IPUMS
Working Papers

The IPUMS Multigenerational Longitudinal Panel: Progress and Prospects

Steven Ruggles, University of Minnesota†

Julia A. Rivera Drew, University of Minnesota

Catherine A. Fitch, University of Minnesota

J. David Hacker, University of Minnesota

Jonas Helgertz, Lund University

Matt A. Nelson, University of Minnesota

Nesile Ozder, University of Minnesota

Matthew Sobek, University of Minnesota

John Robert Warren, University of Minnesota

July 2024

Working Paper No. 2024-05
DOI: https://doi.org/10.18128/IPUMS2024-05

†Address correspondence to Steven Ruggles: ruggles@umn.edu. This research was supported by a grant from the National Institute on Aging (R01-AG057679).

The IPUMS Multigenerational Longitudinal Panel: Progress and Prospects

Steven Ruggles, Julia A. Rivera Drew, Catherine A. Fitch, J. David Hacker, Jonas Helgertz, Matt A. Nelson, Nesile Ozder, Matthew Sobek, and John Robert Warren

Abstract:

The IPUMS Multigenerational Longitudinal Panel (MLP) is a longitudinal population panel that links American censuses, surveys, administrative sources, and vital records spanning the period from 1850 to the present. This article explains the rationale for IPUMS MLP, outlines the design of the infrastructure, and describes the linking methods used to construct the panel. We then detail our plans for expansion and improvement of MLP over the next five years, including the incorporation of additional data sources, the development of a "linkage hub" to connect MLP with other major record linkage efforts, and the refinement of our technology and dissemination efforts. We conclude by describing a few early examples of MLP-based research.

Keywords: census; record linkage; longitudinal data; IPUMS

Author Information: Steven Ruggles, University of Minnesota, Department of History, https://orcid.org/0000-0001-5353-2578, @HistDem, (corresponding author: ruggles@umn.edu); Julia A. Rivera Drew, University of Minnesota, Institute for Social Research and Data Innovation; Catherine A. Fitch, University of Minnesota, Institute for Social Research and Data Innovation; J. David Hacker, University of Minnesota, Department of History and Minnesota Population Center, https://orcid.org/0000-0002-5971-955X ; X(Twitter): @jdavidhacker612; Jonas Helgertz, Lund University, Department of Economic History and Centre for Economic Demography, https://orcid.org/0000-0002-2200-9095; Matt A. Nelson, University of Minnesota, Institute for Social Research and Data Innovation, https://orcid.org/0000-0002-8849-4628, X(Twitter): @mattnelson; Nesile Ozder, University of Minnesota, Institute for Social Research and Data Innovation; Evan Roberts, University of Minnesota, Program in the History of Medicine, https://orcid.org/0000-0001-5621-4823, X(Twitter): @evanrobertsnz; Matthew Sobek, University of Minnesota, Institute for Social Research and Data Innovation; John Robert Warren, University of Minnesota, Department of Sociology, https://orcid.org/0000-0001-6079-5122.

Introduction

Over the past five years, the IPUMS Multigenerational Longitudinal Panel (MLP) has constructed a massive longitudinal population panel by linking together U.S. censuses, surveys, administrative sources, and vital records spanning the period from 1850 to the present. MLP is designed as a general-purpose resource for studying longitudinal processes and long-run social and economic change. By linking individuals across generations from birth through death using data from multiple sources, MLP enables reproducible research on shifting life course patterns and intergenerational processes. The large scale of the MLP database allows researchers both to study particular communities, small dispersed populations, and to conduct big studies spanning many places and periods. Data that allow investigators to examine simultaneously the broad sweep of time and fine spatial detail will yield new insights into ongoing transformations of demographic behavior.

This article describes the creation of the IPUMS MLP database and its potential uses for research. We begin with a brief overview of the history of our record linkage efforts over the past two decades and describe the current MLP linking strategy. We then detail our plans for expansion and improvement of MLP over the next five years, including the incorporation of additional data sources, the development of a "linkage hub" to connect MLP with other major record linkage efforts, and the refinement of our technology and dissemination efforts. Finally, we conclude by describing a few early examples of MLP-based research.

Background

Historians have been linking individuals between censuses since the 1930s (e.g., Malin 1935; Owsley and Owsley 1940; Curti 1959; Thernstrom 1964; Katz 1975). Starting with a base census year, investigators identified a population of men residing in a locality and searched the census listings for the same locality in a subsequent census year to locate the same individuals. These studies linked only a small percentage of men, partly because of high geographic mobility, and there was little means to assess the quality of the links. Beginning in the 1980s there were several attempts to develop national samples of linked data capitalizing on genealogical indexes (Guest 1987, Steckel 1988, Ferrie 1996), but the costs were high, linkage rates were low, and false links were difficult to ferret out (Ruggles, Fitch, and Roberts 2018).

In 2003 IPUMS released a full count microdata file covering the entire United States Census of 1880, comprising over 50 million person records. Over 1,000 volunteers of the Church of Jesus Christ of Latter-Day Saints (LDS) logged 11.5 million hours of data entry over a 17-year period to transcribe the census. Between 1999 and 2002 the IPUMS team partnered with LDS to clean the data and make it available for both genealogical and demographic research (Ruggles 2023).

The 1880 full count census file made it feasible to develop new automated record linkage strategies. The IPUMS Linked Representative Samples (LRS) linked the 1880 census to each of the 1% IPUMS samples for the period from 1850 to 1930. Each linked sample consisted of two observations for each linked individual, one from the census sample and the other from the complete 1880 census.

Record linkage is subject to two kinds of errors: false matches (Type I errors) and missed matches (Type II errors). Our primary goal for LRS was to minimize false links (Type I errors). False matches are the greatest concern because they introduce systematic upward biases in transition rates, such as migration, economic mobility, family transitions, or fluidity in racial identification. Suppose, for example, that an investigator seeks to measure migration. Falsely matched cases usually appear to be migrants, since two incorrectly linked individuals are extremely unlikely to reside in precisely the same place. We developed a conservative two-step machine-learning algorithm that rejected any link where a potential alternative link existed (Goeken et al. 2011, Ruggles et al. 2011). IPUMS-LRS yielded lower estimates of migration and intergenerational occupational mobility than did the Long-Ferrie linked samples (Baskerville et al. 2014). This result suggests that the extremely high geographic and occupational mobility found in other linkage studies may be partly ascribed to false matches (Long and Ferrie 2013; Bailey et al. 2020).

