UNIVERSITY OF MINNESOTA

Bricks without Straw: The use of linked census data to estimate child mortality in the pre-registration era of the United States

J David Hacker[1]
Department of History,
University of Minnesota

Jonas Helgertz
Minnesota Population Center, University of Minnesota
Institute for Social Research and Data Innovation, University of Minnesota
Department of Economic History and Centre for Economic Demography,
Lund University

March 2024

Working Paper No. 2024-01
DOI: https://doi.org/10.18128/MPC2024-01

[1]Address correspondence to J David Hacker (email: hacker@umn.edu) or Jonas Helgertz (email: jonas.helgertz@ekh.lu.se). This project was supported by a research grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01-HD082120-01Q3) and the Minnesota Population Center (P2C HD041023).

Bricks without Straw: The use of linked census data to estimate child mortality in the pre-registration era of the United States

March 13, 2024

J David Hacker (hacker@umn.edu) [1, 2, 3]
Jonas Helgertz (jonas.helgertz@ekh.lu.se) [2, 3, 4]

[1] Department of History, University of Minnesota
[2] Minnesota Population Center, University of Minnesota
[3] Institute for Social Research and Data Innovation, University of Minnesota
[4] Department of Economic History and Centre for Economic Demography, Lund University

Abstract

A national death registration system was not started in the United States until 1901 and not completed until 1933. As a result, the onset and early decades of the mortality transition are poorly understood. In this article we describe an indirect method to estimate white and black child mortality using linked census data for each decade 1850-1880 and 1900-1940. We compare the estimates to several external sources. For each decade, we aggregate the number of child deaths and survivors to the national level and compare the implied level of mortality to published life tables. For the period 1900-1910, we compare our estimates to alternative estimates using the number of children ever born and number of children surviving reported by women in the 1910 census and compare models of child mortality using both estimation methods. For the period 1900-1940, we compare random samples of children to data on year of death provided from FamilySearch.org. We conclude that the new method represents a viable way to estimate and model child mortality over the course of the mortality transition. In the final section of the paper, we illustrate a few applications of the new method. We map mortality estimates by race and decade, convert the estimates to abridged life tables using model life tables, construct a public-use dataset based on the results, and describe possible applications.

Introduction

Prior to the onset of the mortality transition circa 1875, approximately one-in-four children in the United States died before reaching their fifth birthday. Fifty years later, thanks in large part to the implementation of a range of public health measures, the ratio had fallen to less than one-in-ten (Arias 2014; Cutler, Deaton and Lleras-Muney 2006; Haines 1998). Despite its importance, the onset and first half century of the mortality transition are poorly understood. We lack a good understanding of the contribution of individual and contextual factors to mortality decline and changes in their relative importance over time. Our knowledge is especially lacking in the period before 1900, when a "national" Death Registration Area (DRA) was first established for 10 states and the District of Columbia. It is also lacking, although to a lesser extent, in the period between 1900 and 1933, when states were being added to the national DRA.

In this paper we describe and evaluate a method of estimating child mortality at the individual level using millions of linked census records for the period between 1850 and 1940. Our primary data source is the new IPUMS Multigenerational Longitudinal Panel (IPUMS MLP) datasets (Helgertz et al. 2023). The IPUMS MLP is composed of individuals in the IPUMS full-count census datasets of the United States (Ruggles et al. 2021) linked between the 1850-1860, 1860-1870, 1970-1880, 1900-1910, 1910-1920, 1920-1930, and 1930-1940 censuses (Helgertz et al. 2023; Ruggles et al. 2021).[1] We supplement the IPUMS MLP datasets with links recently published by the Census Tree Project (Buckles et al. 2023), which uses additional sources to establish probable links of the same individual across multiple censuses. These data are used to infer child deaths and survivors between each pair of decennial censuses and to identify census undercounts.

There are many challenges to inferring child survival and mortality using linked datasets, not the least of which are well-known quality issues associated with early census enumerations—undercounting, age reporting errors, inconsistent spelling of names, and a variety of errors in other variables needed for

[1] Unfortunately, the manuscript records of the 1890 census were destroyed by fire, preventing the construction of IPUMS MLP linked datasets for the decades 1880-1890 and 1890-1900.

linking (e.g., birthplace, sex, race, etc.), so we proceed cautiously. Census quality improves over time, but child mortality declines as well, perhaps making census quality a greater challenge in estimating child mortality in the latter decades of our study than in the early decades. We are, in effect, trying to measure a signal in the presence of significant noise, which is more difficult as the signal strength wanes.

Unfortunately, some of the motivations for conducting this study—the lack of death records in the United States history prior to the completion of the DRA in 1933 and the lack of digitized death records linked to socioeconomic information for the period in which death records exist—make it impossible to verify our estimates in most decades and difficult in others. We suspect that our selection criteria—we limit our analysis to children of parents who were linked between adjacent censuses—biases child mortality estimates downwards. This bias, however, is likely offset to some degree by an unknown number of surviving children who, for whatever reasons, we were unable to track between subsequent censuses and incorrectly assumed to have died. Although we contend that these biases were likely modest and offsetting, their probable existence makes it imperative that the results are both plausible and verified where possible.

We evaluate the quality of our mortality estimates in several ways. We begin by comparing our estimates of the proportion of children dying in each year to published national life tables. Next, we compare a random selection of children in our twentieth-century datasets to mortality data maintained by FamilySearch, yielding very encouraging results Finally, we make several comparisons of our estimates in the period 1900-1910, the approximate middle of the period we study, using information collected on the number of women's children ever born (CEB) and the number of those children surviving (CS) in the 1910 census and now available in the IPUMS full-count dataset (Ruggles et al. 2022). We conclude that our method of estimating child mortality using linked census data represents a viable approach to measuring and modeling child mortality in the pre-vital registration era of the United States.

Because no standard for comparison exists, we have not attempted to verify our census undercount estimates. The undercount estimates agree, however, with published age distributions of the population and

qualitative assessments, which suggests that infants and young children were particularly likely to be missed by historical censuses and that black children were more likely to be undercounted than white children.

After evaluating the quality of our estimates, we aggregate the number of at-risk, deceased, and undercounted children aged 0-5 by race, sex, and decade at five different levels of geography: county, state economic area (SEA), state, census division, and nation. We calculate ten-year cohort mortality and survivorship rates, convert these rates to conventional life table parameters, and construct complete life tables with the assistance of model life tables. These estimates, we believe, represent a major contribution to our understanding of the timing and spatial variations in the demographic transition in the United States and will prove to be a valuable resource for other scholars needing contextual information on mortality, such as researchers interested in the effects of early life conditions on population health in older ages, mapping spatial variations in mortality, making forward and backward population projections, and for adjusting own-child fertility estimates (e.g., Dwyer-Lindgren et al. 2017; Elman et al. 2023; Hacker 2016). As a simple illustration of the potential of these data, we map white and black child mortality estimates by county in the period 1900-1910 and briefly discuss the results.

Background

The United States was late to establish a national system of vital registration. Although a few states and municipalities began registering births and deaths in the mid nineteenth century, the national Death Registration Area (DRA) was first established in 1900 and initially included just 10 states and the District of Columbia, which together represented just 26.3% of the nation's population. Initially, the DRA was unrepresentative of the national population, being predominantly composed of states in the Northeast census region that were more urban and industrial than non-DRA states. DRA states were also characterized by significantly higher proportions of foreign-born residents, lower proportions of black residents, and lower fertility rates than non-DRA states. States were added to the DRA over time, making it more

representative of the nation, but it was not completed until 1933, when Texas was added. Even then a significant percentage of deaths are believed to have been unregistered (Haines 2000; Preston and Haines 1991: 49-50).

Because of the lateness and deficiencies of vital registration, our understanding of the mortality and transition in the United States rests on a weak empirical foundation. Researchers of the demographic transition have needed to be creative, constructing demographic rates using limited and unrepresentative vital registration data, indirect estimation methods based on census data documenting the living population, or from studies of genealogical records and special populations, including retrospective mortality information collected with the 1850-1900 censuses and CEB and CS data collected for each mother in the 1900 and 1910 censuses (e.g.,Dribe, Hacker and Scalone 2020; Haines 1998; Pope 1992).

Michael R. Haines' (1998) life tables for the years preceding the 1850, 1860, 1870, 1880, 1890, and 1900 censuses were based on retrospective mortality censuses conducted alongside the population censuses. Households enumerated in the population census were to report any household members who had died in the year prior to the census, and these decedents were recorded on a separate schedule. These mortality censuses clearly under-reported deaths—perhaps as much as 40% or more overall—and the year in which mortality was recorded may have not been representative of the decade. Haines, however, relied solely on children aged 5-14 dying in the year prior to each census, who appeared to be reasonably well reported. He then fitted the age-specific mortality estimates for children aged 5-14 to model life tables to infer estimates to all age groups. Hacker (2010) also used model life tables to construct nineteenth-century life tables, fitting adult life expectancy estimates at age 20 from published genealogical studies to a relational model life table system based on life tables constructed for the 1900-1902 DRA (Glover 1921). No attempt was made by Haines or Hacker to estimate regional or sub-regional estimates of mortality. Both sets of life tables were for the white population only, although Haines included a few life tables for the black population at the turn of the twentieth century based on his analysis with Samuel Preston of CEB and CS data in the 1900 and 1910 censuses (Haines and Preston 1997; Preston and Haines 1991).

Preston and Haines used a low-density public use sample constructed for the 1900 census to analyze CEB and CS collected by that census for ever-married women. These data allowed them to estimate child mortality and construct empirical models of mortality with suspected covariates (Preston and Haines 1991). CEB and CS data were also collected in the 1910 census and have been analyzed by researchers using higher-density and full-count public use microdata samples of the 1900 and 1910 censuses in similar ways (Dribe et al. 2020; Preston et al. 1994).

Identifying Child Deaths and Undercounts

We use individuals linked between US censuses 1850-1940 by the IPUMS MLP (Helgertz et al. 2020) and Census Tree (Buckles et al. 2023) – hereafter referred to as CT – projects to identify mortality and undercounts. MLP links were made between adjacent censuses using a supervised machine-learning algorithm and individual time-invariant (e.g., given name, surname, year of birth, place of birth, and race) and time-variant information that tends to change slowly from decade to decade (e.g., spouses and other household members). As shown elsewhere (Helgertz et al. 2022), the use of household-level information in the construction of the MLP datasets results in high linkage rates and low shares of false positives, which are critical in measuring demographic events with relatively low rates of occurrence such as mortality. These links are supplemented with links made using any of the three CT methods that were associated with the highest precision, namely "Family Tree", Family Search Direct Hint" and "Family Search Profile Hint."

We restrict the MLP datasets to successfully linked couples, then develop an algorithm to infer the survival or mortality of children aged 0-5 in the first census based on the children present in the household in the second census. This procedure is supplemented with newly released data from the CT Project, primarily through the identification of survival through links to censuses further forward in time. Finally, we estimate census under-enumeration in the first census from children present in the second census who were not linked back to a record ten years earlier despite being alive at the time of the first census. Hereafter,

when discussing individuals linked across adjacent censuses, we will consistently refer to the first census as "census A" and the second census as "census B."

The method developed in this paper identifies child deaths and census undercounts across the decennial census pairs between 1850-1880 and between 1900-1940.[2] While the method is computationally demanding as it involves a substantial number of record comparisons, the underlying logic is rather straightforward. Initially, we only use MLP links to select coresident couples in census A – defined through being married or through the presence of joint children – where both partners were successfully linked to and co-residing in census B. Naturally, imposing the restriction that both partners were successfully linked means that we exclude couples where one or both partners die, emigrates, or for whatever other reason are not linked. Overall, between 34.1 (1850) and 46.6 (1930) percent of census A couples were linked to the subsequent census, with differences over time consistent with declining mortality and improving census enumeration (see table A1). Differences between the linked population and the baseline population of couples were generally small and consistent with expected differences in mortality (e.g., the greater linkage of white couples relative to Black couples) and linkage rates more generally (e.g. geographically immobile groups such as farmers (owners/managers/foremen) were linked more frequently than more mobile groups, although this difference might also reflect differences in adult mortality). Although there is no way to conclusively identify "potentially linkable" couples (i.e. surviving and enumerated in census B) using only the cross-sectional datasets, period estimates of ten-year survival rates and census coverage errors suggest that the IPUMS MLP project is linking approximately two-thirds of the potentially linkable couples (see discussion in Hacker, Dribe and Helgertz 2023).[3]

[2] Unfortunately, the original manuscript returns from the 1890 Census manuscripts do not survive and there is therefore no corresponding 1890 IPUMS full-count dataset. Although the IPUMS MLP project provides links for individuals across the twenty-year interval 1880-1900, the interval is too large to estimate child mortality.
[3] We constructed a similar table to Appendix Table A1 for children by comparing the universe of children aged 0-5 in census A who are the children of successfully linked parents to all children aged 0-5 in census A (Appendix Table A2). Overall, 36.2-56.5 percent of all children belong to the population at risk, unsurprisingly with similar patterns of selectivity as for the parents.

