Methods for Working with Incomplete Administrative Mortality Records: Updating CenSoc from 2006 to 2020

Technical documentation for CenSoc-DMF Data Version 4.0 weights (April 2025 Release)*

Maria Osborne †

Updated: May 17th, 2025

Summary

This technical report describes the creation of statistical weights for version 4.0 of the CenSoc-DMF mortality dataset, which links the Social Security Death Master File with the 1940 US Census. This version extends the previous version of the CenSoc-DMF, which contained deaths from 1975-2005, by adding deaths and weighting deaths from 2006-2020. Since 2005, the public version of the Death Master File (DMF) has been incomplete, including fewer than 15% of deaths recently. Here, we describe changes in the DMF after 2005 and the weighting strategy used to account for this decline in completeness. We find that during the period of overall decline in the DMF, there is likely substantial variation in death reporting at the state level. We weight the new CenSoc-DMF linked data up through 2020 to mortality data from the National Center for Health Statistics on state of birth, race, age of death, and year, to account for changing probabilities of inclusion on these variables over time.

*CenSoc is supported by National Institute of Aging grants R01AG05894 and R01AG076830.
†Department of Demography, University of California, Berkeley. mariaosborne@berkeley.edu.

Contents

1 Overview

The CenSoc project produces large mortality datasets by linking public Social Security Administration (SSA) death records to the 1940 Census. This report describes the creation of weights for the CenSoc-DMF 4.0, a mortality dataset that links the Social Security Death Master File (DMF) to the 1940 Census.

Previous versions of the CenSoc-DMF covered years from 1975-2005. This is because the DMF is an extremely good source of age 65+ mortality data for these years, capturing over 95% of deaths that occurred each year. Recently, we obtained more recent data by purchasing a monthly subscription to the DMF through the National Technical Information Service. However, this version of the DMF is highly incomplete after 2005. As shown in Figure 1, the proportion of all deaths recorded in the DMF declines steadily from around 2005-2015, then drops to under 15% after 2015.

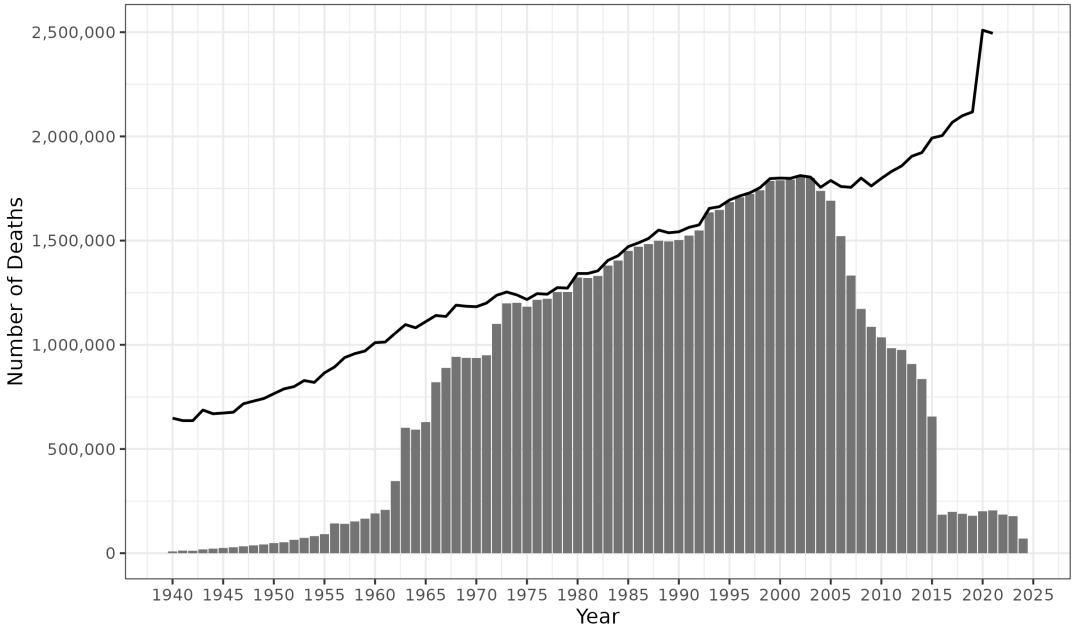

Figure 1: Yearly recorded deaths at aged 65+ in the DMF (bars) compared to the Human Mortality Database (lines). The DMF is nearly complete for years 1975-2005, but far less so outside this period.

