Carolina
	Primary	Shelagh Harris Ogilvie	(0.20/0.53)	31	North Carolina

Notes: Name commonality scores are computed as a ratio, a/b, where a is the log count of the first or last name under consideration in the 1940 Census and b is the log count of the most common first or last name in the 1940 Census. These examples are truncated to a handful of candidate links for ease of presentation. Training sets typically included up to 30 individuals.

Table 2.

LIFE-M Hand-linked Data by Trainer Decisions

Link Type	(1) Linkable Cases	(2) Agreement (2-0)	(3) Agreement with Trainer Error (4-1 split)	(4) Ambiguous (2-3 split)	(5) Linked but Overturned
			Panel A. Ohio		
G2-Marriage	42,617	*99.01%*	*.39%*	*.21%*	*.39%*
G2-Death	45,325	*95.17%*	*3.27%*	*1.43%*	*.13%*
G2-Census1940	36,007	*85.91%*	*6.02%*	*6.68%*	*1.39%*
G1-Marriage	31,661	*91.4%*	*3.44%*	*3.06%*	*2.1%*
G1-Death	23,645	*92.13%*	*4.03%*	*2.87%*	*.97%*
G1-Census1940	31,815	*88.44%*	*4.13%*	*3.64%*	*3.8%*
G1-Census1920	31,669	*88.47%*	*4.63%*	*2.95%*	*3.95%*
G1-Census1910	31,605	*89.35%*	*4.37%*	*3.09%*	*3.2%*
G1-Census1900	23,946	*85.91%*	*4.86%*	*3.64%*	*5.59%*
G1-Census1880	8,704	*89.04%*	*4.17%*	*4.32%*	*2.47%*
			Panel B. North Carolina		
G2-Marriage	62,587	*89.83%*	*5.07%*	*4.71%*	*.39%*
G2-Death	68,265	*96.91%*	*1.61%*	*.86%*	*.62%*
G2-Census1940	54,304	*82.41%*	*7.19%*	*8.14%*	*2.25%*
G1-Marriage	44,246	*83.75%*	*6.34%*	*3.88%*	*6.03%*
G1-Death	28,898	*90.29%*	*3.77%*	*2.82%*	*3.13%*
G1-Census1940	44,428	*79.07%*	*7.99%*	*7.71%*	*5.24%*
G1-Census1920	44,440	*86.56%*	*5.53%*	*5.17%*	*2.74%*
G1-Census1910	41,436	*86.77%*	*5.51%*	*5.43%*	*2.28%*
G1-Census1900	27,870	*81.86%*	*7.37%*	*4.73%*	*6.04%*
G1-Census1880	5,504	*83.25%*	*8.85%*	*4.16%*	*3.74%*

Notes: Linkable cases (1) include all records presented to trainers for decisions. Columns (2)-(5) show the proportion of linkable cases by trainer decisions. Agreements are the cases where the first two trainers agreed (either link or no link). Agreement with trainer error are cases where the first two trainers disagreed but the additional three trainers reached an agreement (4-1 in favor of a link or no link). Ambiguous cases are those for which the first two trainers disagreed and the additional three trainers failed to reach an agreement (3-2, split decision). Linked but overturned cases include those where the LIFE-M team overturned the link due to conflicts after linking.

Table 3:

Link Rates for G2 Hand-Links from Different Data Sources

	(1)	(2)	(3)	(4)	(5)
	Siblings	**1940 Census**	**Death**	**Marriage**	**Children**
Panel A: Ohio					
All	87.3	35.4	29.9	21.9	47.7
	31,299/35,857	17,534/49,519	14,945/49,940	10,941/49,940	23,803/49,940
Males	87.7	45.0	37.8	21.9	50.6
	13,260/15,126	10,977/24,377	9,295/24,570	5,390/24,570	12,434/24,570
Females	87.8	29.9	24.4	25.1	51.4
	11,901/13,562	6,557/21,931	5,394/22,104	5,551/22,104	11,369/22,104
Panel B: North Carolina					
All	83.5	31.9	27.8	15.2	32.5
	28,831/34,548	22,264/69,832	19,546/70,219	10,691/70,219	22,795/70,219
Males	82.8	42.3	35.8	14.6	34.6
	13,730/16,588	14,750/34,903	12,574/35,110	5,114/35,110	12,133/35,110
Females	83.8	23.2	21.5	17.2	32.8
	12,731/15,186	7,514/32,328	6,972/32,484	5,577/32,484	10,662/32,484

Notes: Link rates are calculated as the number of links divided by the number of linkable people in the hand-linked sample. Linkable people for column (1) are those in the sample who have non-missing parents' names (father's first name, last name, and mother's first name). Linkable people for column (2) are those who were born by 1940 and have non-missing names (both first and last names). Linkable people for columns (3)-(5) are G2s with non-missing names. Links to children include all G2s with at least one child (G3s) found from either birth records or the 1940 Census. The number of links and linkable people for males and females do not sum up to the "All" number, because some G2s have unknown sex not displayed for brevity). Link rates and counts are updated with each data release. All counts in this table and subsequent tables/figures come from version 1 of the public LIFE-M data. The most updated tabulations are reported on our website at https://life-m.org/linking/.

Table 4.

Link Rates for G2s in the LIFE-M Data

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

Notes: Link rates are calculated as the number of links divided by the number of linkable people. Linkable people for column (1) are G2s who were born from 1900 to 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 before 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 males and females do not sum up to the "All" category, because some G2s have unknown sex (not displayed for brevity). Link rates and counts are updated with each data release, and the most updated tabulations are reported on our website at https://life-m.org/linking/.

Table 5.