A decade after the release of the full count 1880 Census, IPUMS released a full count microdata file for the 1940 census based on a collaboration with the commercial genealogical firm Ancestry.com. The new file created an opportunity to link individuals in ongoing longitudinal surveys back to their childhood homes in the 1940 Census. Building on the LRS linking strategies, we linked the 1940 census to the Health and Retirement Study (HRS); the National Social Life, Health, and Aging Project (NSHAP); and the Wisconsin Longitudinal Study (WLS) (Warren, Lee, and Osypuk 2022; Modrek et al. 2022; Warren et al. 2020). These ongoing longitudinal studies are key elements of data infrastructure for interdisciplinary research on aging and health, including physical and mental health, cognition, disability, and well-being. A key limitation of these surveys, however, was their lack of information about participants' social, economic, family, neighborhood, and environmental circumstances in childhood and young adulthood. By linking survey respondents to their childhood homes in the 1940 census, we not only observe their families' socioeconomic and other circumstances, but we also see their households' precise spatial locations—allowing for linkages to environmental, policy, economic, and other contextual exposures. The linked survey data allow direct prospective analyses of the effects of early life educational opportunities, environmental toxins, neighborhood characteristics, and other exposures on later life cognitive, health, and economic outcomes.

Between 2013 and 2019 IPUMS collaborated with genealogical organizations to develop full count data for all the remaining U.S. census years between 1850 and 1940, a total of almost 700 million records (Ruggles 2023). As soon as the full count data with names became available, economists and sociologists began developing methods for automated record linkage to construct longitudinal panels. This work resulted in a flood of new research papers on a wide range of topics (Ruggles, Fitch and Roberts 2018). The ferment of new research is exciting, but serious technical problems have arisen. Few investigators are well versed in advanced record linkage methods, and few have access to the software or high-performance computing needed for reliable large-scale linkage. Almost all these studies use off-the-shelf statistical packages to do the matching, forcing compromises that reduce the reliability of the links.

Recent analyses have demonstrated that the most commonly used methods of automatic record linkage—which are usually based on deterministic rules and phonetic classifications of names—have false match rates ranging from 20% to 70% (Anbinder et al. 2021; Bailey et al. 2020; Ghosh, Hwang, and Squires 2023; Helgertz et al. 2022; Massey 2017; Ruggles, Fitch, and Roberts 2018). Such high error rates will yield invalid estimates of transition rates. Missed matches (Type II error) are also problematic since they can introduce selection bias and reduce the representativeness of longitudinal panels.

In 2018 IPUMS began developing MLP based on novel automatic record-linkage technology that substantially improves on prior methods (Helgertz et al. 2022). The core of MLP now consists of nine censuses covering the entire U.S. population enumerated between 1850 and

1940. The first version, released in 2020, consisted of 208 million links across these censuses (Helgertz et al. 2020). From 2022 to 2024 we published substantial improvements to the infrastructure, including links to the Social Security enrollment database and to the vital records from the Longitudinal, Intergenerational Family Electronic Micro-Database (LIFE-M) (Helgertz et al. 2023; Bailey et al. 2023).

MLP Linking Strategy

By using all available information and leveraging artificial intelligence—including information on household characteristics and geographic location—MLP achieved substantially higher precision (lower Type I errors) and recall (lower Type II errors) than earlier methods of automated record linkage. An independent evaluation of eight algorithms that have been used to link U.S. censuses reached the following conclusion:

> The main finding is that one particular algorithm, proposed in Helgertz et al. (2020) and referred to as the MLP algorithm henceforth, scores significantly higher than any other algorithm on both quality measures over the entire range of legibility. This result holds for a variety of subsamples, and similar results hold after applying inverse probability weighting (Ghosh, Hwang, and Squires 2023: 6).

Ghosh, Hwang, and Squires compared MLP's linking approach with previous leading linkage methods and found that—by a broad margin—MLP provides the highest linkage rate, the highest share of validated links, and the most accurate estimates of 10-year migration rates. When Helgertz et al. (2022) compared alternative methods against a genealogical gold standard, they found that MLP Type I errors were 68% lower and Type II errors were 65% lower than the next best algorithm.

This improved performance comes at the cost of greater complexity; the MLP algorithm is based on a much richer set of features (population characteristics) than the alternative methods. Moreover, MLP uses supervised machine learning instead of the more commonly used deterministic algorithms. To implement the algorithm, we developed an end-to-end Python package called "Hlink" that provides a flexible, configuration-driven solution to probabilistic record linking at scale. Hlink includes a high-level Applications Programming Interface for Python as well as a standalone command line interface for running linking jobs with minimal programming. Hlink supports the linking process from beginning to end, including preprocessing, filtering, training, model exploration, blocking, feature generation and scoring. The Hlink system uses a distributed processing engine built on Apache Spark to spread the workload across a cluster of computers and leverages the column-store Apache Parquet system to optimize sequential processing. We released the open source Hlink software on GitHub in May 2022 (IPUMS 2022).

The specific methods and procedures MLP used to link individuals across census records, surveys, administrative records, and vital records vary depending on the characteristics of the source, but several elements are common to all our supervised machine learning record linkage. First, every pair of records drawn from two files is either a match referring to a single individual or is a non-match describing two different people. Optimal matching requires that every individual be compared with every possible match (Felligi and Sunter 1969), but this is not computationally feasible for linking the full count census data, since it would require some 3.8 x 10^{25} comparisons to construct the links across all census years. Second, to reduce the computational requirements, we introduce "blocking factors" that define a subset of the population and limit comparisons to persons who share the same blocking factors. We currently define a set of potential matches for each individual as persons who share the same sex, state of birth, and age within +/- three years. Following Christen (2012), we also block by surname bigrams—which are successive pairs of letters in the surname—thereby reducing the number of needed comparisons by 75% while retaining virtually 100% of true matches. Finally, we limit potential matches using the Jaro-

Winkler string comparison metric developed by the Census Bureau to aid record linkage (Porter and Winkler 1997); candidates for linkage must have a Jaro-Winkler similarity score of 0.7 or greater on both first name and surname. As we improve the performance of our Hlink record linkage software, we expect it will be feasible to relax these blocking factors to consider a broader pool of potential matches, thus further reducing both Type I and Type II errors.

The precision of the machine-learning models depends on accurate training data that identifies true and false matches (Bailey et al. 2020). To obtain high-quality "true" links, we use genealogical tools. Trained research assistants identify cases with high confidence that the same individual appears across two sources. These links are independently reviewed by two trainers, and any discrepancies are arbitrated by a third trainer. The training data are then divided in half; one half is used to train the model, and the other half to evaluate accuracy of the model.