Our population of interest is children aged 0-5 in census A. Our method intends to identify both those who survived until the subsequent census, when they were age 10-15, and those who died sometime between the two censuses. Our limitation to younger children is motivated by the expectation that children with surviving parents would almost invariably continue to co-reside with their parents until at least age 15. This is important because children who left their parental homes to live elsewhere are more difficult to link and therefore at greater risk of being considered by our method to be deceased. Consequently, there is greater risk of overestimating mortality among older children, who were more likely to have left home than their younger siblings. This is not to say that the percentage of older teenage children who left their parents' homes was large. While there are exceptions to central tendencies, such as girls marrying at a young age or leaving the family home at an early age to board in another household, the typical age of leaving home was well above 20 in the United States during the nineteenth and early twentieth centuries. Steckel (1996), for example, estimated that the median age that white males left home was 25.2 years in 1860 and 24.4 years for white females. Additionally, the census data confirm that the most precipitous decline in the share of individuals residing with at least one parent occurred after age 15. Published estimates of adult mortality, moreover, can account for the small percentages of children were not residing with their parents before age 15.[4] Nevertheless, because ten-year mortality rates are typically low among children aged 5-9, even a small percentage of children who left home in their mid and late teenage years could bias mortality estimates. A comparison of age-specific mortality rates estimated with our method to rates from model life tables indicated that the bias is minimal for children aged 4 and younger (Hacker, Dribe and Helgertz (2023). Estimated ten-year mortality rates increasingly departed from predicted rates among children aged 5 and older in census A, however, consistent with an increasing tendency for children to leave home with each year of age.

[4] It is, however, difficult to be precise about the contribution of adult mortality to children not living with their parents. Available mortality estimates are not based on marital status and do not consider possible clustering of deaths among marital partners.

Originating from the population of successfully linked couples, we initially identify surviving children aged 0-8 in census A who were successfully linked to a census B record.[5] For this, we relied on the universe of MLP links supplemented with CT links. Among the unlinked children, there are two probabilities: (1) they died; or (2) they survived to census B but were not linked by the MLP or CT data. It is likely that some of the unlinked children survived but—for a variety of possible reasons—were unlinked. One reason for surviving children being unlinked to a census B record stems from the prioritizing of low type I errors in both sources of linked data. For example, while within-household links allow for record discrepancy in terms of several characteristics, including age and gender, the MLP algorithm still excluded many record pairs that a human would consider to be the same person. There are many linked couples, for example, with forwards unlinked children in census A and backwards unlinked children aged 10-18 in census B who were almost certainly the same children. Especially common are cases characterized by a large name discrepancy due to the use of initials in one census and a full first name in the other census. Also common are cases with a child enumerated with their probable first name in one census and probable middle name in the other census. A typical example is a couple with an unlinked daughter aged 3 named "Julia A." in census A and an unlinked daughter aged 13 named "Anna" in census B. In addition, age discrepancies, caused by age heaping or other types of age reporting and transcription errors, occur occasionally, with accurate record pairs typically being within a few years of each other but which were often unlinked underlying datasets.

For these reasons, we implemented a strategy to "force" links between children in census A who were not linked forward to children in census B who were not linked back to a child in census A. Here, we also relied on the CT data providing links to censuses further ahead in time than census B, allowing us to designate unlinked children in census A as "definite survivors" (i.e., children linked by CT to a census C/D/E record but not to a census B record) and "potential survivors" (i.e., children not linked to any future

[5] Note that children aged 0-8 in Census A are selected, rather than the age range we ultimately focus on. The reason for this is to avoid inflating the census undercount.

census). Beginning with the population of definite survivors in census A, and in descending order of age from age five[6], we forced a link to the closest same-sex and estimated birth year (+/- 3 years) backwards unlinked child in census B. We limited census B children to those aged 10-18, meaning that in practice we compared children aged 5 in census A to children aged 12-18 in census B, while we compared infants in census A to children aged 10-13 in census B. This asymmetry was motivated by wanting to avoid including census B children who were born after census A. Despite significant age misreporting errors in earlier censuses, we thought it very unlikely that a child aged 7-9 in census B was truly aged 10 or older. Age heaping—the preference of respondents to report some ages, typically those ending in the digits "5" and "0," over other ages—was relatively rare among children compared to age heaping among older adults, and in most cases would have resulted in too many children reported aged 10 and too few children aged 7, 8, or 9.[7] We then repeated this process for the population of potential survivors, and then again but without restricting links to children of the same sex. At its conclusion, successfully linked census A children were added to the population of survivors, whereas those that remained unlinked were considered to have died during the census interval.[8]

Although this method greatly reduced the number of children aged 10-15 in census B who were not linked backward to a child aged 0-5 in census A, some couples in the analytical datasets still had children aged 10-15 in census B who remained unlinked after the forcing process. There are four major potential sources for these unlinked census B children. We believe that census under-enumeration in census A is the most likely source of these unlinked children. Census undercounting of young children is a well-known problem in historical census data and in census data for developing nations in the twentieth century (Ewbank 1981; Hacker 2013). Coale and Zelnik's classic estimates of age-specific net undercounts of the native-

[6] This is motivated by an older child aged 0-5 at Census A being more likely to survive to the next Census, all else equal.

[7] We note one significant exception. The 1920 Census was enumerated earlier in the calendar year (January 5) relative to the 1910 Census (April 15). It this therefore likely that a significant percentage of 9 years olds in the 1920 Census were born prior to the 1910 Census and enumerated age 0. We are currently evaluating ways to force links between children of these ages but have not implemented any forcing in this paper.

[8] See Appendix Table A2 for a comparison of all census A children aged 0-5 to the study population of census A age 0-5 deaths and survivors.

born white population in the 1880-1940 censuses indicated that younger children were much more likely to be undercounted than older children, averaging 7.1% for children aged 0-4, 3.7% for children aged 5-9, and 3.3% for children aged 10-14 (Coale and Zelnik 1963: 179-180). Although Black individuals likely suffered higher census undercounts in all census years (Robinson et al. 1993), there are no good estimates of age-specific Black undercounts before 1930. In that year, Preston et al. (2003) estimated that Black children aged 0-4 were undercounted 14.2%, more than two times the omission rates estimated for Black children aged 5-9 (5.5%) and aged 10-14 (7.0%). White and Black children who were undercounted in census A but enumerated in census B when aged 10-15 survived the intercensal interval.

A second possible source of unlinked children aged 10-15 in census B is from other errors in one or both linked datasets. Potential errors that could have resulted in unlinked children includes the enumeration of coresident nephews, nieces, other relatives, or non-relatives (farm hands, boarders, servants, etc.) as parents' own children in census B or an enumeration or transcription error in an individual's name, age, sex, race, birthplace, or other variables used by the linking algorithm in one or both censuses. Children aged 9 and younger who were reported as aged 10—a round number that appears to have been reported more frequently than children aged 9 and 11—in census B were not yet born before census A and therefore cannot be linked. Children aged 16 and above who are listed incorrectly in census B as aged 10-15 could also be unlinked, although our forcing procedure would link these children if within three years of the assumed birth year of an unlinked census A child. Third, even if ages and household members were accurately reported and transcribed in both censuses, an unlinked child in census B could result of changes in the nominal date of enumeration for different censuses, leading to greater discrepancies in their calculated year of birth and greater difficulties linking.[9] This was particularly problematic among children enumerated as aged 10 years in the 1930 census. The nominal date of the 1920 census was January 5, while the nominal date of the 1930 census was April 1. Children born in the period January 5 – March 31, 1920 and surviving

[9] The nominal date of each enumeration was June 1 for the 1850-1880 censuses, April 15 for the 1910 census, January 5 for the 1920 census, and April 1 for the 1930 and 1940 censuses.

to the 1930 census, therefore, should have been enumerated as aged 10 years in the 1930 census but should not have been enumerated by the 1920 census. Finally, unlinked children in census B could represent children who for whatever reason were living apart from their parents in census A when aged 0-5, but who were coresident with their parents in census B when aged 10-15.

Although there may have been alternate and multiple sources of error resulting in unlinked children in census B, we assumed that all remaining children aged 10-15 who were not linked backwards were "undercounted" in census A. We added a record for these undercounted children to their parents' records at their estimated age in census A (age in census B – 10 years) and assumed all these children were at risk of mortality and survived the interval. For the 1920-1930 interval, which is the only interval in study in which the nominal date of census B occurred later in the calendar year than in census A, we randomly selected the percentage of undercounted children aged 10 in 1930 that were assumed to have been born between January 5 and March 31, 1920, and removed them from the at-risk dataset. Note that in the case of young children living apart from their parents in census A, our undercount assumption results in the correct inference of their survival to census B.

Naturally, our method is unable to generate a sample of representative couples and children. Because children of unlinked parents likely experienced higher rates of mortality, we constructed and applied propensity weights as described by Bailey et al. (2020) to yield estimates consistent for all children of baseline couples in census A[10]. Table 1 shows the share of the children considered as undercounted and

[10] We used logistic regression to model whether children aged 0-5 of coresident parents in the IPUMS full-count datasets were children of parents linked to the subsequent census. Independent variables included children's race; mother's age, nativity, and literacy; urban-rural residence, size of place, and census region; and father's occupation. Propensity weights were constructed in Stata using a program developed by Mark Lunt (2014). Similar propensity weights could be constructed for orphaned children or children with only a mother or father present in the household in census A using fewer covariates. Children without coresident parents are much more difficult to link to census B records, however, and we found that the higher error rates associated with linking these children significantly biased our resulting mortality estimates. We therefore limited the study to children of two linked parents. Orphaned children, of course, likely suffered significantly higher mortality rates. In a recent study of the impact of parental death on child mortality in the Netherlands, 1850-1940, Quanjier et al. (2023) found that children aged 0-1 at the time of their mother's death experienced more than 4 times of the risk of death compared to other children. Among children aged 1-5, a father's death was associated with a 1.39 times higher risk of death while the death of a mother was associated with a 2.30 times higher risk (both sexes combined). Fortunately, the proportion of children orphaned was low in most census years. The percentage living aged 0-5 in census A without a mother present in the household

dying between the censuses in the completed datasets by age in census A, after the inclusion of propensity weights. Combined, the data cover 35.2 million at-risk children born between around 1845 and 1930, of whom about 1.5 million are considered to have died during the subsequent ten-year period. Obviously, these proportions hide considerable heterogeneity by socioeconomic status, residence characteristics, and other assumed correlates. The average proportion of children aged 0-5 in census A dying is plotted by race and census year in figure 1. Because slaves were enumerated on separate census schedules in 1850 and 1860 and were not named, estimates shown for Black children for the 1850s and 1860s represent only the children born to free Black parents (about 10% of the Black population) who could be linked between the 1850-1860 and 1860-1870 censuses. The overall pattern indicates a long-term decline in mortality among both white and Black children, but with substantially higher mortality rates among Black children in all census years. Although the race differential in the average mortality rate between Black and white children narrows between the 1870s and 1930s, the ratio of black to white child mortality increases. Interesting, the results indicate an increase in white child mortality during the 1860s, which corresponds with the years of the American Civil War. This increase has long been suspected—the war was associated with large numbers of refugees and the probable spread of infectious disease—but has never been confirmed (Cashin 1996; Hacker 2011). While the exact timing of the onset of the mortality transition is difficult to pinpoint due to the gap in the data between 1880-1900, substantial declines in the 10-year mortality risk across all ages can be observed from circa 1870. Although not shown in the plot, we find the highest mortality rates for children aged 0 in census A, lower mortality rates among children aged 1, and consistently lower rates for children aged 2-5 in most years and race combinations, generally consistent with published life tables.

[Table 1 and Figure 1 here]

Turning to census undercounts, or unlinked census B children whose age in census B suggests that they were alive at the time of census A, we observed a nontrivial presence in each census. Although there

ranged between a high of 6.0% to a low of 3.1% over the course of the study period, while the percentage living without a father ranged from 10.7% to 6.1%. Many of these children may have had a surviving mother or father who was not coresident (e.g., fathers who worked and were enumerated elsewhere).

is some heterogeneity across census years and by race, we find that undercounts tended to be much larger for children aged 0 in census A and larger in the earlier censuses, except for the 1920 census, which was recorded earlier in the calendar year. The very large undercounts among children aged 0 suggests that we should be cautious about interpreting results for this age. Among white children aged 1-5, the average undercount was typically 2-3%, but more elevated in the 1870 census (5.5%). This census has long had the reputation as suffering from higher undercounts than other nineteenth-century censuses because of the unsettled conditions after the American Civil War. Undercount rates among Black children were typically 2-3 times higher than white children, averaging about 6-7% among children aged 1-5 in the twentieth century.

Verification of the Results

How accurate is our method of identifying child mortality using linked census data? Unfortunately, the lack of a complete and accurate death registration system in United States until the end of the period of this study means we lack a reliable standard to verify whether the children we infer as having died in an intercensal interval truly did so. We are, however, able to compare our estimates to a variety of direct and indirect estimates. We begin with an overview of the study period by comparing our results at the national level to published life tables. We then move to a direct comparison of our inferences of survival or death of individual children in the four decades between 1900 and 1940, using links provided by FamilySearch.org. Finally, we conduct several comparisons for the period 1900-1910, the approximate middle of our analysis, using children ever born and children surviving data in the 1910 census. For reasons of concision and for comparability to the results for children ever born and children surviving data in the United States—which do not distinguish the sex of ever born and surviving children—we limit our discussion to the mortality of both sexes combined.