Representativeness of Linked G2 Records

	(1) G2-Sibling	(2) G2-1940 Census	(3) G2-Death	(4) G2-Marriage	(5) G2-G3
Day of Birth	0.00000485*** (0.00000144)	−0.000246*** (0.00000190)	−0.0000265*** (0.00000188)	0.00000955*** (0.00000172)	−0.0000140*** (0.00000200)
Number of Siblings	0.0674*** (0.0000675)	0.00619*** (0.0000854)	0.00908*** (0.0000862)	0.00589*** (0.0000791)	0.0135*** (0.0000920)
Length of Name	0.00101*** (0.0000401)	0.0104*** (0.0000457)	0.00853*** (0.0000458)	0.00711*** (0.0000407)	0.0165*** (0.0000477)
Length of Father’s Name	0.00781*** (0.0000408)	0.00187*** (0.0000548)	0.00168*** (0.0000540)	0.000294*** (0.0000482)	0.00101*** (0.0000587)
Length of Mother’s Name	0.0146*** (0.0000328)	−0.000297*** (0.0000441)	0.00186*** (0.0000444)	0.0000652* (0.0000377)	0.00201*** (0.0000468)
Share of Birth Records with Misspelled Father’s Name	0.299*** (0.000687)	−0.00351*** (0.00109)	0.00525*** (0.00110)	−0.0133*** (0.00101)	−0.0158*** (0.00116)
Share of Birth Records with Misspelled Mother’s Name	0.382*** (0.000646)	0.0526*** (0.000978)	0.0292*** (0.000978)	0.0574*** (0.000916)	0.0269*** (0.00104)
Link in Ohio	0.0751*** (0.000397)	0.0848*** (0.000458)	0.00233*** (0.000506)	0.103*** (0.000394)	0.0577*** (0.000513)
Female	0.0000647 (0.000304)	−0.0409*** (0.000394)	−0.0870*** (0.000394)	0.0174*** (0.000362)	−0.0194*** (0.000420)
Constant	−0.0239*** (0.000778)	0.00497*** (0.000956)	0.0471*** (0.000941)	−0.0614*** (0.000823)	−0.109*** (0.000993)
Observations	4,297,569	4,297,569	4,297,569	4,297,569	4,297,569
R^2	0.511	0.033	0.027	0.029	0.043
F-statistic	553612.5	19394.9	14700.0	19632.3	25877.8

Notes: Each column reports the regression coefficients from a linear probability model using the sample of all G2s in the LIFE-M data regardless of whether the observation is linked. The dependent variable is an indicator variable equal to one if the G2 record is linked to the data source appearing in the column title and zero otherwise. The covariates for G2 links include the birthday 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 dummy variables for sex and state. Regressions pool Ohio and North Carolina as well as men and women. F-statistics are reported for a heteroskedasticity-robust Wald-test of joint significance of covariates. Robust standard errors are reported in parentheses. *** indicates the variable is statistically significant at the one-percent level.

Table 6.

Representativeness of Linked G2 and G1 Records, 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: Each F-statistic (and corresponding p-value in parenthesis) in this table is from a separate, linear regression, in which the dependent variable is an indicator equal to one for observations in the subsect of G2 or G1 linked samples. The F-statistic is from a heteroskedasticity-robust Wald test of the joint significance of all covariates. The regressions use either all G2s or G1s in the LIFE-M data, regardless of whether they are linked. The regressions in the top row pool men and women, and the regressions in the bottom two rows are restricted to men or women. All regressions pool Ohio and North Carolina. See Table 5 for more information on covariates for G2s. The covariates for G1s include number of children in G1-G2 family, name length, name length of spouse, share of birth records with a misspelled name, share of birth records with misspelled spouse name, and dummy variables for sex and state.

Table 7.

Characteristics of the Linked Sample, with and without Weights

	Linked G2s, Unweighted (1)	Linked G2s, Weighted (2)	1940 Census (3)	Difference (1)-(3) (4)	Weighted difference (2)-(3) (5)
		Panel A: Ohio			
Age in 1940	20.191	19.784	19.789	0.402***	−0.005
Male	0.562	0.504	0.504	0.058***	0.000
White	0.983	0.970	0.970	0.013***	0.000
Urban Status	0.617	0.647	0.648	−0.031***	−0.000
Farm Status	0.191	0.159	0.159	0.032***	0.000
High School Graduate	0.536	0.478	0.478	0.058***	0.001
Never Married	0.688	0.653	0.653	0.036***	−0.000
Occupational Income Score	23.496	23.677	23.671	−0.175***	0.006
Migration: 5-year	0.036	0.047	0.048	−0.012***	−0.000
		Panel B: North Carolina			
Age in 1940	20.584	19.731	19.738	0.846***	−0.007
Male	0.630	0.516	0.517	0.113***	−0.001
White	0.829	0.726	0.725	0.104***	0.001
Urban Status	0.280	0.282	0.283	−0.003*	−0.001
Farm Status	0.445	0.455	0.454	−0.010***	0.001
High School Graduate	0.396	0.298	0.298	0.098***	0.000
Never Married	0.679	0.631	0.632	0.047***	−0.000
Occupational Income Score	19.884	18.715	18.660	1.224***	0.055
Migration: 5-year	0.046	0.045	0.044	0.002**	0.000

Notes: Column (1) reports the unweighted means for characteristics in the linked sample of G2s born between 1900 and 1940 to the 1940 Census. Column (2) reports the weighted mean of these characteristics for the same G2s in column (1). Column (3) reports the means of characteristics for the 1940 Census population for those who were born in Ohio and North Carolina between 1900 and 1940. Five-year migration status is directly reported in the 1940 Census for people born by 1935. We exclude individuals having missing years of schooling in both the linked sample and population because we include years of schooling as a weighting variable. Columns (4) and (5) report differences in means between indicated columns. *** indicates the variable is statistically significant at the one-percent level.