For linking across censuses, MLP uses a two-step process:

- Step 1 model. The first step identifies a set of high confidence matches among all potential matches. Our approach is unusual in its use of an extended set of characteristics—or "features" in the machine-learning literature—when comparing a potential match. In addition to individual-level characteristics used by most historical census linkage projects, our approach exploits an extended set of individual, household and contextual characteristics when predicting the degree of similarity between a given set of census records. The variables used vary slightly depending on the particular pair of censuses being matched, but in most cases the model uses about 26 variables. Hlink works with any supervised machine-learning algorithm. The initial version of MLP used logistic regression, but the evidence suggests that the XGBoost open-source library currently provides the highest accuracy for cross-census record linkage

(Chen and Guestrin 2016; Synced 2017). For each individual, a link to the potential match with the highest predicted similarity is confirmed when two conditions are fulfilled: (1) The predicted similarity of a combination of census records exceeds a given threshold (α), and (2) the predicted similarity of the most probable census record combination (α1) exceeds the predicted similarity of the second-most probable combination of census records (α2) by at least (β). Consequently, α1/α2 > β. The parameter estimates used to predict the similarity between two records, as well as the thresholds for α and β, are obtained through calibration of the algorithm with the training data. In addition, training data are used to evaluate the expected quality of the resulting linked data.

- Step 2 model. To maximize the number of accurate links, in Step 2 we link other household members of individuals who were linked in Step 1. The universe of individuals whom we attempt to link in Step 2 is restricted to those residing in a household with at least one other individual who was successfully linked in Step 1. In essence, Step 2 blocks by household, and potential matches may have an age discrepancy of +/- ten years. The procedure employed to confirm a match follows a similar strategy to Step 1 but uses a different set of linking characteristics. In addition, point estimates and α/β threshold values are calibrated based on specialized training data and used to determine which links constitute a match.

The early releases of the linked data were in the form of crosswalks, which required users to do their own matching of data between sources using a statistical package. In February 2023 we introduced the Historical Identification Key (HIK), which assigns a unique identifier to individuals that is consistent across all sources. We also integrated MLP into the IPUMS data access system so that users can easily select individuals who are linked in any specified combination of sources.

These innovations dramatically simplify the construction of panels of linked data that are customized for particular research problems.

MLP Expansion and Improvement

Over the next five years we plan expansion, improvement, and dissemination of the system of linked data from censuses, surveys, administrative sources, health systems, and vital records spanning the period from 1850 to the present. We will expand the scope of MLP record linkage to incorporate newly available data from the 1950 census, the National Vital Statistics System, ongoing surveys of aging, and data derived from clinic and hospital records. We will integrate MLP with other major record linkage projects to make them interoperable and simplify analyses combining data from multiple projects. To make this massive infrastructure sustainable and accessible, we also plan refinement of software and methods for automatic record linkage, maintenance of the system of linked records, and dissemination of linked data.

Figure 1. Existing and planned MLP infrastructure and links to collaborating projects

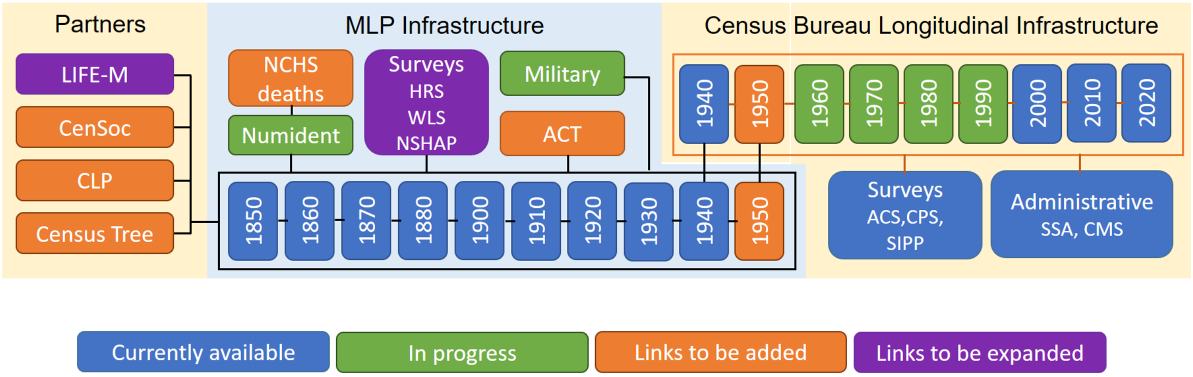

Figure 1 summarizes the data components of MLP and partner linkage projects, identifying those elements that are currently available or in progress (blue and green, respectively) and those we now plan to add or expand (orange and purple). With the core MLP infrastructure now in place, a central activity for the next five years will be adding data linkages to rich new sources of population and health data. The following paragraphs outline the significance of each of these resources.

The 1950 Census. IPUMS is collaborating with genealogical organizations to produce a complete set of microdata based on the handwritten enumeration forms from the 1950 Census of Population. With 151 million individuals, 43 million households, and 62 data fields, this is the largest handwriting transcription project from a single source ever undertaken. Despite the powerful scientific impact of the 1940 census, the new database comprising the entire 1950 census will have even more profound consequences for research. The most important factor is timing: children residing in their parental households in 1950 are now 74 to 91 years old. This age group offers far better opportunities for prospective analysis than do older cohorts that can be observed in the 1940 or earlier censuses. The 1950 census provides a key baseline for analyzing the impacts of social and economic change on later life health outcomes: it was the first U.S. census to ask about total income, self-employment income, and unearned income. Used in conjunction with other census records, these new data will help us understand long-run shifts in life-course patterns of migration, economic opportunity, fertility, and family transitions. Individual and family records from 1950 will allow investigators to enrich recent sources of data on older Americans. The database will cover the entire population with full geographic detail, providing *contextual* information on childhood neighborhood characteristics, labor-market conditions, policy contexts, and environmental exposures for persons now in later life. Precise geographic information will enable new kinds of spatial analyses, including identifying geographic clusters of causes of death.

Administrative records and strategies for linking women. Names are the single most important linking variable, so linking women across census years is complicated by name changes at marriage. In the current MLP methodology, it is usually only feasible to link women who do not marry during the interval between censuses. Accordingly, to trace women from childhood to old age we must turn to other sources, most importantly, the public version of the Social Security Applications and Claims Index (Numident) (National Archives and Records Administration 1936-2007), which includes persons with a Social Security number who either died or would have reached age 110 by December 31, 2007. The Numident records include each Social Security applicant's full name, maiden name, exact dates of birth and death, place of birth, place of death, sex, father's name, mother's maiden name, and race. We will use the Numident to follow women from their families of origin to the families in which they reside as adults. We can further improve links of women across census years by leveraging vital records from LIFE-M (Bailey et al. 2020) and crowd-sourced genealogical data from the Census Tree (described below) (Buckles et al. 2023). We plan to use these data to refine and augment the IPUMS MLP links, particularly for women who change their names at marriage.

Expansion of Historical Identification Keys (HIKs). For each individual, we will construct variables identifying the linking keys for spouses, mothers, fathers, up to four grandparents, and up to eight great-grandparents. These interrelationship keys will be available in all census years. For example, a future spouse will be identifiable when that individual is still a child. Siblings will be identified because they share a parental linking key, cousins because they share a grandparent key, and so on. Accordingly, the family interrelationship variables will not only be valuable for assessing family influences across multiple generations but will also allow study of lateral kin relationships beyond the household.