Comparison of national estimates to published life table estimates

We begin by comparing our estimates of child mortality at the national level to published life tables. For the white population prior to 1900, we use life tables constructed by Haines (1998) and Hacker (2010). After 1900, we compare our estimates to life table estimates made for the white population residing in the DRA (Foudray and Davis 1923; Glover 1921; Hill, Glover and Foudray 1936). There are no reliable life table estimates for the black population in the period 1850-1900, but we can compare our estimates for the period after 1900 to DRA life tables. As noted earlier, however, we caution that the black population in the DRA in the early part of the twentieth century was unrepresentative of the overall black population. It was far more urban, which results in a significant upward bias in the DRA child mortality estimates (Preston and Haines 1991: 81-83). For the decades before the 1900 and 1910 censuses, we also include national estimates of white and black child morality made by Haines and Preston (1997) using information on the number of children ever born (CEB) and the number of those children surviving (CS) collected for all ever-married women by the 1900 and 1910 censuses.

For each age, race. and census combination, we aggregated the total number of at-risk, surviving, and deceased children estimated using the methods described above. For children aged 0-5 in census A, we calculated cohort mortality ratios by dividing the total number of children at each age dying in the intercensal interval by the total number of at-risk children. Using conventional life table parameters, this is the equivalent to $(L_x - L_{x+10})/L_x$. We then used Coale and Demeny's (1983) "Model West" family of model life tables to convert these values to the proportion of children dying before age 5 (q_5) for each age.[11] A preliminary analysis indicated a small but significant upward bias in our mortality estimates among children aged 4 and 5 in census A compared to children aged 1-3 above what might be expected by secular trends (likely the result of a small proportion of children who left the parents household and who we were unable

[11] We evaluated several different model life table systems, including Coale and Demeny's North and South life tables, the United Nation's Far-East and General model life tables, Preston et al.' (1993) models for high mortality populations (based on the experience of African Americans arriving in Liberia in the late nineteenth century and the United Nation's General Model), and Brass relational models based on white and black DRA life tables with fixed slopes. Model West is the most used model for U.S. populations and resulted in q_5 values approximately midway between the estimates based on other models.

to link).[12] Together with our concern about the large proportion of undercounted children aged 0, we decided to rely solely on the average of the fitted q_5 values for children aged 1-3. These values are plotted in Figure 2 together with estimates from Hacker (2010), Haines and Preston (1997), and the DRA where available. Not surprisingly, our estimates show a similar trajectory to the proportions of children aged 0-4 dying in the ten-year intervals plotted in Figure 1, but now are converted to a standard life table parameter and comparable to other estimates.

[Figure 2 here]

Encouragingly, our estimates for the white population closely approximate the level and the long-term decline in child mortality estimated by Hacker (2010) before 1900 and the mortality for the white population in the DRA after 1900. For the one decade in which our estimates of q_5 for the white population overlap those made by Preston and Haines with CEB and CS data in the 1910 census, our estimate of the proportion of white children dying before age 5 (0.148) closely corresponds with Preston and Haines' estimate (0.147). This close correspondence suggests that our concern that reliance on children of linked couples might bias our estimates downwards was either unfounded or else the bias was compensated by offsetting biases (e.g., surviving children we failed to link). The few exceptions when our estimates vary from other published estimates can be explained. For the 1850s, Hacker's estimates (2010), which were based on estimates of adult life expectancy made by other researchers using genealogical data and a relational model life table based on the 1901 DRA to estimate child mortality, are likely too high. The average of Haines's estimates (1998) (not shown) of q_5 for the 1850 and 1860 censuses of mortality (for the years 1849-50 and 1860-61 and which includes data from the 1849 cholera pandemic year), suggests a

[12] Very few children of living parents appeared to have left their parents' homes before age 15. Steckel (1996) estimated that the median age that white males left home in 1860 was 25.2 years and 24.4 years for white females and published estimates of adult mortality can account for the small percentages of children in cross-sectional census datasets who were not residing with their parents before age 15. Mortality rates are low in these age groups, however, and even a small percentage of children who left home in their mid to late teenage years and who could not be located in the full-count datasets could bias mortality estimates. An earlier comparison of age-specific mortality rates estimated with our method to rates from model life tables indicated that the bias was negligible for children aged 4 and younger but more significant at older ages (Hacker, Dribe and Helgertz 2023).

value (0.274) closer to our estimate (0.239). As discussed earlier, the American Civil War in the 1860s resulted in large disruptions in the economy and in family life, and the movement of troops and refugees likely contributed to the spread of diseases, which disproportionately contributed to higher mortality rates among children. Hacker's estimates, which are based on adult life expectancies, do not reflect this possibility. Our data, therefore, represents an exciting opportunity to examine the human cost of war more fully than has been possible before. Finally, for the disagreement with our rates and the DRA rates in the 1920s and 1930s, we note that our estimation method likely suffers a modest but increasing bias towards overestimating child mortality with each decade closer to 1940, the endpoint of the available MLP and CT data. In the 1930s, there is only one opportunity in the estimation method to observe a child's survival: the 1940 census. For earlier decades, there were more opportunities. In the decade 1900-1910, for example, we can observe child survival through links to any census between 1910 and 1940.

The accuracy of the results for the black population is more difficult to assess. Higher black child mortality rates in all census years and the long-term decline in mortality rates are consistent with our expectations, which is encouraging. Our estimate for the percentage of free black children dying in the 1850s and 1860s—about 34%—is lower than Steckel's estimate for the enslaved population (about 50%), but this difference is readily explained by slave women being forced back into field work and the early weaning of children (Steckel 1986). There is little doubt that children born to free black women enjoyed higher survival rates than children born to slave women. When our estimates can first be compared to an alternative source, namely Haines and Preston's estimates for the first decade of the twentieth century, they are quite close—our estimate (0.251) is just 4.1% higher than Haines and Preston's estimate (0.241). Both estimates, of course, are lower than the estimate for the black population residing in the DRA (0.316) (Glover 1921; Preston and Haines 1991), but that difference is explained by the highly unrepresentative black population in the DRA, which were concentrated in urban areas with high mortality rates (Preston and Haines: 49-50, 81-83). More worrisome in the divergence of our rates from the DRA rates in the 1920s and 1930s, when the DRA included a more representative share of the black population. It is possible, of

course, that black deaths were under-registered in the DRA, but black births were also likely under-registered as well, providing a countervailing bias (Eriksson, Niemesh and Thomasson 2018). As we contended with the white population, we suspect an increasing bias towards overestimating black child mortality with each decade closer to 1940, the endpoint of our available data. It may also be the case that Model West life tables are not appropriate for the black population after circa 1910. Use of the Coale and Demeny's Model North or the United Nation's Far East model life table system, the latter of which Douglas Ewbank (1987) has suggested is the best fit for the African American population in this period, lowers the estimated q_5 values substantially. In the period 1930-1940, for example, the estimated q_5 for our data falls from 0.179 to 0.150, still significantly higher than the value estimated using the DRA (0.099), but much closer than the estimate based on Model West.

Case validation using Family Search data

Ideally, we would like to compare a representative sample of our results to reliable death records on a case-by-case basis. Although there is no ground truth dataset available for verification, some non-representative comparisons can be made by overlapping our data with census-to-mortality records from FamilySearch.org, accessible through a collaboration with the Record Linkage Lab at Brigham Young University.[13] Family Search (FS) is a free online genealogical resource created and supported by the Church of Latter-Day Saints. The FS website allows amateur and trained genealogists to construct family trees and link individuals in the trees to censuses and extant birth, marriage, and death records. We randomly selected 500,000 cases from each of the four twentieth-century child mortality datasets for linking with FS, available from 1900. From this group we removed undercounted children who, by definition, are known to have

[13] We thank Joseph Price, Tommy Morgan and the team of researchers at the Record Linkage Lab for their assistance.

survived. Of those, around 50% of the 1900-1910, 1910-1920, and 1920-1930 children were linked to death records by FamilySearch, but only 39% of the children in the 1930-1940 dataset.

Unfortunately, our estimate of a child's survival was strongly and positively associated with the record being linked to FS, while our estimate of a child's death was negatively associated with the record being linked. Only about one-third of the children our method inferred as having died before census B were linked, compared to about one-half of the children the method inferred as having survived to the census. Assuming our inference of these children's survivals and deaths were correct, the smaller percentage of inferred deceased children who were linked to FamilySearch data is likely because of their appearance in only one twentieth-century census and the relatively poor coverage of the death registration system in the early twentieth century. Simply put, children who died young appear in fewer records, especially early in the century, when many state vital registration systems were incomplete or non-existent. Other possible reasons include poor data quality, which might have prevented a link to an existing census or death record.

Conditional on deaths being observed in the FamilySearch data, we defined *confirmed FamilySearch survivors* as children having an observed year of death after census B and *confirmed FamilySearch deaths* as children having an observed year of death between census A and census B. We then compared confirmed FamilySearch deaths and survivors to our inferred deaths and survivors. Table 2 shows the results of this exercise. Among children we identified as surviving each decade, FS almost always agreed, with agreement rates above 99%. Among those children we inferred as having died, the agreement is lower, falling from around 89% in 1900-1910 to 69% in 1930-1940. The greatest difference, therefore, was in the decade with the lowest mortality rate.

[Table 2 here]

There are several reasons to believe that the accuracy of our mortality inferences is considerably higher than suggested by the agreement rates with FS. As noted, around 70% of the children we designated as deceased in census B—could not be linked to FS. A likely reason they could not be linked was that these

children in fact died early in their life course. Deceased children do not appear in future censuses, so they cannot be confirmed in a family tree as living by their appearance in census B, or as living in census C, D, E, or F. Deceased children do not marry, which reduces their possibility of appearing on a family tree through their appearance on a marriage record. And deceased children do not reproduce and therefore cannot be direct ancestors of future genealogists. Although some children's deaths were recorded in FS, there is a bias towards individuals who lived long enough to reproduce. It is for this reason that estimates of life expectancy based on genealogical evidence typically ignore children—infant and child mortality estimates are far too low relative to adult mortality estimates—and are restricted to estimates to adult life expectancy (e.g., Pope 1992). Second, as noted, the death registration system in the United States was not considered "complete" until 1933 and suffered from under-registered deaths for many years thereafter. If a child could not be linked to a death record, it is thus very likely because they died earlier in the century and in a state without a death registration system.

To consider this possibility, the final two rows of Table 3 show the adjusted accuracy of our inferences under the assumption that all children not linked to FS were accurately inferred as deceased or surviving. The results suggest that the percentage of deaths accurately inferred by our method could be as high as 96.5% in 1900-1910 and 1910-1920. Even under this best-case scenario, however, the accuracy falls to 93.8% in 1920-1930 and 90.4% in 1930-1940. Interestingly, this finding agrees with our comparison to national life tables above, which found good agreement for the decades prior to 1920 and some overestimation of mortality for the periods 1920-1930 and 1930-1940. As we noted above, a probable reason is the end of our available data in 1940. A non-random check of cases in the period 1930-1940 in which we believed the child was dead while FS believed the child survived found a disproportionately large number of children aged 4 and 5 in 1930, suggesting that many of these cases are the result of children who left home before age 14 or 15 and who we could not confirm as survivors in the 1950 or subsequent censuses.

Comparison of model results using our estimates and an alternative estimate based on Children Ever Born and Children Surviving in the 1910 census

We anticipate that a major use of our new indirect estimation method is for modeling individual-level, household-level, and area-level correlates of child mortality similar to that pioneered by Samuel Preston and Michael Haines in *Fatal Years: Child Mortality in the Late Nineteenth-Century America* (Preston and Haines 1991). Although Preston and Haines' analysis –and similar studies by other researchers (Preston et al. 1994; Dribe et al. 2022)–was based on an index of child mortality constructed using the number of women's children ever born (CEB) and number of children surviving (CS) data collected by the 1900 and 1910 censuses, not an analysis of individual children's survival or death over a ten-year interval, it is possible to construct similar models using both methods for comparison. Child mortality estimates based on CEB and CS data pertain to mortality in the period before the census. Because our estimates are not available for the periods 1880-1890 and 1890-1900, we use only the IPUMS 1910 full-count dataset for estimation of child mortality in the period 1900-1910 for comparison to our estimates for the same period.

Details of the use of CEB/CS for child mortality estimation are provided elsewhere (Dribe et al. 2020; Haines and Preston 1997). Briefly, we relied on all mothers present in the 1910 census aged 20-39 with spouses present and valid CEB and CS data and estimated the child mortality index using the proportion of each mother's children dying relative to the *expected* proportion of her children dying.[14] The expected proportion dying was obtained using the mother's age in 1910 (a proxy for her children's length of exposure to the risk of dying), parity progression ratios for women in her birth cohort, and a life table standard (we relied on Model West life table 13.5, which is a good approximation of average mortality rates by age during the period).