This decline after 2005 is due to the exclusion of certain death records from the public version of the DMF, as implemented by the SSA in 2011. This raises significant concerns about potential selection in the DMF after 2005, especially as little is known about the nature of death records that are included or excluded from the public version of the file. The DMF itself contains almost no information on decedents, making it difficult to assess which records are released to the public or the possible consequences of this undercoverage of deaths.

To create this version of the CenSoc-DMF, we first link the 1940 Census to the DMF using a conservative variant of the ABE record linkage algorithm (Abramitzky et al., 2021). We do not find strong evidence that the decline in DMF completeness introduces

notable selection on demographic characteristics or socioeconomic status as measured in the census. As with version 3.0 of the CenSoc-DMF dataset, we create weights using vital statistics data from the National Center for Health Statistics (NCHS). Records are weighted for each year on age at death, race, and location of birth. This allows us to account for differing inclusion probabilities over time by these characteristics, most importantly state. For the majority of records (decedents born in the contiguous United States and who die 1979-2020), weights are calculated using logistic regression and raking. Alternate weights are constructed for other populations for which weighting directly to NCHS counts of death is inappropriate or impossible, including the foreign-born.

The remainder of this report is organized as such: in Section 2, I further describe DMF data. Section 3 describes data linkage and the unweighted CenSoc-DMF dataset. In Section 4 I discuss the weighting methodology in detail. In Section 5 I summarize the weights generated and demonstrate use in regression analyses. Finally, Section 6 summarizes the findings of this report and concludes with considerations for researchers.

2 Data

2.1 The Death Master File

The Death Master File (DMF) is a record of over 100 million deaths to persons assigned social security numbers (SSNs), created by the Social Security Administration (SSA). Maintained since 1962, it contains records of deaths dating to about 1937 and is updated monthly, with new deaths typically appearing in the DMF within months of their occurrence. The SSA receives notification of death from numerous sources, including states, federal agencies, family members, funeral homes, hospitals, postal authorities and financial institutions (Social Security Administration, 2012).

DMF records first became available to the public in 1980 due to a Freedom of Information Act (FOIA) lawsuit (Social Security Administration, 2012). We use the public DMF purchased on a subscription basis from the National Technical Information Service (NTIS). It contains regularly-updated, publicly-available death data. The only variables in this version of the DMF are SSN, name, date of birth, and date of death.

Unfortunately for researchers, policy changes surrounding state-provided death records have drastically affected the nature of public DMF data after about 2005. While a complete version of the DMF is used internally by the Social Security Administration and other government agencies, the public version of the DMF accessible to non-governmental entities (henceforth referred to simply as "the DMF") now includes only a small percentage of deaths that occur each year. In 2011, the SSA determined that state-owned death records not covered by the FOIA due to section 205(r) of the Social Security Act had been improperly included in the DMF (Levin et al., 2019; Da Graca et al., 2013). This

led to the removal of about 4.2 million protected state records prior to November 1, 2011, largely affecting years after 2005, and millions fewer deaths per year added to the file going forward (National Technical Information Service, nd).

This has drastically impacted overall completeness of the DMF overall in recent years (refer to Figure 1). Because changes to the DMF were related to inclusion of state death records, individual state policies and methods of death reporting are likely responsible in part for shaping the DMF after 2005. As shown in Figure 2, the timing of this decline happened at different points between 2005-2015 for individual states. Here, we use the first 3 digits of the SSN to determine what state that SSN was assigned in. While not equivalent to state of death or state or birth, this gives us some insight into how the geography of state death records may be changing over time in the DMF.

The number of records included in the DMF from states such as Minnesota and New Hampshire begin to significantly decline as early as 2004-2005. For others, such as North Carolina, a substantial drop is not observed until 2016. For many states, the number of records falls sharply in the span of only a few years. After the period of about 2005-2015 where individual state coverage declines at different rates, coverage across most states is universally poor from 2016 onward. Our findings are consistent with Navar et al. (2019), who find that undercoverage in the DMF varies over both state and time.