Death certificate data. The project will link more than 30 distinct variables on official death certificates to repeated observations on the full U.S. population spanning the past 100 years, including information on family circumstances, demographic characteristics, occupation, and geographic location. Linked death certificate information will encompass the underlying cause of death, up to 20 different contributing causes of death, the place (e.g., at home, as an inpatient) and manner of death, the geographic location of death, and demographic information about the decedent. Combining information about an individual's death with data collected while they were alive creates a powerful resource for understanding the social determinants of health and mortality, including how earlier life context and exposures shape their risk and cause of death. Linked mortality data allow researchers to produce customized group-specific death rates beyond what is available in published vital statistics. Such data also provide the unique opportunity for researchers to infer the effects of environmental or other contextual exposures on mortality and to reconstruct life biographies to understand the timing and causes of death. The linked death certificates will comprise the largest and most comprehensive resource available for individual-level analyses of mortality (Centers for Disease Control and Prevention 2022), enabling investigation of a vast array of topics central to the study of health and aging. For example, the data will identify adults whose underlying or contributing cause of death was Alzheimer's Disease and related dementias. Researchers can exploit the rich set of MLP family relationship and sociodemographic information to assess the contributions of early- and mid-life environmental exposures to later-life mortality, and to investigate fatal disease clusters among kin groups, adults who lived on the same block when they were children, or occupational groups.

Linking strategy for death certificate data. Linking MLP to death certificates will rely on the Social Security Numident file, using the Hlink machine-learning algorithm. We will extract approximately 45 million Social Security Numbers, names, and month and year of birth from the Numident and submit them to the National Death Index (NDI) for identification. NDI will provide a death certificate number for each death found. Based on this search, IPUMS will provide a “finder file” to the Mortality Statistics Branch of the National Center for Health Statistics (NCHS) consisting of the death certificate number, year of death, and state of death.

The NCHS will match the IPUMS finder file to the Mortality Historical File, a comprehensive database with death certificate information provided to NCHS by states and other local reporting jurisdictions. The fields in the death certificate data include death certificate number, date of death, underlying cause of death, multiple causes of death, demographic characteristics, tobacco use, and contextual information. Under our Data Use Agreement with NCHS, we will produce two data products: (1) A restricted dataset with more than 30 variables from the death certificate linked to the MLP data, for which NCHS will serve as the data steward, and (2) a more limited version disseminated directly through IPUMS. The restricted version will include full information on causes and manner of death and a variety of demographic variables. The IPUMS version will include a 113-category recode of underlying cause of death and multiple cause of death (MCOD) flags for selected contributing causes of death, such as Alzheimer's Disease and related dementias, diabetes, falls, sepsis, and influenza/ pneumonia. Incorporating MCOD flags in addition to the underlying cause of death allows for a much more inclusive identification of deaths from causes of interest than available in other linked data sources. For example, during the 2018-2021 period, 497,159 individuals had Alzheimer's Disease listed on the death certificate as an underlying cause of death, and an additional 122,380 individuals had Alzheimer's listed as a multiple cause of death (Centers for Disease Control and Prevention 2022). Adding an Alzheimer's MCOD flag over this four-year period increases the number of identified Alzheimer's-related deaths by 25%.

Longitudinal surveys and Adult Changes in Thought. We will link MLP to population-representative cohort studies that follow large and diverse samples of Americans across the life course. Among these, the Health and Retirement Study (HRS), the Wisconsin Longitudinal Study (WLS), and the National Social Life, Health, and Aging Project (NSHAP) are among the cornerstones of research on the lifelong set of exposures and conditions that shape population health, aging, cognitive functioning, mortality, and other important outcomes. As noted, we have linked these three sample surveys of older Americans to the 1940 census—in which sample members are observed in their families and communities in that year (Warren, Lee, and Osypuk 2022; Modrek et al. 2022; Warren et al. 2020). In just a few years these links have led to high-profile scholarly articles on the effects of early-life environmental, policy, and socioeconomic exposures on later-life outcomes. We now plan to extend this work and link the three longitudinal surveys to the 1950 census and to the rest of the MLP infrastructure. Participants in these surveys will now be observable in every decennial census conducted during their lifetimes and in vital records (adding, for example, important new information from birth certificates and new details about underlying causes of death).

We also plan to produce an entirely new (for MLP) class of linked data, adding records from a large health system that were collected as part of the Adult Changes in Thought (ACT) study (Montine et al. 2012). ACT has enrolled and followed several thousand older adults who receive care as Kaiser Permanente Washington members, creating comprehensive longitudinal datasets that can support a wide range of research on factors that affect brain aging. While lacking the representative national coverage of survey-based studies like the HRS or NSHAP, the ACT study is unique in its high-resolution information about later-life health and cognition. ACT combines measures from study-specific surveys and visits; electronic health records; abstracted medical records; accelerometers; genomic and blood biomarkers; MRI-based neuroimages; and neuropathological assessments from autopsies. What ACT lacks, however, is anything more than rudimentary information (all collected retrospectively) about early life exposures, conditions, and experiences. We will link ACT records to MLP records from decennial censuses, vital records, and mortality files. The resulting linked files will be an unprecedented resource for aging research, enabling transformative analyses of how early demographic, socioeconomic, environmental, and policy factors shape disparities in adult cognition, the biology of brain aging, and ADRD risk.

Linking strategies for surveys of aging and ACT. Since 2016 we have successfully linked records from HRS, WLS, and NSHAP to the 1940 census; of survey participants alive as of 1940, we linked 48% of HRS records, 86% of WLS records, and 33% of NSHAP records. All linkages were accomplished using early versions of the machine learning algorithms described above. These surveys differ slightly with respect to the information available about sample members (e.g., for how many women they have last name at birth) and their co-residents at various timepoints (e.g., for how many cases they know parents' or siblings' names).

In the next five years, we will add health system linkages through the ACT study (Montine et al. 2012) and expand and improve linkages to HRS, WLS, and NSHAP in several ways. First, all linkages will employ improved state-of-the art, high precision and high recall machine learning methods as described above. Second, all will be linked to MLP and all its data elements—not just the 1940 census; panelists may thus be observed in every decennial census in their lifetimes, in vital records, and in mortality files. Third, both HRS and NSHAP have completed new rounds of survey data collection in recent years—and both added survey content designed explicitly to facilitate more and better linkages to MLP (e.g., by collecting women's names at birth, state of birth, parents' names). The rates at which we linked HRS, WLS, and NSHAP records to the 1940 census were very good, but we expect higher linkages rates going forward (without increasing Type I error rates).