[14] It is impossible to be precise about the period in which children were born and died using CEB and CS data. Among mothers aged 20-34 in 1910, however, most child deaths should have occurred in the decade prior to the census. According to the calculation of the mortality reference date associated with the method, the average date of death for children born to women in each of these age groups was after 1900.

In Table 3 we show the results. Although based on different estimation methods, data sources, and models—model 1 uses the 1910 IPUMS full-count dataset and a weighted OLS regression of the child mortality index constructed for each mother, while model 2 uses our data and a logistic regression model of the survival or death of each child prior to the 1910 census —the coefficients are interpreted in the same way, as the risk of child mortality relative to the reference group.

[Table 3 here]

In general, there is excellent agreement between the two models. Black child mortality was 60% higher in the 1910 cross-sectional dataset using the CEB/CS mortality index measure and 65% higher in our dataset compared to white children. Relative to children whose father was a farmer, children of blue-collar workers suffered 16% higher mortality in the CEB/CS model and 15% higher mortality using our estimates, while children of white-collar workers enjoyed similar or lower rates than farm children in both models. According to our estimates, children of literate parents enjoyed about 19.3% lower mortality than children of illiterate parents, while the CEB/CS estimates indicate that the advantage was 20.2%. Both models agree that urban areas were more deadly than rural areas, although the risk appears to be somewhat greater in the models based on our data. Both models indicate higher mortality among the children of Irish and Canadian parents relative the native-born parents and lower mortality in the Midwest census region relative to the Northeast. The most significant differences in the results are the findings for the West census region, which the model based on the CEB/CS data and methods suggests modestly lower mortality than the Northeast region while the model based on our data indicates similar mortality to the reference group. The West census region had few cases (just 4.4% of the births in our dataset), however, and was in the process of being populated by recent migrants. Overall, the strong agreement between the two models increases our confidence that our method of estimating child mortality using linked census data can be used to model child mortality in other decades. Other comparisons are possible. In an online appendix we show good correlation between our estimates and the state and county level with estimates made using CEB and CS data.

Geographic Differentials and Abridged Life Tables

Although our primary goal in this paper is to describe and evaluate our method of estimating child mortality with linked census data, in this section we briefly discuss two possible uses of the new estimates: mapping geographic differentials in child mortality and the construction of abridged life tables. In Figures 3 and 4, we map child mortality in the period 1900-1910 by county of residence in 1900. Although all counties are mapped, many contained small numbers of at-risk children and even fewer deaths. To reduce random noise, we limited unique estimates to counties with 1,000 or more at-risk children (ages 0-5, both sexes combined). If a county did not have a sufficient number of cases, we replaced the mortality estimate with the estimate from the next higher level of geography with 1,000 or more cases (SEA, state, census region, or nation). Figure 3, which maps child mortality for the white population, shows substantial geographic differentials. The proportion of children dying before age 5 (q_5) in many midwestern counties was less than 0.10 (low for the period). Proportions of children dying were significantly higher mortality in the South and in developed counties in the Northeast. Many counties along the lower Mississippi River, including counties in the states of Missouri, Tennessee, Mississippi, Louisiana, and especially in Arkansas, suffered two- or three-times higher rates of mortality than the low mortality areas of the Midwest. Coastal counties in Virginia, North Carolina, and South Carolina also experienced higher child mortality rates than inland counties. These patterns are suggestive to high death rates from malaria, which was known to be endemic in counties with year-long warm temperatures and ponding, ideal conditions for the *Anopheles* mosquito vector (Elman, McGuire and London 2019; Maxcy 1923). White children in the Southwestern states of Arizona and New Mexico also experienced high rates of death. It is difficult to see on the map, but counties dominated by large urban areas also had high child mortality rates.

[Figure 3 and Figure 4 here]

Figure 4, which maps child mortality patterns for the black population, shows higher mortality rates overall but similar geographic differentials to that observed for the white population. The strong differences in counties bordering state lines apparent in many areas are in most cases an artifact of our decision to replace county estimates based on fewer than 1,000 children at risk of mortality with the estimate from the next higher level of geography with 1,000 or more observations. Very few black children resided in northern states prior to the Great Migration that commenced in the 1910s and those that did tended to live in large urban cities with high mortality rates, not rural counties. The high mortality rates shown for all New York counties, for example, was dominated by the black population residing in New York City and a few small urban places around the state. In southern states, however, there are many counties and SEAs with a sufficient number of cases to map unique values. These show similar geographic differentials as observed for the white population, with higher mortality in the coastal counties of the Carolinas and in counties bordering the Mississippi River.

Our child mortality estimates can be the basis for constructing complete life tables using a model table system. There is, of course, significant potential bias in inferring a complete life table from a single life table parameter such as q_5. In addition to being reliant on the accuracy of the estimate, life tables estimated from a single parameter should ideally be based on a model that accurately reflects the true age-pattern of mortality for the place and time, particularly the relationship between child and adult mortality rates. The true relationship is unknown, however, and different models yield different results for mortality at older ages and associated life expectancies. Given the lack of county, state, and even regional life tables for this period, and their usefulness in many applications, however, we decided to construct complete life tables for each decade, sex, and race at five levels of geography: county, state economic area (SEA), state, census region, and nation.

Because the results shown in Figure 2 corresponded closely with published life tables and other estimates prior to 1920, we made no adjustment to our estimates in those decades. As we noted above, however, our estimates of child mortality were higher than published life table estimates in the 1920s and

1930s, a period when the DRA can be considered reasonably representative and complete. This overestimation of mortality was confirmed in our analysis of cases linked to FT. Given these findings, we decided to standardize our results for the 1920s and 1930s to the national estimate of q_5 for the white and black populations residing in the DRA. A correction factor was established for each sex and race using the ratio of q_5 estimated from the observed ten-year mortality rates to the observed q_5 values in the DRA and applied to all county, SEA, state, and census division estimates. We then fitted the adjusted q_5 estimates—and the unadjusted estimates for other decades—to Coale and Demeny's Model West and other model life tables to infer complete life tables.[15]

Given the large amount of data, there is too much information to summarize here. We have created a database with the estimated q_5 values for each county and the following life table parameters for each of the seven models: life expectancy at birth (e_0), infant mortality ($_1q_0$), and the number of survivors of a hypothetical cohort of 100,000 births at ages 0, 1, 5, and every five years through age 85 and older (l_0, l_1, l_5, l_{10}, … l_{85+}). We also include the number of at-risk children used to produce the estimates and life tables values estimated at four additional levels of geography: SEA, state, census division, and nation. For consistency, we use Model West for all decades prior to 1920 for both races. For the decades 1920-1930 and 1930-1940, however, we also constructed estimates for the white population using a relational model based on the observed average DRA life tables for the white population in each decade, and for the black

[15] As observed in note 11 above, we evaluated seven different model life table systems to generate full life tables and estimated life expectancy at birth. We limited our discussion and presentation to Model West for ease of discussion and for several other reasons. First, Model West was based in part on U.S. data and prior researchers have confirmed that the model results in a close match for the United States circa 1900 (Preston and Haines 1991). Second, results were approximately midway between those of other models. A strong case can be made for the use of Preston et al.'s (1993) "Liberian" models for the black population of the United States prior to 1900, which is based on the experience of the African American population in late nineteenth-century Liberia, but results for that model are very similar to those for Model West. Other model suggestions have been offered for the black population in the early twentieth century (e.g., Ewbank 1987, suggests that the United Nation's Far East model is the best fit). Without alternative standards to test the results against, we preferred using a consistent model. A test is possible at the national level for period after 1920, when the DRA is reasonably complete. We found very good correspondence in the 1920s and 1930s between the life tables constructed with Model West and the DRA life tables for the white population, but the best correspondence—unsurprisingly—with a Brass relational model based on the DRA life tables. We also found the best correspondence at the national level between the life tables we constructed for the black population with a Brass relational model based on Preston et al.'s 1935-1940 life table for the black population.

population using a relational model based on Preston et al.'s (2003) 1935-1940 life table. Unsurprisingly, these relational models resulted in modestly better fits at the national level than when using Model West, but their applicability to all counties—we used a fixed slope and adjusted the intercept to fit the logistic model to our county estimates—is unknown.

Conclusion

In this paper we described a new method to estimate white and black child mortality using linked census data. We used linked data for the periods 1850-1880 and 1900-1940 to construct individual-and geographic-level estimates of mortality and census undercounts and compared the mortality results to several external standards. Despite some concern that our reliance on children of linked couples might bias mortality estimates downwards, we found close agreement between the new estimates and published life tables in most decades. Our method appears to overestimate mortality in the 1920s and 1930s, however, likely because of the current lack of linked census data beyond 1940. A comparison of our estimates to a non-representative dataset constructed by linking children to available Family Tree data at FamilySearch also revealed close agreement at the individual level, although accuracy appears to fall in the 1920s and 1930s consistent with our comparisons to national life tables. Perhaps most importantly, our comparison of models of child mortality using estimates based on the linked data and estimates derived using children ever born and children surviving data from the 1910 census found remarkably similar results to those reported by other researchers (e.g., Preston and Haines 1991). These results suggest that mortality estimated with linked census data is a viable approach to investigating the contribution of suspected correlates over the course of the mortality transition from its onset in the late nineteenth century to the mid twentieth century.

References

Arias, E. 2014. "United States life tables, 2009."

Bailey, M., C. Cole, and C. Massey. 2020. "Simple strategies for improving inference with linked data: a case study of the 1850–1930 IPUMS linked representative historical samples." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53(2):80-93.

Buckles, K., A. Haws, J. Price, and H.E. Wilbert. 2023. "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project." National Bureau of Economic Research.

Cashin, J.E. 1996. "Into the trackless wilderness: The Refugee experience in the Civil War." Pp. 29-54 in *A Woman's War: Southern Women, Civil War, and the Confederate Legacy*, edited by J. Campbell, Edward D. C. and K.S. Rice. Richmond, Va.: The Museum of the Confederacy and the University of Virginia Press.

Coale, A.J.and P. Demeny, with B. Vaughan. 1983. *Regional Model Life Tables and Stable Populations*. New York: Academic Press.

Coale, A.J.and M. Zelnik. 1963. *New Estimates of Fertility and Population in the United States: A Study of Annual White Births from 1855 to 1960 and of Completeness of Enumeration in the Censuses from 1880 to 1960*. Princeton: Princeton University Press.

Cutler, D., A. Deaton, and A. Lleras-Muney. 2006. "The determinants of mortality." *Journal of economic perspectives* 20(3):97-120.

Dribe, M., J.D. Hacker, and F. Scalone. 2020. "Immigration and child mortality: Lessons from the United States at the turn of the twentieth century." *Social Science History* 44(1):57-89.

Dwyer-Lindgren, L., A. Bertozzi-Villa, R.W. Stubbs, C. Morozoff, J.P. Mackenbach, F.J. van Lenthe, A.H. Mokdad, and C.J. Murray. 2017. "Inequalities in life expectancy among US counties, 1980 to 2014: temporal trends and key drivers." *JAMA internal medicine* 177(7):1003-1011.

Elman, C., S.A. Cunningham, V.J. Howard, S.E. Judd, A.M. Bennett, and M.E. Dupre. 2023. "Birth in the US Plantation South and Racial Differences in all-cause mortality in later life." *Social Science & Medicine* 335:116213.

Elman, C., R.A. McGuire, and A.S. London. 2019. "Disease, Plantation Development, and Race-Related Differences in Fertility in the Early Twentieth-Century American South." *American Journal of Sociology* 124(5):1327-1371.

Eriksson, K., G.T. Niemesh, and M. Thomasson. 2018. "Revising infant mortality rates for the early twentieth century United States." *Demography* 55(6):2001-2024.

Ewbank, D.C. 1981. *Age misreporting and age-selective underenumeration: Sources, patterns, and consequences for demographic analysis*: National Academy Press.

Ewbank, D.C. 1987. "History of black mortality and health before 1940." *Milbank Quarterly* 65(Supplement 1 (Part 1).):100-128.

Foudray, E.and W.H. Davis. 1923. *United States Abridged Life Tables, 1919-1920*: US Government Printing Office.

Glover, J.W. 1921. *United States Life Tables, 1890, 1901, 1910, and 1901-1910*. Washington, D.C.: GPO.

Hacker, J.D. 2010. "Decennial Life Tables for the White Population of the United States, 1790-1900." *Historical Methods* 43(2):45-79.

Hacker, J.D. 2011. "A census-based count of the Civil War dead." *Civil War History* 57(4):307-348.

—. 2013. "New estimates of census coverage in the United States, 1850–1930." *Social Science History* 37(1):71-101.

—. 2016. "Ready, willing, and able? Impediments to the onset of marital fertility decline in the United States." *Demography* 53(6):1657-1692.

Hacker, J.D., M. Dribe, and J. Helgertz. 2023. "Wealth and Child Mortality in the Nineteenth-Century United States: Evidence from Three Panels of American Couples, 1850–1880." *Social Science History*:1-34.