The reasons for these state-specific time trends are not well understood by researchers, as very little public record of changes to the DMF exists. Some patterns may be attributable to the early adoption of electronic death registration (EDR) systems in states such as California and Montana (Social Security Administration Office of the Inspector General, 2017), as deaths reported using EDR systems were retroactively removed from DMF data. However, use of EDR cannot fully explain these patterns – North Carolina did not launch an EDR system until 2020 (North Carolina Medical Board, 2020), but deaths for persons assigned a SSN in North Carolina drop drastically after 2015.

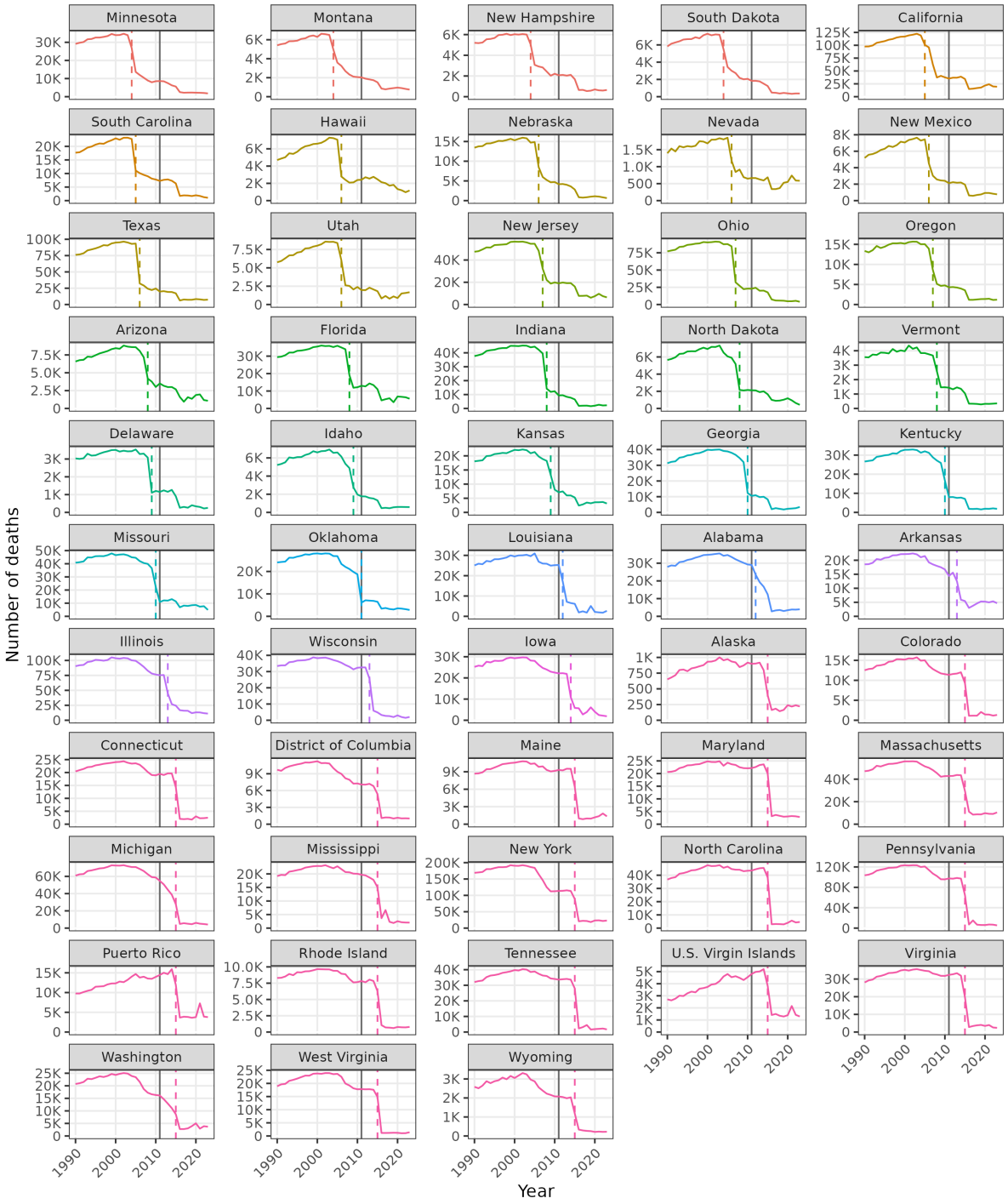

Figure 2: Number of deaths per state, based on state of SSN assignment. Dashed lines indicate the first year where the state experienced a significant year-to-year (>15%) decline in death records. Top left states (such as Minnesota) experience this drop early (around 2004-2005) compared to states in the bottom left, such as Virginia, which do not drastically decline until about 2016. 2011, the year that policy changes affecting completeness of the public DMF were implemented, is marked by a solid black line on each plot.