MLP Linkage Hub

During the past five years, several teams of investigators around the country have developed linked data infrastructure leveraging the IPUMS full count census data. We plan to make these sources more easily interoperable with one another and with MLP. The MLP Linkage Hub will provide tools and documentation for using linked data from multiple sources. This will allow investigators to assess the robustness of their results to alternative linking algorithms, simplify replication of past research, and enrich the corpus of linked data. Our efforts fall into two broad categories. First, we will work with the Census Bureau to improve interoperability between MLP and the linked data infrastructure being developed within the Census Bureau, which will be disseminated through the network of Federal Statistical Research Data Centers (FSRDCs). Second, we will work with leading public data linking projects to simplify the use of their linked data in conjunction with MLP.

Census Bureau longitudinal infrastructure. Over the past two decades IPUMS collaborated with the Census Bureau to create a verified version of all the census microdata files held within the Census Bureau for the period since 1960, convert these data into harmonized IPUMS format, and disseminate that harmonized data through the FSRDCs (Ruggles et al. 2003, 2011).

In February 2014, representatives of the Census Bureau met with the IPUMS team in Minneapolis to discuss collaborative strategies for linking the IPUMS full count 1940 census to recent data sources housed in the Census Bureau, including the 2000 and 2010 censuses. The Census Longitudinal Infrastructure Project had the goal of linking the IPUMS 1940 full count census to the existing Census Bureau longitudinal infrastructure by adding Protected Identification Keys (PIKs) to the microdata records. Over the next two years, the 1940 census was successfully linked and made available through the network of Federal Statistical Research Data Centers (FSRDCs) (Alexander et al. 2015; Massey et al. 2018).

A major limitation of this project was the lack of linked observations for censuses between 1940 and 2000, a period when respondent names were not originally digitized. A new collaborative effort, the Decennial Census Digitization and Linkage project (DCDL), is now filling the six-decade gap in observations by using automatic handwriting recognition technology to capture approximately 850 million names from the original microfilm of the 1960 to 1990 census returns (Genadek and Alexander 2022; Alexander, Fisher, and Genadek 2022; Alexander and Genadek 2022). IPUMS will contribute an additional 152 million observations from the 1950 census. Leveraging administrative records from Social Security and the Internal Revenue Service, DCDL will add Protected Identification Keys (PIKs) to each record, providing linked observations across all censuses from 1940 to 2020 that will be available through the FSRDCs. MLP will work closely with DCDL to ensure that the entire collection of MLP data—including the linked censuses from 1850 to 1950—can be easily used in conjunction with more recent linked censuses within the FSRDC network. The entire series of MLP/DCDL censuses includes approximately 2.6 billion individual-level observations; used in conjunction with a vast array of administrative and survey data that also have PIKs, the MLP/DCDL data will create an extraordinary range of new opportunities for life-course analysis of health.

We will collaborate with the Census Bureau to develop a crosswalk between the Census Bureau's PIKs and the MLP Historical Identification Keys (HIKs). We also aim to fully document the linked data and to make that documentation accessible through a public-facing system to allow researchers to assess project feasibility and prepare proposals for FSRDC access. There is presently minimal public-facing documentation for the linked IPUMS-format data resources available through the FSRDC network. The internal documentation is available only to researchers who have already been approved to use the data, making it difficult to formulate proposals to gain data access. Moreover, the internal documentation is not in a standardized format and provides no information about the variables except their codes. There is no information on comparability over time, universe, questionnaire text, or frequencies. Accordingly, researchers must make proposals for access without knowing the full content of the data they are proposing to use, and once they are approved to use the data, they must cope with documentation that is cumbersome and incomplete. In addition to creating this badly needed public-facing documentation, we will also work with Census Bureau staff to develop tools and procedures to simplify using MLP in conjunction with linked internal censuses, surveys, and administrative data within the FSRDC network.

To better understand the comparability of MLP with the Census Bureau's linked data, we also plan an analysis of the Numident files. The Census Bureau's linking procedure is entirely based on an internal version of the Social Security Numident file. This internal file is more comprehensive than the public version of the file that we are using to link death certificates to MLP, in part because it includes persons who are still living or who died after 2007. There are apparently additional omissions in the public file prepared by the National Archives and Records Administration, but documentation is scant. Any systematic omissions in the public Numident could have implications for analyses that rely on that data. We will work with the Census Bureau to compare the two versions of the Numident and fully document the differences.

Public linking partners. Over the past five years, four other teams of researchers have established links to IPUMS full count census microdata. The LIFE-M project linked the IPUMS censuses of 1880 through 1940 to birth, death, and marriage certificates from Ohio and North Carolina (Bailey et al. 2023). The CenSoc Project linked the Social Security Death Master File (DMF) and the public Numident to the IPUMS census of 1940 (Goldstein et al 2021). The Census Linking Project (CLP) linked individuals from census to census between 1850 and 1940 (Abramitzky et al 2020, 2021). The Census Tree project combines links from MLP and CLP with genealogical data to improve linkage rates (Buckles et al 2023). Our goal is to develop systems that will make it easy to incorporate new linked data sources as they become available, and to make it easy to use multiple linked sources in combination.

- *The Longitudinal, Intergenerational Family Electronic Micro-Database (LIFE-M)* is linking 20th-century state vital registration data with census data. Beginning with a set of birth certificates from the first four decades of the 20th century, LIFE-M links to siblings' birth records, marriage records, and death records, matching on names and birthplaces as well as parental names and birthplaces. The birth records and marriage records are not available from any other source, and the death records in LIFE-M generally derive from an earlier period than alternative sources of mortality information. Because the records include women's birth names as well as their married names, these data are invaluable for tracing women over the life course. The reconstituted families are then linked to MLP census data; LIFE-M currently includes 5.4 million links to MLP. The LIFE-M data are currently limited to vital records from Ohio and North Carolina, but additional states are planned; approximately 185 million state vital records from the 20th century have been digitized and are candidates for inclusion in the database.

  The MLP data access system already identifies cases that link to LIFE-M. Accordingly, users can request a data extract that includes only individuals (and their families, if desired) that appear in the LIFE-M database, dramatically reducing the work needed to match the databases and capitalize on LIFE-M variables. We will incorporate new additions to the LIFE-M database and work with LIFE-M to define new summary variables, such as dates of birth, marriage, and death and number of children born.