Haines, M.R. 1998. "Estimated life tables for the United States, 1850-1910." *Historical Methods* 31(4):149-169.

—. 2000. "The white population of the United States, 1790-1920." Pp. 305-370 in *A population history of North America*, edited by M.R. Haines and R.H. Steckel. Cambridge ; New York, NY: Cambridge University Press.

Haines, M.R.and S.H. Preston. 1997. "The use of the census to estimate childhood mortality: Comparisons from the 1900 and 1910 United States census public use samples." *Historical Methods* 30(2):77-97.

Helgertz, J., J. Price, J. Wellington, K.J. Thompson, S. Ruggles, and C.A. Fitch. 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., S. Ruggles, J.R. Warren, C.A. Fitch, J.D. Hacker, M.A. Nelson, J.P. Price, E. Roberts, and M. Sobek. 2023. "IPUMS Multigenerational Longitudinal Panel: Version 1.1 [dataset]." edited by IPUMS. Minneapolis, MN.

Hill, J.A., J.W. Glover, and E. Foudray. 1936. *United States Life Tables: 1929 to 1931, 1920 to 1929, 1919 to 1921, 1909 to 1911, 1901 to 1910, 1900 to 1902*: US Government Printing Office.

Maxcy, K.F. 1923. "The distribution of malaria in the United States as indicated by mortality reports." *Public Health Reports (1896-1970)*:1125-1138.

Pope, C.L. 1992. "Adult mortality in America before 1900: A view from family histories." Pp. 267-296 in *Strategic Factors in Nineteenth Century American Economic History: A Volume to Honor Robert W. Fogel*, edited by C. Goldin and H. Rockoff. Chicago: University of Chicago Press.

Preston, S.H., I.T. Elo, M.E. Hill, and I. Rosenwaike. 2003. *The Demography of African Americans 1930-1990*. Dordrecht , Netherlands: Kluwer Academic Publishers.

Preston, S.H., D. Ewbank, M. Hereward, and S.C. Watkins. 1994. "Child mortality differences by ethnicity and race in the United States: 1900–1910." *After Ellis Island: Newcomers and natives in the*:35-82.

Preston, S.H.and M.R. Haines. 1991. *Fatal years: Child Mortality in Late Nineteenth-Century America*. Princeton, New Jersey: Princeton University Press.

Robinson, J.G., B. Ahmed, P. Das Gupta, and K.A. Woodrow. 1993. "Estimation of population coverage in the 1990 United States census based on demographic analysis." *Journal of the American Statistical Association* 88(423):1061-1074.

Ruggles, S., C.A. Fitch, R. Goeken, J.D. Hacker, M.A. Nelson, E. Roberts, M. Schouweiler, and M. Sobek. 2021. "IPUMS Ancestry Full Count Data: Version 3.0." edited by IPUMS. Minneapolis, MN: IPUMS.

Steckel, R.H. 1986. "Birth weights and infant mortality among American slaves." *Explorations in Economic History* 23(2):173-198.

Table 1: Unadjusted mortality and Undercount rates by age and race, both sexes combined

	1850-1860		1860-1870		1870-1880		1900-1910		1910-1920		1920-1930		1930-1940	
Age	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount
0	0.107	0.216	0.122	0.202	0.103	0.196	0.080	0.100	0.073	0.066	0.042	0.202	0.037	0.111
1	0.094	0.060	0.106	0.067	0.083	0.089	0.052	0.034	0.044	0.029	0.034	0.037	0.029	0.029
2	0.072	0.046	0.089	0.054	0.063	0.083	0.042	0.027	0.035	0.023	0.029	0.027	0.025	0.023
3	0.064	0.038	0.083	0.043	0.056	0.075	0.037	0.022	0.031	0.020	0.027	0.019	0.024	0.017
4	0.059	0.043	0.077	0.047	0.052	0.084	0.035	0.023	0.030	0.022	0.027	0.020	0.023	0.018
5	0.056	0.038	0.074	0.043	0.051	0.086	0.035	0.021	0.029	0.021	0.030	0.017	0.026	0.015
N	1,413,945		1,974,578		2,738,441		5,281,300		6,267,776		7,647,211		7,745,708	
N_{adj}	3,316,022		4,615,518		5,386,875		9,025,319		10,601,227		12,118,076		11,818,900	

Black Population

	1850-1860*		1860-1870*		1870-1880		1900-1910		1910-1920		1920-1930		1930-1940	
Age	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount	Dead	Undercount
0	0.173	0.248	0.174	0.284	0.143	0.260	0.130	0.157	0.134	0.126	0.075	0.203	0.064	0.164
1	0.153	0.090	0.150	0.159	0.127	0.123	0.094	0.069	0.083	0.071	0.059	0.085	0.050	0.068
2	0.134	0.110	0.127	0.182	0.100	0.129	0.077	0.063	0.068	0.064	0.052	0.071	0.046	0.059
3	0.146	0.081	0.135	0.131	0.097	0.089	0.069	0.046	0.062	0.051	0.049	0.050	0.045	0.050
4	0.155	0.080	0.147	0.141	0.101	0.100	0.069	0.050	0.060	0.056	0.057	0.054	0.047	0.048
5	0.137	0.070	0.145	0.131	0.103	0.103	0.076	0.043	0.066	0.052	0.065	0.047	0.055	0.044
N	13,221		16,683		193,222		404,246		466,316		469,118		530,319	
N_{adj}	52,073		59,942		730,226		1,174,696		1,307,253		1,220,398		1,250,647	

* The 1850-1860 and 1860-1870 linked dataset include only the free black population only. N is the total number of at-risk children aged 0-5 in the first of the two linked censuses including undercounted children. N_{adj} is the adjusted number of at-risk children weighted by propensity weights to represent all children of married couples (linked and not linked) in the first census.

Figure 1: Average ten-year mortality rate by race, children aged 0-5 in first of two linked censuses, both sexes combined