3 Linkage of the DMF to the 1940 Census

We link the DMF to men in the 1940 Census using a conservative variant of the ABE automated record linkage algorithm (Abramitzky et al., 2021). Records are linked on standardized first name, last name, and age at time of the 1940 Census, which in the DMF is calculated using date of birth information. Due to lack of information on surname changes in the DMF, we link only men between the datasets.

For men born after 1900, the census linkage rate (percent of men in the census that are linked the DMF) is around 10%, with some variation by cohort. This rate is unadjusted for mortality; people observed in the 1940 Census but who did not die between 1975 and 2020 are impossible to link, which limits the potential match rate. The census linkage rate peaks for birth cohorts circa 1910-1920, whose members are likely to die in the observable window.

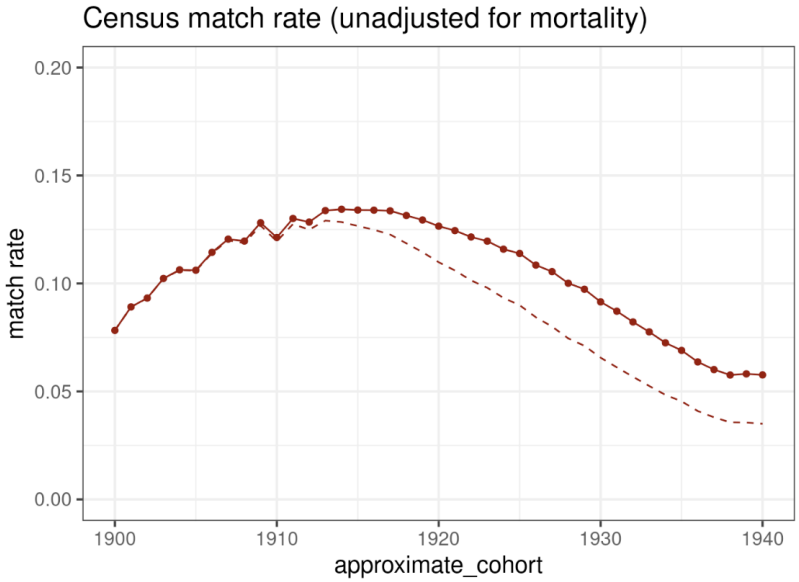

Figure 3: CenSoc-DMF Match Rate (proportion of males in the 1940 Census linked to the DMF) by cohort. The solid line indicates linkages through 2020, in comparison to linkages through 2005 as indicated with the dashed line. Linking records through 2020 primarily increases the census linkage rate for the cohorts after about 1910.

3.1 Assessment of the Unweighted Linked DMF

Undercoverage in the DMF after 2005 has clear impacts on the age distribution of deaths for later cohorts in the CenSoc-DMF. Figure 4, for example, compares the unweighted distribution of deaths for an earlier birth cohort (1905) and a more recent cohort (1930). Deaths for the 1905 cohort, which was largely extinct by 2005, peak around age 80 and taper off afterwards. Conversely, for the 1930 cohort, which turned 75 in 2005, the distribution of older death ages is primarily shaped by the decline in DMF coverage.