- *The CenSoc Project* linked the Social Security Death Master File (DMF) to the 1940 census. The DMF includes 83 million deaths from Social Security records of deceased persons, which represents approximately 93 to 96% of deaths reported after 1973 for persons age 65 or older (Hill and Rosenwaike 2002). As described above, the IPUMS team plans to link death certificate data from the National Vital Statistics System to MLP, leveraging the public Social Security Applications file (Numident) and the National Death Index. For many cases we will have comparable information from our Death Certificate linkage, but the results will differ both because of varying coverage of the two sources and because of differences in the linking methodology. Having measures available from both sources will greatly simplify comparisons and allow researchers to test the robustness of their findings under different linking approaches. For many studies of mortality disparities and neighborhood-level determinants of mortality, it will be useful to combine information from both sources. We will therefore make the CenSoc data directly available through MLP alongside the death certificate data.
- *The Census Linking Project (CLP)* has constructed links across censuses from 1850 to 1940. The CLP linking methods, which apply a deterministic algorithm limited to invariant characteristics, differ significantly from those used by MLP. MLP performs better than CLP with respect to both Type I and Type II errors (Helgertz et al. 2022; Ghosh, Hwang, and Squires 2023), but CLP may still be preferred for some kinds of analyses (Ruggles 2002). Specifically, because the MLP linking algorithm uses characteristics of households, it disproportionately links people who remained in the same household across periods. Although weighting can restore balance in the observable composition of the linked panel, the data may still be

unbalanced in terms of unobservable characteristics. The CLP linkage strategy is based on immutable characteristics, and therefore avoids these kinds of biases. To simplify making comparisons between the two linking approaches, we will add variables to MLP identifying cases that are linked according to each of the seven available alternative linking methods provided by CLP.

- *Census Tree* is an exciting new initiative (first released in September 2023) that builds on MLP and CLP links by adding crowd-sourced information from genealogical sources, their own linking algorithm using XGBoost, and proprietary profile hints generated by FamilySearch. By combining information from multiple sources, Census Tree significantly improves recall without degrading precision and representativeness of the links. We will assess the Census Tree links and explore the potential of using this innovative approach to improve the quality of the core MLP intercensal links.

In addition to creating new variables and associated documentation, the linking hub will include guidance for users by describing the methods used to establish links, the potential pitfalls for particular kinds of analyses, and the appropriate weighting procedures when combining linked data from multiple sources.

Infrastructure Improvement and Dissemination

We plan to develop a sustainable framework for individual-level and family-level linked records in the United States that can be continuously expanded and improved to incorporate new data sources as they become available. Meeting this goal will require continued refinement of our record-linkage software infrastructure, new software to automate maintenance and updating of the infrastructure, and new dissemination tools and activities to make MLP readily accessible to the research community.

Refinement of Hlink. Our Hlink package provides a flexible end-to-end solution for large-scale probabilistic record linking at scale. Hlink supports the entire linking workflow, including preprocessing, filtering, training, model exploration, blocking, feature generation, and scoring (Helgertz et al. 2022). We use a distributed processing engine to spread the workload across a cluster of computers, and we leverage a column-store database to optimize processing speed.

- *Performance*. A major goal is to further improve performance of the Hlink package. As described above, Hlink uses blocking strategies to reduce the computational load, which means that some true matches are never considered. By further improving performance we will be able to consider a broader range of potential matches, increasing the number of true matches and reducing the number of incorrect links. We will experiment with a variety of strategies to improve performance, focusing on massive parallelization. An especially promising path is algorithms that exploit graphics processing units (GPUs) (Li et al. 2015; Chen et al. 2017).
- *Machine-learning algorithms*. Research in artificial intelligence is progressing rapidly. We believe that the XGBoost machine-learning framework currently provides the highest accuracy for cross-census record linkage. Hlink is designed to accommodate a wide range of linking algorithms. We will evaluate alternative approaches to maximize accuracy and efficiency, including CatBoost (Prokhorenkova et al. 2018; Bentéjac, Csörgő, and Martínez-Muñoz 2021)
- *User interface.* Hlink is designed to be operated by subject-matter experts with minimal coding skills or expertise in artificial intelligence. The system can be controlled through Python via an Application Programming Interface or through a standalone command line interface. We will refine the Hlink command-line interface to simplify use and expand flexibility.

Software and workflows for maintenance and updates. We have largely automated our processes for record linkage, but MLP still relies on manual processes for maintaining and updating the system. The system of links is constantly improved, either because we gain access to cleaner or more detailed source data or because we refine the linking procedures. In some instances, making a correction can have ripple effects through multiple sources. Under our current workflow, updating MLP takes a month and requires coordination of multiple experts to ingest new data and generate updated links throughout the system. As we expand the system to incorporate additional sources, this burden will increase. Accordingly, to make MLP sustainable we must automate as much of the process as possible.

We will automate the pipeline by developing new tools and workflows to maintain and update MLP. We envision a cohesive system for generating a master link file with automated end-to-end propagation of updates through the entire system. We will develop tools to synchronize this work and document changes to data and metadata, allowing us to coordinate multiple streams of work. We will extend our data production workflow and utilize version control tools for metadata, allowing us to integrate independent streams of work into a single copy when appropriate. We will implement project management software for documenting higher level issues, processes, and procedures. The metadata versioning system will complement our existing software versioning system, allowing us to roll back to earlier versions. This approach will ensure a complete, permanent record of the evolution of the MLP data, which is essential for supporting replication of findings.

Dissemination tools and activities. Data sharing is central to the project: effective dissemination is essential if the data are to be widely used. The linked data and documentation from the 1850-1950 censuses and the public use version of the death certificate data, as well as the links to LIFE-

M, CenSoc, CLP, and Census Tree will all be made freely available to the public through the IPUMS web-based data access system. Other components of the MLP infrastructure will be restricted access, controlled by our partners. NCHS will disseminate a restricted version of the death certificate data. The FSRDC network will disseminate the data from recent censuses, surveys, and administrative data linked to MLP. Access to the MLP-linked survey data—HRS, WLS, and NSHAP—will be controlled by each survey organization. Finally, access to the linked ACT clinical and hospital data will be controlled by Kaiser. Each of these organizations has their own protocols for gaining access; we will develop public-facing guidance and individualized support for users to help them overcome procedural hurdles.

Because of the large scale of the public data, efficient subsetting and data manipulation capabilities are essential. We plan additional capabilities for merging data from multiple sources and constructing customized variables that capitalize on the longitudinal structure of the data. We disseminate user-specified customized datasets as ASCII text files and in the proprietary formats of the major statistical packages (Stata, SAS, SPSS, and R), together with customized codebooks in Data Documentation Initiative (DDI) structured XML format. As with other IPUMS data collections, we will provide online data analysis tools as well. We also plan to release all software used for the project under an open-source license, both for purposes of documentation and to enable researchers to apply our linking methods to other datasets**.**

The substantial investment in MLP is only justified if the data are widely used to produce new discoveries; accordingly, effective outreach is essential. We will develop training materials for new user-facing features of the infrastructure, including specialized documentation (user guidance notes, how-to blog posts, website FAQs, and help text); tutorials (video demonstrations and sample exercises); curricular materials to aid in using IPUMS for teaching; and training (webinars and workshops). We will also provide direct support to individual data users through our User Support team, exhibit hall booths at academic conferences, and Virtual Office Hours, which allow users to connect with IPUMS staff via video conferencing.