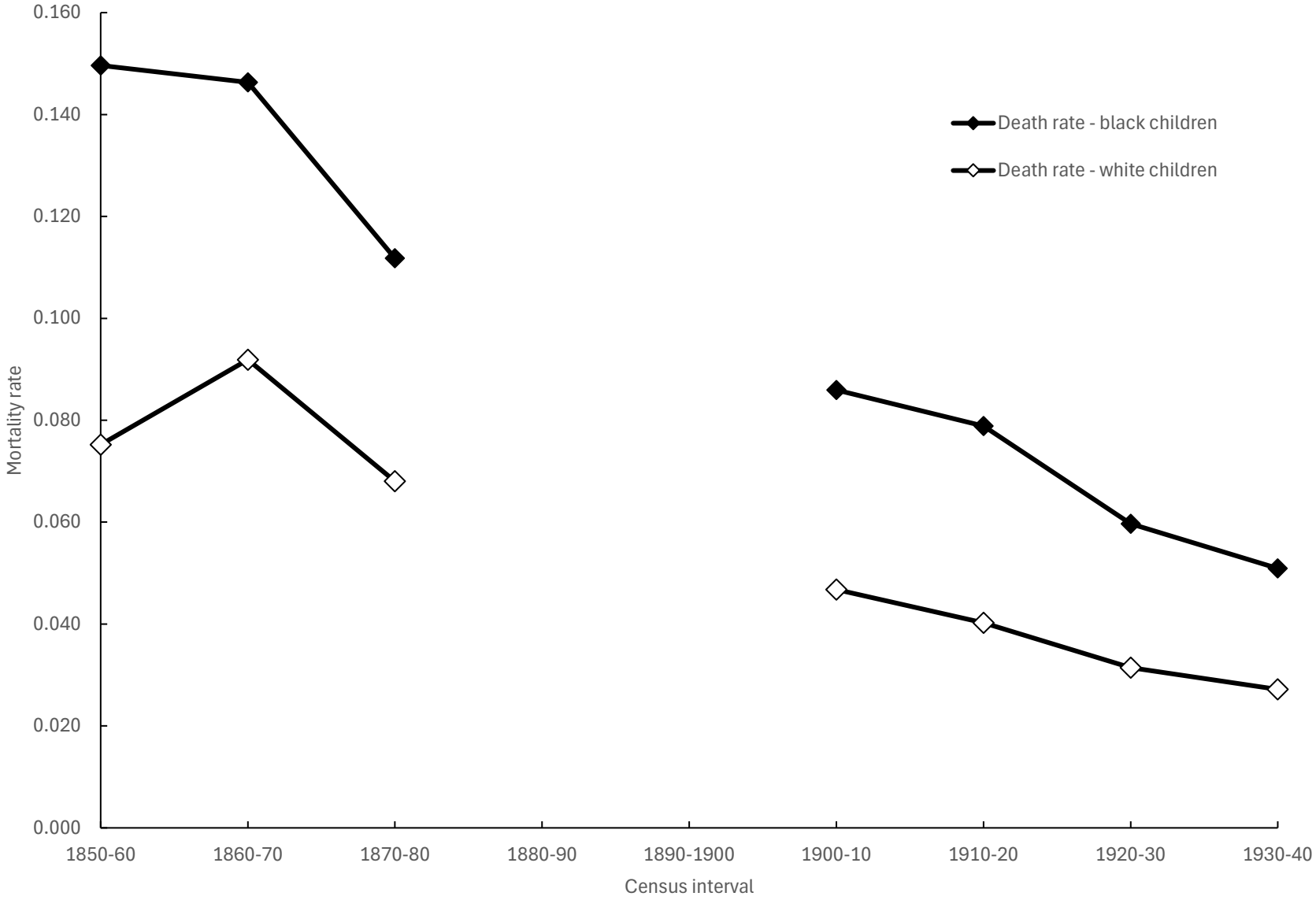


Figure 2: Proportion of children dying before age 5 (q_5), both sexes combined

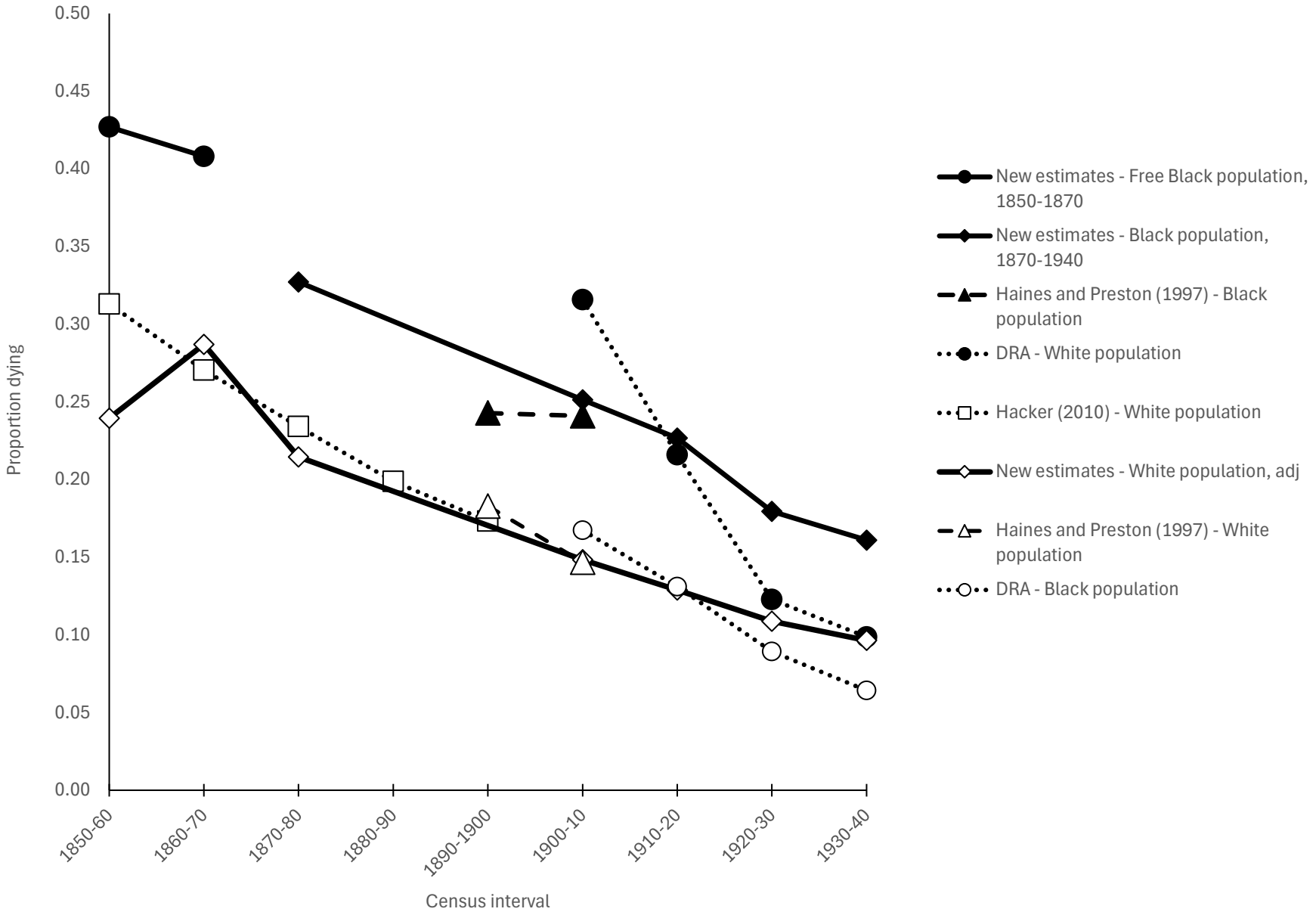


Table 2. Comparion of IPUMS MLP results to Family Search data

	1900-1910	1910-1920	1920-1930	1930-1940
Randomly selected cases for the IPUMS MLP dataset for comparison	500,000	500,000	500,000	500,000
Deaths in selected cases according to our method	24,985	21,421	16,775	14,424
Survivors according to our method	475,015	478,579	483,225	485,576
Undercounted cases in selected cases removed from comparison	20,507	16,475	36,116	19,059
Cases sent to Record Linkage Lab at Brigham Young University	479,493	483,525	463,884	480,941
Deaths in cases sent for comparison according to our method	24,985	21,421	16,775	14,424
Survivors in cases for comparison according to our method	454,508	462,104	447,109	466,517
Cases linked to death record by Family Search	239,575	250,784	234,177	186,781
Deaths in linked cases according to our method	7,660	6,469	5,182	4,350
Survivors in linked cases according to our method	231,915	244,315	228,995	182,431
Deaths according to our method confirmed as deaths by Family Search	6,795	5,764	4,167	2,988
Deaths according to our method confirmed as survivors by Family Search	865	705	1,015	1,362
Deaths according to our method not linked to Family Search	17,325	14,898	11,568	10,050
Survivors according to our method confirmed as deaths by Family Search	1,340	1,190	957	538
Survivors according to our method confirmed as survivors by Family Search	230,575	243,125	228,238	181,893
Survivors according to our method not linked to Family Search	222,593	217,702	218,030	283,992
Perc. of deaths confirmed as deaths by Family Search conditional on link	88.7%	89.1%	80.4%	68.7%
Perc. of survivors confirmed as survivors by Family Search conditional on link	99.4%	99.5%	99.7%	99.7%
Perc. deaths correctly inferred if cases not linked to Family Search are accurate	96.5%	96.5%	93.8%	90.4%
Perc. survivors inferred if cases not linked to Family Search are accurate	99.7%	99.7%	99.8%	99.9%

Table 3. Comparison of regression results, child mortality circa 1900-1910 by dataset, esimation method, and type of model, both sexes combined

Model number	1	2
Dataset	IPUMS 1910 full-count	IPUMS MLP 1900-1910
Estimation method	CEB/CS	Link observation
Model type	Weighted OLS	Logistic
Dep. Variable	Child mortality index	Death of child in interval
Child's characteristics		
Age census A		
0	*ref.*	*ref.*
1	-	0.595 ***
2	-	0.474 ***
3	-	0.415 ***
4	-	0.411 ***
5	-	0.411 ***
Race		
White	*ref.*	*ref.*
Black	1.599 ***	1.650 ***
Mother's characteristics		
Age group		
20-24	1.095 ***	0.898 ***
25-29	0.991 ***	0.906 ***
30-34	*ref.*	*ref.*
35-39	1.029 ***	0.973 ***
Father's characteristics		
Occupation		
Farmer	*ref.*	*ref.*
White Collar	0.997 **	0.967 ***
Blue Collar	1.160 ***	1.152 ***
Other	1.254 ***	1.211 ***
Parent's characteristics*		
Age difference	1.004 ***	1.002 ***
Literacy		
Illiterate	*ref.*	*ref.*
Literate	0.798 ***	0.807 ***
Nativity		
Native Born	*ref.*	*ref.*
Canadian	1.096 ***	1.076 ***
British	1.023 ***	1.016
Irish	1.064 ***	1.066 ***
German	0.975 ***	1.003
Other foreign born	1.001	0.983 *

Residential characteristics		
Size of place		
Rural	*ref.*	*ref.*
Small urban	1.041 ***	1.055 ***
Medium urban	1.066 ***	1.111 ***
Large urban	1.128 ***	1.189 ***
Region		
Northeast	*ref.*	*ref.*
Midwest	0.911 ***	0.839 ***
South	1.081 ***	1.073 ***
West	0.946 ***	1.019
Number of cases	24,855,777	5,246,146

Notes: The analystical population in model 1 is currently married couples in the IPUMS 1910 full-count dataset with wives aged 20-39, one or more children ever born, and marital durations of 10 or more years. For model 2 the universe is all white and black couples in the IPUMS MLP 1900-1910 dataset with wives aged 20-39 and one or more at-risk children aged 0-5 in the 1900 census. Couples are considered literate only if both partners can read and write. Nativity is based on the wife's nativity, unless the wife is native born and the husband is foreign born, in which case nativity is based on the husband's place of birth. Small urban are urban places with fewer than 25,000 residents, medium urban places are cities with 25,000-99,999 residents, and large urban places are cities with 100,000 or more residents. The number of cases for OLS model is the weighted number of children ever born to women in universe and the OLS regression is weighted by the number of children ever born.

Appendix

J. David Hacker and Jonas Helgertz, " Bricks without Straw: The use of linked census data to estimate child mortality in the pre-registration era of the United States."

March 13, 2024

This appendix has three numbered sections: (1) additional tables; (2) additional figures; and (3) information on a public use dataset constructed with the method, including details on its construction and features and instructions on how to download the data.

1. Additional Tables.

Table A-1 (for couples) and A-2 (for children of linked couples) compare the baseline populations in census A with the populations in the linked dataset.

2. Additional Figures.

As noted in the main text, we constructed an alternative estimate of child mortality at the mother level using the number of women's children ever born (CEB) and number of children surviving (CS) data collected by 1910 censuses are available in the 1910 IPUMS full-count dataset (Ruggles et al. 2021). Although CEB and CS data cannot be used to verify the death or survival of individual children, they allowed us to construct similar models of child mortality to models constructed with our estimates for comparison. Results are described in the main text. As briefly mentioned, CEB and CS data also allow us to construct county- and state-level estimates of child mortality for comparison with our estimates aggregated to the same levels, which we do here.

Child mortality estimates based on CEB and CS data pertain to mortality in the period before the census. Because our estimates are not available for the periods 1880-1890 and 1890-1900, we use only the IPUMS 1910 full-count dataset for estimation of child mortality in the period 1900-1910 for comparison to our estimates for the same period.

We created scatter plots and estimated simple OLS regressions for white and black children at the state and county levels. We limited the comparison to states and counties with 1,000 or more children in our dataset and weighted the regression results by the number of children aged 0-5 at risk.

State and county scatter plots for the white population are shown in Figures A1 and A2. At the county level, the correlation is weaker (R2 = 0.46) than at the state level (R2=0.77), but it is stronger when the results are limited to larger counties. In both plots the slope is less than 1.0, averaging about 0.85 for linear regressions without a constant), indicating that the IPUMS MLP is somewhat less sensitive than estimates based on CEB and CS data. In general, however, we believe that these results—which are based on different methods, selection criteria, and data sources—are very encouraging. Although we would have liked to have seen a stronger correlation at the county level, most counties have a relatively small number of children dying and the results are subject to random errors. And both methods are subject to non-random biases from a variety of data errors.

Largely because of its long use and analysis by different researchers, we have more confidence in the measurement of child mortality using CEB and CS data, but it is clearly an imperfect standard

subject to many possible sources of error, including errors in enumeration and transcription errors. There are additional reasons for disagreement between the two measures. Child mortality estimates based on CEB and CS data rely on information reported for all mothers in the 1910 census, not just linked mothers, and the information may reflect mortality occurring before 1900 and among children older than age 10 (e.g., the deceased child of a 39-year old mother of three children ever born and two children surviving in 1910 may have died as early as the late 1870s, when his or her mother was an older teenage mother, or as late as the day prior to the 1910 census, and anywhere between the ages of 0 and approximately 20). The estimates based on our data, in contrast, are limited to couples linked across the 1900-1910 interval, and therefore is known to occurred in the decade between the two censuses and before the age of 10 (for a child aged 0 in 1900) to age 15 (for a child aged 5 in 1900).

Figures A3 and A4 indicate that the correlation between the two different estimates for Black children is weaker. Both measures likely suffered greater errors related to poorer enumeration quality, both in the CEB and CS data collected, and among all variables associated with the quality of census linkages and the indirect measurement of child mortality. The results, nevertheless, indicate the expected large differences in child mortality by race suggested by other studies (e.g., Karbeah and Hacker 2023; Preston et al. 1994) and a positive relationship between the two measures. For the relatively smaller number of counties with 1,000 or more at-risk Black children aged 0-5 in 1900, the weighted regressions in models without a constant indicated a slope of 0.946 (Figure 5). For the state regressions (Figure 6), the slope was 0.931.

3. Information on a public use dataset constructed with the method

As mentioned in the text, we believe other researchers will find the life tables constructed with our estimates to be useful for a variety of research. We have therefore made these life tables available for free downloading at LINK.

At present, we make all data available in one STATA file. Table A3 shows the variables available. Each row in the dataset represents a unique combination of year (census A year), sex (1=male, 2=female, 3=both sexes combined), race (1=white, 2=black), life table model, and geographic area. We evaluated seven possible life table model systems. We decided to use Coale and Demeny's Model West for all life tables prior to 1920. Model West is the most frequently used life table model for U.S. populations and yields results in the approximately middle of the models we evaluated. For the white population in the decade 1920-1930, however, we relied on a relational model based on the life table for the white population in the DRA for the 1920s (the average of the 1920 and 1930 DRA life tables), which (unsurprisingly) yielded a modestly better fit to the DRA life table. Similarly, for the white population in the decade 1930-1940, we used a relational model based on the DRA life table for the white population in the 1930s. Although we could have used a similar approach in the decades 1900-1910 and 1910-1920, the DRA life tables for those decades are based on less representative, more subject to differential undercounts, and potentially biased. After 1920 we judged that the DRA for the white population was generally representative and complete. We continued to rely on Model West for the black population in the 1920s. For the decade 1930-1940, however, we used a relational model based on Preston et al.'s (1993) re-estimated life table for the black population 1935-1940. These model choices are noted in the variable "model".

Figure 3. White population, q_5, 1910 - 1920

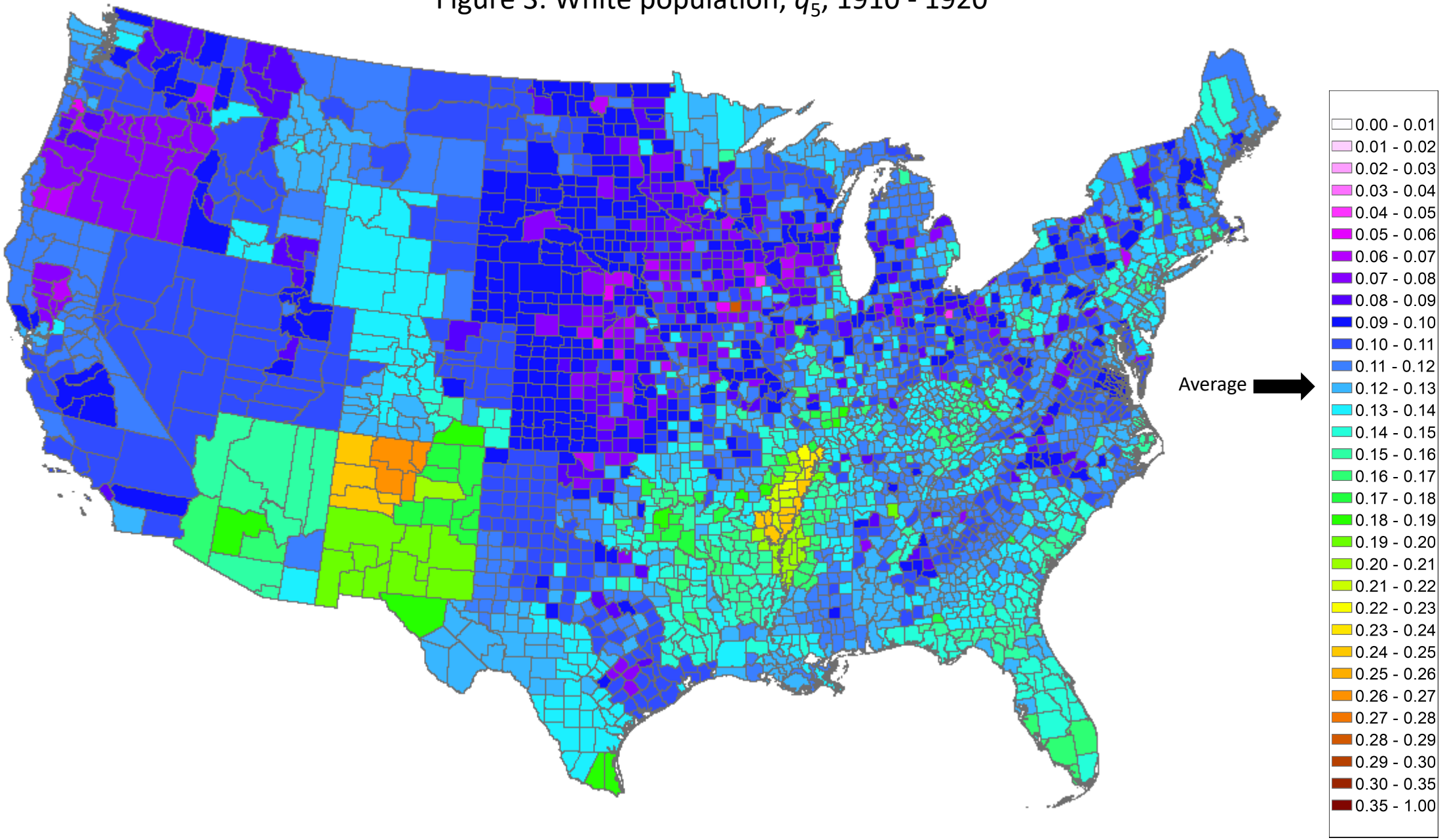


Figure 4. Black population, q_5, 1910 - 1920

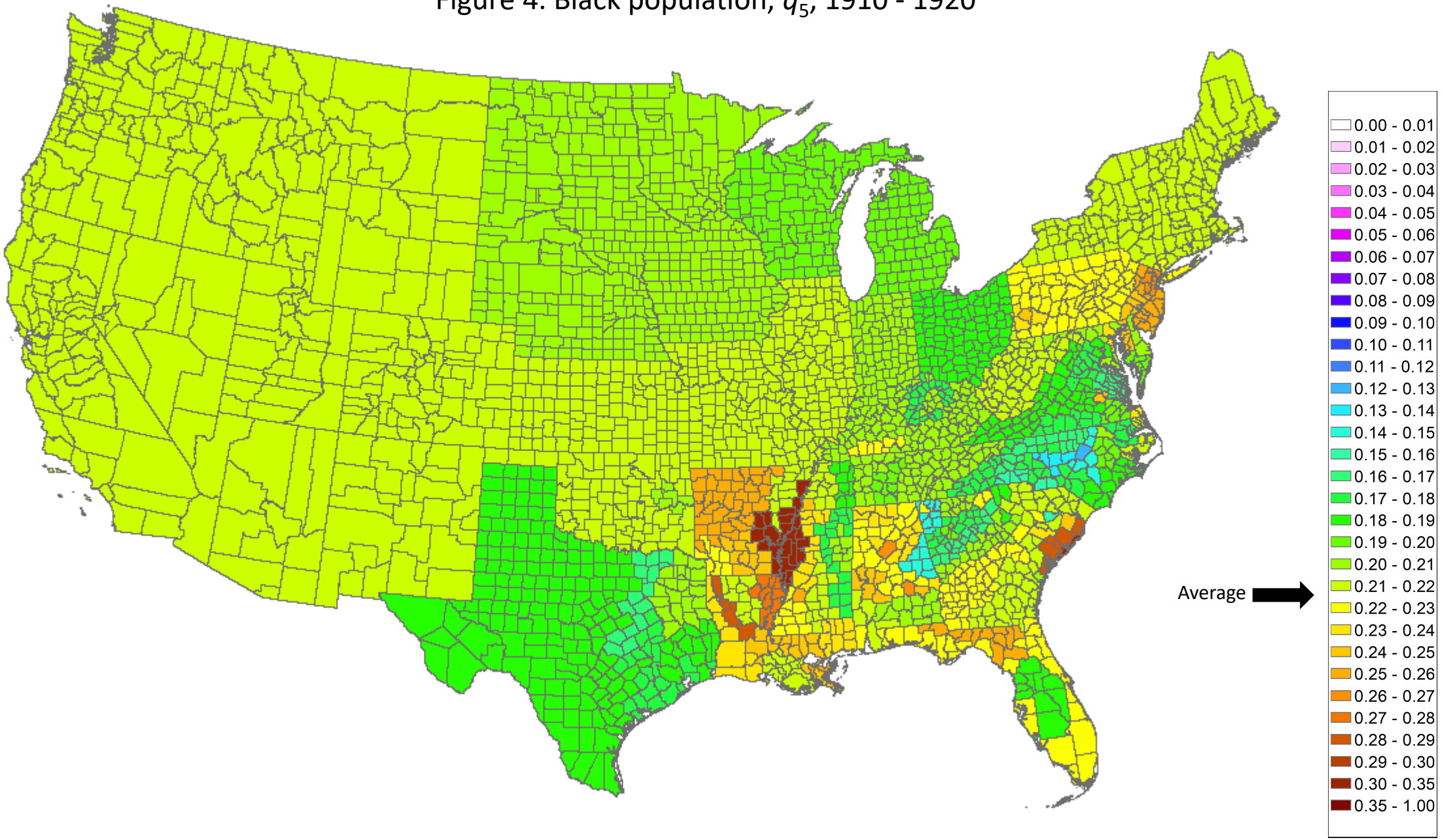


Table A1: Comparison of married couples in the baseline and linked populations

	1850-1860				1860-1870				1870-1880				1900-1910			
	Baseline population		Linked couples		Baseline population		Linked couples		Baseline population		Linked couples		Baseline population		Linked couples	
	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Age																
20 or less	1.2	6.8	0.4	4.5	1.1	6.4	0.4	4.6	1.3	6.3	0.6	4.7	0.7	4.9	0.6	4.1
21-30	26.7	35.7	23.8	35.2	25.4	35.6	23.1	35.2	23.8	33.6	22.8	34.3	21.1	30.3	21.6	32.9
31-40	31.0	27.6	35.6	32.7	31.7	28.1	34.9	32.0	29.1	28.1	32.5	31.8	29.1	28.2	34.3	32.9
41-50	21.3	17.1	25.0	19.0	21.3	16.8	24.6	18.6	22.6	18.0	25.4	19.2	23.1	19.4	25.3	19.6
51-60	11.7	8.4	11.2	6.9	12.6	8.8	12.7	7.7	14.1	9.2	13.7	7.8	14.8	11.1	13.1	8.3
61 or more	8.1	4.4	4.1	1.8	7.9	4.3	4.3	1.9	9.2	4.8	5.0	2.1	11.2	6.1	5.2	2.2
Race																
White	98.3	98.3	99.2	99.2	98.6	98.6	99.3	99.3	88.7	88.7	94.8	94.8	89.7	89.7	94.9	94.9
Black	1.7	1.7	0.8	0.8	1.4	1.4	0.7	0.7	11.2	11.2	5.2	5.2	10.0	10.0	5.1	5.1
Other	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.3	0.0	0.0
Children age 0-5 in household																
0	37.4		28.7		38.6		32.1		42.8		35.3		52.1		42.7	
1-3	61.0		69.2		59.8		66.1		55.9		63.3		46.7		55.8	
4 or more	1.6		2.0		1.6		1.8		1.3		1.4		1.2		1.5	
Occupation of male																
N/A	1.7		0.8		10.4		8.8		4.6		3.6		14.8		13.3	
White collar	8.3		8.1		9.1		9.2		8.6		9.2		9.4		9.8	
Farmers (owners/managers/foremen)	51.4		58.5		40.5		46.3		39.3		45.0		33.3		36.7	
Blue collar	38.2		32.4		36.7		33.0		36.5		34.4		38.4		36.8	
Farm laborers	0.3		0.2		3.3		2.8		11.0		7.8		4.1		3.4	
Region of residence																
New England	15.0		18.7		12.7		15.7		10.2		12.2		7.7		8.2	
Middle Atlantic	30.6		29.6		28.2		28.4		23.9		25.8		21.1		21.1	
East North Central	23.2		23.7		25.9		28.0		24.7		28.0		22.6		25.5	
West North Central	3.8		3.5		7.6		7.4		10.0		10.0		13.8		15.6	
South Atlantic	13.7		13.2		11.6		10.4		13.9		11.8		12.7		10.9	
East South Central	10.6		9.7		8.9		7.2		10.4		7.6		9.4		7.7	
West South Central	2.7		1.5		3.7		1.9		4.9		2.8		7.6		6.3	
Mountain	0.3		0.1		0.4		0.2		0.7		0.5		2.1		1.9	
Pacific	0.1		0.1		1.1		0.8		1.4		1.3		3.1		2.9	
Couples	3,112,446		1,060,063		4,420,733		1,543,261		6,294,108		2,398,823		13,201,563		5,676,195	

1910-1920				1920-1930				1930-1940			
Baseline population		Linked couples		Baseline population		Linked couples		Baseline population		Linked couples	
Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
0.8	4.9	0.6	4.1	0.8	4.4	0.7	3.7	0.8	4.3	0.6	3.8
21.9	30.4	22.5	32.8	20.3	29.1	22.2	32.4	19.2	26.5	20.7	29.7
28.8	28.3	33.8	33.1	28.9	28.6	33.9	33.2	28.0	28.5	33.8	33.6
22.6	19.4	24.9	19.7	23.4	20.2	25.3	20.3	23.8	21.1	25.8	21.2
15.1	10.9	13.2	8.2	15.1	11.4	13.1	8.3	16.1	12.5	13.8	9.3
10.9	6.1	4.9	2.2	11.5	6.4	5.0	2.2	12.1	7.0	5.3	2.4
89.8	89.8	95.0	95.0	90.5	90.5	95.6	95.6	91.1	91.1	95.5	95.5
9.8	9.7	4.8	4.8	9.2	9.2	4.3	4.3	8.6	8.6	4.3	4.3
0.5	0.5	0.1	0.1	0.3	0.3	0.2	0.2	0.4	0.3	0.2	0.2
54.3		45.5		57.9		48.7		63.5		54.4	
44.5		53.1		41.2		50.3		35.8		44.7	
1.2		1.4		0.9		1.1		0.7		0.9	
18.9		16.7		8.1		5.7		19.7		17.6	
9.5		10.4		12.4		13.6		11.6		12.3	
28.7		31.8		25.5		27.4		19.1		19.9	
38.8		37.8		50.4		50.3		46.2		47.2	
4.0		3.4		3.6		3.1		3.4		3.0	
7.1		7.6		7.0		7.2		6.4		7.0	
21.1		21.3		21.1		21.8		21.3		21.7	
21.1		23.7		21.5		23.8		21.8		24.2	
12.9		14.7		12.2		12.7		11.1		12.7	
12.4		10.7		12.0		10.9		11.8		9.9	
8.9		7.4		8.2		7.3		7.8		6.5	
9.2		7.6		9.3		7.5		9.9		8.1	
2.7		2.6		3.3		3.4		2.9		2.9	
4.7		4.5		5.5		5.5		7.0		7.0	
16,836,482		7,174,216		20,482,629		9,145,729		24,914,172		11,606,104	

Table A2: Comparison of children aged 0-5 in the baseline and linked populations

	1850-1860				1860-1870				1870-1880			
	Baseline population		Sample children		Baseline population		Sample children		Census population		Sample children	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Age												
0	15.5	15.4	15.1	14.9	16.4	16.4	16.3	16.3	17.0	17.1	17.3	17.2
1	17.2	16.9	16.9	16.5	16.4	16.3	16.4	16.0	16.7	16.6	16.7	16.4
2	17.5	17.5	17.5	17.3	17.7	17.8	17.6	17.5	17.7	17.7	17.8	17.6
3	16.6	16.8	16.8	17.1	17.1	17.2	17.2	17.3	17.2	17.4	17.3	17.5
4	17.0	17.1	17.1	17.4	16.8	16.7	16.8	17.0	16.6	16.6	16.5	16.7
5	16.3	16.4	16.6	16.9	15.6	15.6	15.7	15.9	14.7	14.7	14.5	14.8
Race												
White	98.0	97.9	99.1	99.2	98.3	98.3	99.3	99.3	85.9	85.6	93.6	93.9
Black	2.0	2.1	0.9	0.9	1.6	1.6	0.7	0.7	14.1	14.4	6.4	6.1
Other	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.0
Occupation of father												
N/A	10.4	10.5	0.6	0.5	18.5	18.5	8.6	8.5	14.8	14.8	2.9	3.0
White collar	6.8	6.9	7.3	7.3	7.1	7.2	8.1	8.0	6.5	6.5	7.9	7.9
s (owners/managers/foremen)	48.5	48.1	59.3	59.4	37.4	37.0	62.7	46.3	36.3	36.3	45.6	45.8
Blue collar	34.1	34.4	32.7	32.7	33.8	34.1	34.3	34.2	31.7	31.7	34.4	34.6
Farm laborers	0.2	0.2	0.1	0.1	3.2	3.2	3.1	3.1	10.7	10.7	9.2	8.9
Region of residence												
New England	10.7	10.8	13.0	13.3	9.0	9.1	10.9	11.2	6.8	6.9	8.2	8.4
Middle Atlantic	27.5	27.9	27.3	27.8	25.5	25.7	26.3	26.6	20.6	20.9	23.0	23.7
East North Central	24.6	24.6	25.9	25.8	27.0	27.1	29.7	29.8	24.3	24.3	28.5	28.9
West North Central	4.6	4.5	4.3	4.2	8.7	8.7	8.9	8.7	11.3	11.1	11.7	11.5
South Atlantic	15.8	15.7	15.2	15.1	13.0	12.9	11.6	11.5	16.1	16.0	13.6	13.2
East South Central	13.0	12.8	12.1	11.8	10.4	10.2	8.8	8.5	12.7	12.6	9.3	8.9
West South Central	3.3	3.2	1.9	1.8	4.5	4.5	2.5	2.4	5.9	5.8	3.6	3.3
Mountain	0.4	0.4	0.1	0.1	0.6	0.6	0.3	0.3	0.8	0.8	0.6	0.6
Pacific	0.1	0.1	0.1	0.1	1.3	1.3	1.1	1.0	1.6	1.6	1.6	1.5
Individuals	1,803,371	1,741,171	702,116	623,861	2,521,224	2,444,303	961,027	837,106	3,281,581	3,197,494	1,380,031	1,192,617