Early Impact of MLP

Although most MLP resources have been available to the research community for just two or three years, they have already yielded dozens of papers on the impact of early life social, economic, policy, and environmental exposures on later life cognitive, health, and economic outcomes. Consider these diverse examples, all using MLP data and appearing in high-profile journals:

- *Impact of childhood lead exposure on late life cognitive functioning*. Leveraging the well-known relationship between water pH and its plumbosolvency, Lee and colleagues showed that older adults who as children lived in cities with lead pipes and acidic or alkaline water—the conditions required for lead to leach into drinking water—had substantially worse baseline cognitive functioning at age 72 (Lee et al 2022a). Children who lived near a lead mine in 1940 had both lower cognitive functioning at age 64 and steeper cognitive decline at older ages (Lee et al. 2022b).
- *Long-run effects of Depression-era work-relief programs on later life health and well-being.* Modrek et al. (2022) evaluated the impact of New Deal work relief programs on long-term outcomes of persons aged 0-3 in 1940. They found that work relief programs boosted adolescent IQ scores, improved class rank, increased attainment of bachelor's degrees, raised midlife income, and led to better late-life cognition.

- *Impact of childhood nutrition on early adult economic outcomes.* Roberts et al. (2023) assessed the relationship between young boys' biometric measures in the early 20th century and subsequent economic outcomes, leveraging sibling pairs to control for unobserved family characteristics. They found that being taller and heavier in childhood is strongly associated with employment and earnings in adulthood, underscoring the importance of childhood nutrition.
- *Influence of kin proximity on demographic outcomes.* Hacker et al. (2021) found that the presence of both coresident and proximate elder kin (mainly parents) was strongly associated with fertility, child survival, and net reproduction. A dramatic long-run decline in proximate elder kin may help to explain declining fertility.

These articles are just the tip of the iceberg of transformative scholarship resulting from existing MLP resources; many other exciting MLP-based projects are underway. To date, we have identified 23 peer-reviewed articles, 12 Ph.D. dissertations, and 3 NIA-funded R01 research projects that apply MLP data. Google Scholar lists over 90 items—mainly working papers—citing MLP, so we anticipate many more publications will appear over the next few years.

Considering the length of the publication cycle and the scale and complexity of the data, the early record of MLP usage is strong. Indeed, the uptake of the MLP infrastructure by the research community is among the most rapid of any IPUMS data infrastructure product of the past three decades. Although we released a preliminary beta-test version of crosswalks between some censuses in August 2020, the broad public rollout of the MLP intercensal links came in early 2022 (Helgertz et al 2022), and that version still required massive computing power and substantial coding expertise. It was not until February 2023 that MLP links were incorporated into the IPUMS data dissemination system, making the infrastructure easily accessible to a broad audience of researchers.

MLP is a direct outcome of three decades of consistent federal investments in developing large-scale microdata resources to uncover the dimensions of demographic and behavioral change. Through strategic investment in research infrastructure, we seek to reduce redundant effort, shrink the expense of life-course and multigenerational research, and improve the reproducibility of research results. By leveraging billions of dollars of federal investments in census and survey data collected over the course of 170 years and combining them with vital records and administrative sources, we are creating powerful new resources for historical analysis. Consistently linked microdata that include information on the characteristics of the entire population over many generations will make a permanent and substantial contribution to the nation’s statistical infrastructure.

References Cited

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J. and Pérez, S. 2021. Automated linking of historical data. *Journal of Economic Literature* 59(3): 865-918.

Abramitzky, R., Boustan, L., Eriksson, K., Pérez, S. and Rashid, M. 2020. Census Linking Project: Version 2.0 [dataset]. https://censuslinkingproject.org

Alexander, J.T., Fisher, J.D., and Genadek, K.R. 2022. Digitizing hand-written data with automated methods: A pilot project using the 1990 US census. *Journal of Economic and Social Measurement* 46(2): 95-108. https://content.iospress.com/articles/journal-of-economic-and-social-measurement/jem220484

Alexander J.T., Gardner, T., Massey, C.G. and O'Hara, A. 2015. Creating a longitudinal data infrastructure at the Census Bureau. CARRA Working Paper 2015-1, U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2015/adrm/2015-alexander.pdf

Alexander, J.T. and Genadek, K.R. 2022. Using administrative records to support the linkage of census data: protocol for building a longitudinal infrastructure of US census records. *International Journal of Population Data Science* 7(4). https://ijpds.org/article/view/1764

Anbinder, T., Connor, D., Ó Gráda, C. and Wegge, S. 2021. The Problem of False Positives in Automated Census Linking: Evidence from Nineteenth-Century New York's Irish Immigrants. UCD Centre for Economic Research Working Paper Series, WP2021/14. University College Dublin. School of Economics. http://hdl.handle.net/10197/12278

Bailey, M.J., Cole, C., Henderson, M. and Massey, C., 2020. How well do automated linking methods perform? Lessons from US historical data. *Journal of Economic Literature* 58(4): 997-1044.

Bailey, M., Lin, P.Z., Mohammed, A.S., Mohnen, P., Murray, J., Zhang, M. and Prettyman, A., 2023. The creation of LIFE-M: The longitudinal, intergenerational family electronic micro-database project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56(3):138-159.

Baskerville P., Dillon L., Inwood K., Roberts E., Ruggles S., et al. 2014. Mining microdata: economic opportunity and spatial mobility in Britain and the United States, 1850–1881. In *IEEE International Conference on Big Data*, 5–13. Washington, DC: IEEE.

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* 54: 1937-1967.

Buckles, K., Haws, A., Price, J., and Wilbert H.E.B. 2023. Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project. NBER Working Paper 31671.

Chen, T. and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* 785-794.

Centers for Disease Control and Prevention. 2022. Historical Mortality File, 1979-2007. National Center for Health Statistics, National Vital Statistics System.

Chen, C., Li, K., Ouyang, A., Tang, Z. and Li, K. 2017. Gpu-accelerated parallel hierarchical extreme learning machine on flink for big data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47(10): 2740-2753.

Christen P. 2012. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering* 24(9): 1537-1555.

Curti M. 1959. *The Making of an American Community: A Case Study of Democracy in a Frontier County*. Stanford, CA: Stanford University Press

Fellegi, I.P. and Sunter A.B. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association* 64: 1183-1210.

Ferrie, J.P. 1996. A new sample of males linked from the public use microdata sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 29(4): 141–56.

Genadek, K.R. and Alexander, J.T. 2022. The missing link: Data capture technology and the making of a longitudinal US Census infrastructure. *IEEE Annals of the History of Computing* 44(4): 57-66.