1900-1910				1910-1920				1920-1930				1930-1940			
Census population		Sample children		Census population		Sample children		Census population		Sample children		Census population		Sample children	
Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
17.0	17.0	17.3	17.1	17.7	17.7	17.7	17.6	16.6	16.6	16.9	16.9	15.7	15.7	16.0	15.8
16.0	16.0	16.3	16.1	15.3	15.3	15.6	15.3	16.4	16.4	16.7	16.6	15.5	15.4	15.8	15.6
16.9	16.9	17.0	16.9	17.3	17.2	17.4	17.2	16.7	16.7	16.8	16.7	16.7	16.7	16.8	16.7
16.7	16.8	16.7	16.8	17.0	17.1	17.0	17.2	16.9	17.1	16.8	17.0	17.2	17.3	17.1	17.3
16.9	16.8	16.6	16.8	16.8	16.7	16.6	16.8	16.6	16.5	16.3	16.3	17.0	16.9	16.8	16.9
16.5	16.6	16.1	16.3	15.9	16.0	15.6	16.0	16.9	16.9	16.4	16.5	18.0	18.0	17.5	17.7
86.6	86.1	93.0	93.0	87.4	87.1	93.2	93.2	89.6	89.3	94.2	94.2	88.8	88.4	93.6	93.4
13.1	13.5	6.9	7.0	12.0	12.3	6.6	6.6	10.0	10.3	5.5	5.6	10.6	11.0	6.0	6.2
0.3	0.4	0.0	0.0	0.6	0.6	0.2	0.2	0.5	0.5	0.2	0.2	0.6	0.6	0.4	0.4
18.7	18.9	11.5	11.5	20.3	20.4	13.8	13.9	10.5	10.6	4.0	4.0	21.0	21.1	15.0	14.9
6.5	6.5	7.5	7.4	6.7	6.7	8.0	7.9	8.7	8.7	10.3	10.1	8.0	7.9	24.2	9.1
35.4	35.0	40.6	40.6	32.2	31.9	36.8	36.8	29.1	28.8	32.4	32.3	21.9	21.9	48.1	24.0
35.0	35.2	36.4	36.6	36.3	36.4	37.2	37.3	47.6	47.8	49.5	49.8	44.5	44.5	47.6	47.8
4.5	4.5	4.0	3.9	4.6	4.6	4.2	4.1	4.1	4.1	3.8	3.8	4.6	4.6	4.2	4.2
6.0	6.1	6.4	6.6	6.0	6.0	6.2	6.4	7.0	7.0	7.0	7.1	6.1	6.0	6.6	6.7
18.4	18.5	18.8	19.1	19.1	19.2	19.1	19.5	21.3	21.3	21.5	21.8	19.3	19.3	20.2	20.5
19.5	19.4	22.7	22.8	18.0	17.9	20.6	20.7	19.8	19.7	22.0	22.2	19.4	19.2	21.7	22.0
13.9	13.8	16.3	16.3	12.4	12.3	14.8	14.9	12.2	12.2	12.9	12.9	10.6	10.5	12.3	12.4
15.8	15.9	13.4	13.3	15.5	15.6	13.6	13.5	12.8	12.8	11.6	11.6	15.0	15.1	12.7	12.5
11.6	11.5	9.5	9.3	10.9	10.9	9.4	9.2	8.9	8.9	8.4	8.1	9.8	9.8	8.5	8.3
9.9	9.9	8.4	8.1	11.7	11.6	10.0	9.7	10.5	10.5	9.0	8.6	11.4	11.5	9.8	9.4
2.2	2.3	2.1	2.1	2.9	2.9	2.8	2.8	3.4	3.4	3.6	3.6	3.3	3.3	3.2	3.2
2.6	2.6	2.5	2.5	3.7	3.7	3.5	3.5	4.1	4.1	4.1	4.1	5.2	5.2	5.0	5.0
5,532,311	5,431,083	2,884,265	2,537,426	6,445,911	6,318,415	3,363,437	2,949,456	7,078,904	6,897,451	3,918,969	3,457,487	7,088,779	6,880,448	4,190,645	3,701,716

Table A3. Variables in the Public Use Dataset

Variable name	Label	Values
aggregation_level	Geographic aggregation level	"COUNTYICP", "SEA", "STATEICP", etc.
year_a	Census A year (census year of first linked census)	1850, 1860, 1870, 1900, 1910, 1920, or 1930
census_division	Census division	Census division codes used by ipums.org
stateicp	State	State ICPSR codes used by ipums.org
sea	State Economic Area	State economic area codes used by ipums.org
countyicp	County	State ICPSR codes used by ipums.org
sex	Sex	1 "Male"; 2 "Female"; 3 "Both sexes combined"
race	Race	1 "White"; 2 "Black"
age_0_ucr	Undercount rate of children aged 0 in area	0.000-1.000
age_1_ucr	Undercount rate of children aged 1 in area	0.000-1.000
age_2_ucr	Undercount rate of children aged 2 in area	0.000-1.000
age_3_ucr	Undercount rate of children aged 3 in area	0.000-1.000
age_4_ucr	Undercount rate of children aged 4 in area	0.000-1.000
age_5_ucr	Undercount rate of children aged 5 in area	0.000-1.000
age_0_risk	Number children aged 0 at risk in area	
age_1_risk	Number children aged 1 at risk in area	
age_2_risk	Number children aged 2 at risk in area	
age_3_risk	Number children aged 3 at risk in area	
age_4_risk	Number children aged 4 at risk in area	
age_5_risk	Number children aged 5 at risk in area	
q5_avg_1_3	Implied average proportion of at-risk children aged 1-3 dying before age 5	
e0_avg_1_3	Implied average life expectancy birth of at-risk children aged 1-3	
lx_0	Number in hypothetical cohort (100,000 live births)	100,000
lx_1	Number of surivivors to exact age 1	0-100,000
lx_5	Number of surivivors to exact age 5	0-100,000
lx_10	Number of surivivors to exact age 10	0-100,000
lx_15	Number of surivivors to exact age 15	0-100,000
lx_20	Number of surivivors to exact age 20	0-100,000
lx_25	Number of surivivors to exact age 25	0-100,000
lx_30	Number of surivivors to exact age 30	0-100,000
lx_35	Number of surivivors to exact age 35	0-100,000
lx_40	Number of surivivors to exact age 40	0-100,000
lx_45	Number of surivivors to exact age 45	0-100,000
lx_50	Number of surivivors to exact age 50	0-100,000
lx_55	Number of surivivors to exact age 55	0-100,000
lx_60	Number of surivivors to exact age 60	0-100,000
lx_65	Number of surivivors to exact age 65	0-100,000
lx_70	Number of surivivors to exact age 70	0-100,000
lx_75	Number of surivivors to exact age 75	0-100,000
lx_80	Number of surivivors to exact age 80	0-100,000
lx_85	Number of surivivors to exact age 85	0-100,000
model	Life table model used to construct lx values from implied q5	"Coale & Demeny Model West" etc.

Notes: The number of at-risk children includes undercounted children in census A

The dataset can be accessed at this link: https://drive.google.com/file/d/1r9WtrJlpFjWAdaU4-inVonagDhaBRToQ/view?usp=drive_link

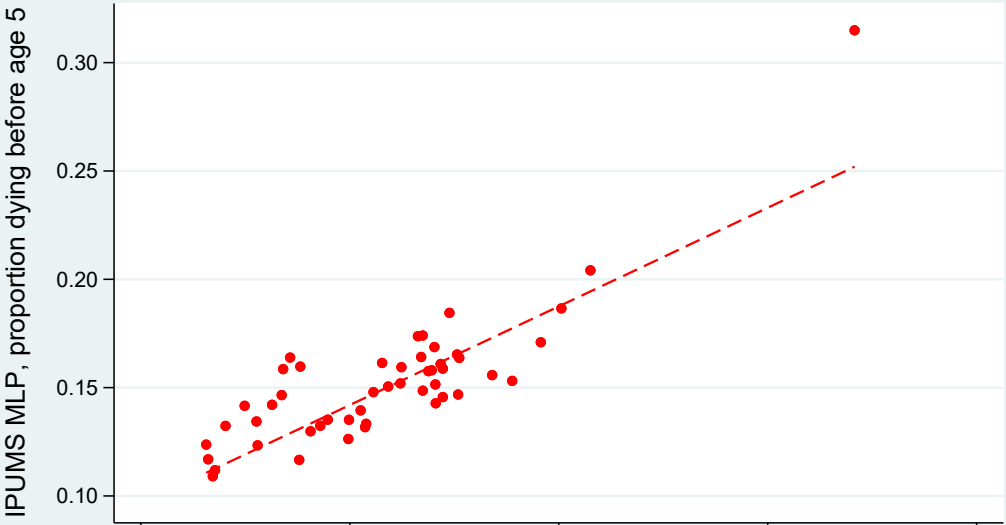
Figure A1: Comparision of indirect estimates of child mortality

White population, 1900-1910 -- County estimates

Weighted regression, R^2 = 0.46, Slope = 0.795 (0.971 without constant)

Figure A2: Comparision of indirect estimates of child mortality

White population, 1900-1910 -- State estimates



Weighted regression, R^2 =0.77, Slope = 0.910 (0.944 without constant)

Figure A3: Comparision of indirect estimates of child mortality

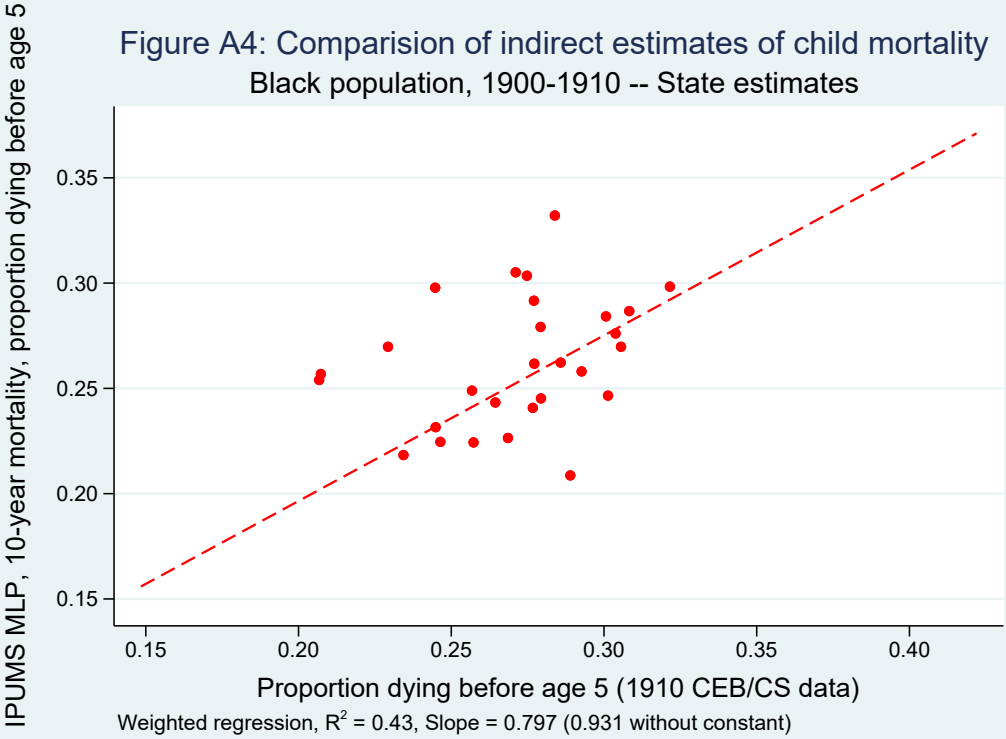


Figure A4: Comparision of indirect estimates of child mortality
Black population, 1900-1910 -- State estimates
IPUMS MLP, 10-year mortality, proportion dying before age 5
0.35
0.30
0.25
0.20
0.15
0.15
0.20
0.25
0.30
0.35
0.40
Proportion dying before age 5 (1910 CEB/CS data)
Weighted regression, R^2 = 0.43, Slope = 0.797 (0.931 without constant)