Ghosh, A., Hwang, S.I.M., and Squires, M. 2023. Evaluating automated linking algorithms without ground truth: A case study of the U.S. Census 1930-1940. Working Paper, University of British Columbia. https://www.dropbox.com/s/xe8cd5h8brj3z04/ghs_final_jbes_wo_fig_tab_longer_version.pdf?dl=0 https://www.dropbox.com/s/0lb5srybihtl7ee/ghs_final_jbes_tables_and_figures_longer_version.pdf?dl=0

Goeken R., Huynh L., Lynch, T.A., and Vick, R. 2011. New methods of census record linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44(1):7–14.

Goldstein, J.R., Alexander, M., Breen, C., González, A.M., Menares, F., Osborne, M., Snyder, M. and Yildirim, U. 2021. CenSoc Mortality File: Version 2.0. Berkeley: University of California.

Guest, A.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.D., Helgertz, J., Nelson, M.A. and Roberts, E., 2021. The influence of kin proximity on the reproductive success of American couples, 1900–1910. *Demography* 58(6): 2337-2364.

Helgertz, J., Ruggles, S. Warren, J.R., Fitch, C.A., Goeken, R., Hacker, J.D., Nelson, M.A., Price, J.P., Roberts, E., Sobek, M. IPUMS Multigenerational Longitudinal Panel: Version 1.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D016.V1.0

Helgertz, J., Price, J.P., Wellington, J., Thompson, K.J., Ruggles, S. and Fitch, C.A. 2022. A new strategy for linking US historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55(1): 12-29.

Helgertz, J., Ruggles, S., Warren, J.R., Fitch, C.A., Hacker, J.D., Nelson, M.A., Price, J.P., Roberts, E. and Sobek, M. 2023. IPUMS Multigenerational Longitudinal Panel: Version 1.1 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D016.V1.1

Hill, M.E. and Rosenwaike, I. 2002. The Social Security Administration's Death Master File: The completeness of death reporting at older ages. *Social Security Bulletin* 64(5): 45-51.

IPUMS. 2022. hlink: Hierarchical record linkage at scale. https://github.com/ipums/hlink#readme

Katz, M.B. 1975. *The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth-Century City*. Cambridge, MA: Harvard University Press

Lee, H., Lee, M.W., Warren, J.R. and Ferrie, J. 2022a. Childhood lead exposure is associated with lower cognitive functioning at older ages. *Science Advances* 8(45). DOI: 10.1126/sciadv.abn5164

Lee, M., Lee, H., Warren, J.R. and Herd, P. 2022b. Effect of childhood proximity to lead mining on late life cognition. *SSM-Population Health* 17:101037. https://doi.org/10.1016/j.ssmph.2022.101037

Li, P., Luo, Y., Zhang, N. and Cao, Y. 2015. Heterospark: A heterogeneous cpu/gpu spark platform for machine learning algorithms. *IEEE International Conference on Networking, Architecture and Storage (NAS)*, 347-348.

Long, J. and Ferrie J. 2013. Intergenerational occupational mobility in Great Britain and the United States since 1850. *American Economic Review* 103(4): 1109–37.

Malin, J.C. 1935. The turnover of farm population in Kansas. *Kansas Historical Quarterly* 4: 23–49, 164–87.

Massey, C.G. 2017. Playing with Matches: An Assessment of Accuracy in Linked Historical Data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50(3): 1-15. https://doi.org/10.1080/01615440.2017.1288598

Massey, C.G., Genadek, K.R., Alexander, J.T., Gardner, T.K. and O'Hara, A. 2018. Linking the 1940 U.S. Census with modern data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51:4: 246-257. DOI: 10.1080/01615440.2018.1507772

Modrek, S., Roberts, E., Warren, J.R. and Rehkopf, D. 2022. Long-Term effects of local-area new deal work relief in childhood on educational, economic, and health outcomes over the life course: evidence from the Wisconsin longitudinal study. *Demography* 59(4): 1489-1516.

Montine, T.J., Sonnen, J., Montine, K., Crane, P. and Larson, E., 2012. Adult Changes in Thought study: Dementia is an individually varying convergent syndrome with prevalent clinically silent diseases that may be modified by some commonly used therapeutics. *Current Alzheimer Research* 9(6): 718-723.

National Archives and Records Administration. 1936-2007. Numerical Identification Files (NUMIDENT). Record Group 47, Records of the Social Security Administration. Electronic and Special Media Records Services Division, College Park, MD.

Owsley, F.L. and Owsley, H.C., 1940. The economic basis of society in the late ante-bellum South. *The Journal of Southern History* 6(1): 24-45.

Porter, E.H. and Winkler, W.E. 1997. Approximate String Comparison and its Effect on an Advanced Record Linkage System. Census Bureau Research Report RR97/02. Washington: U.S. Census Bureau.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada. https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

Roberts, E., Helgertz, J. and Warren, J., 2023. Childhood growth and socioeconomic outcomes in early adulthood evidence from the inter-war United States. *The History of the Family* 28(2): 229-255. https://doi.org/10.1080/1081602X.2022.2034658

Ruggles S. 2002. Linking historical censuses: A new approach. *History and Computing* 14: 213-24.

Ruggles, S. 2023. Collaborations between IPUMS and genealogical organizations. *Historical Life-Course Studies* 13: 1-8. https://doi.org/10.51964/hlcs12920

Ruggles, S., Fitch, C. and Roberts, E. 2018. Historical census record linkage. *Annual Review of Sociology* 44: 19-37.

Ruggles, S., Sobek, M., King, M.L., Liebler, C. and Fitch, C.A. 2003. IPUMS redesign. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 36(1): 9-19.

Ruggles, S., Schroeder, M., Rivers, N., Alexander, J.T. and Gardner, T.K., 2011. Frozen film and FOSDIC forms: Restoring the 1960 US Census of Population and Housing. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44(2): 69-78.

Steckel, R. 1988. Census matching and migration: A research strategy. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 21(2): 52-60.

Synced. 2017. Tree Boosting with XGBoost – Why Does XGBoost Win "Every" Machine Learning Competition? *Synced AI Technology and Industry Review.* https://syncedreview.com/2017/10/22/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/

Thernstrom S. 1964. *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Cambridge, MA: Harvard University Press.

Warren, J.R., Lee, M. and Osypuk, T.L., 2022. The validity and reliability of retrospective measures of childhood socioeconomic status in the health and retirement study: Evidence From the 1940 US Census. *The Journals of Gerontology: Series B* 77(9): 1661-1673.

Warren, J.R., Pfeffer, F.T., Helgertz, J. and Xu, D. 2020. Linking 1940 U.S. Census Data to the Health and Retirement Survey: Technical Documentation – Release 1. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. https://hrs.isr.umich.edu/sites/default/files/restricted_data_docs/HRS-1940-Census-Data-Documentation-Report.pdf