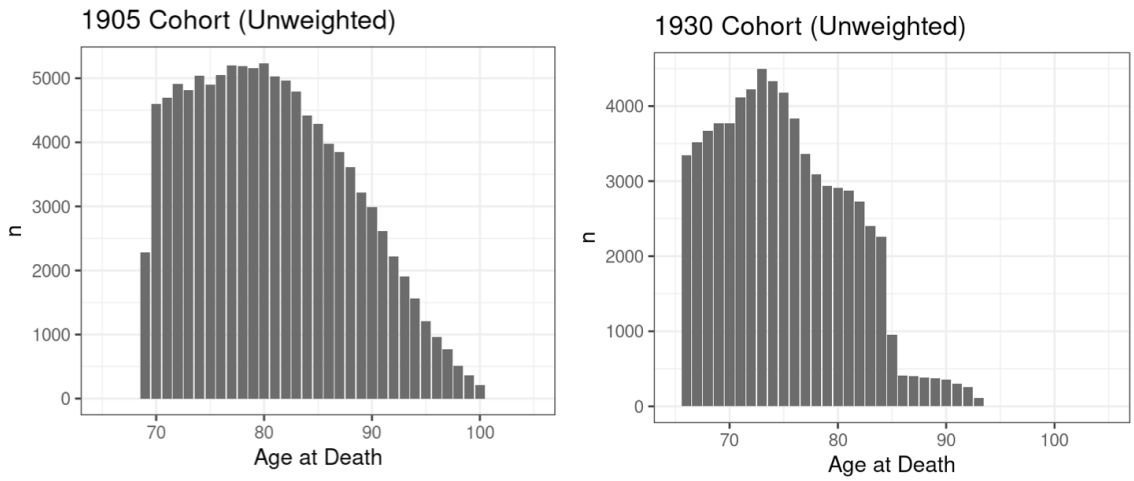

Figure 4: Unweighted cohorts of 1905 (left) and 1930 (right), demonstrating effect of DMF undercoverage on distribution of deaths for more recent cohorts.

Due to lack of variables in the DMF, our ability to assess the DMF on its own is very limited. This makes it difficult to study any selection introduced by undercoverage in the DMF after 2005. By linking to the 1940 Census, however, we can use variables from the census to examine possible changes in the composition of the DMF over time, at least for the linked sample. For example, in Figure 5 and Figure 6 we use race and education as reported in the 1940 Census to assess whether the socioeconomic composition of the linked CenSoc-DMF meaningfully changes after 2005. Across most cohorts, the trends after 2005 appear to align with pre-2005 trends, though small sample sizes after 2005 introduce some noise. Similar results were found for homeownership and urbanicity. While the number of deaths in the DMF after 2005 is certainly much lower that actual deaths in the population, we generally do not observe that this undercoverage causes dramatic changes in representation of certain groups in the linked dataset.

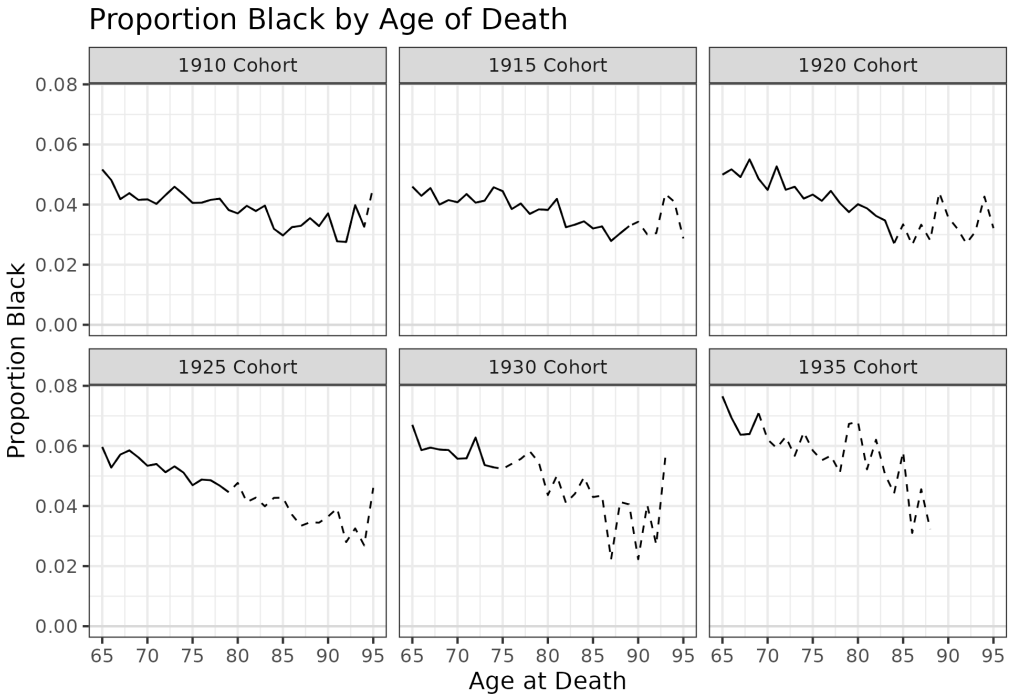

Figure 5: Proportion of decedents at each age who are Black for different birth cohorts. Unweighted data. Solid lines indicate deaths prior to 2005, and dashed lines after 2005.

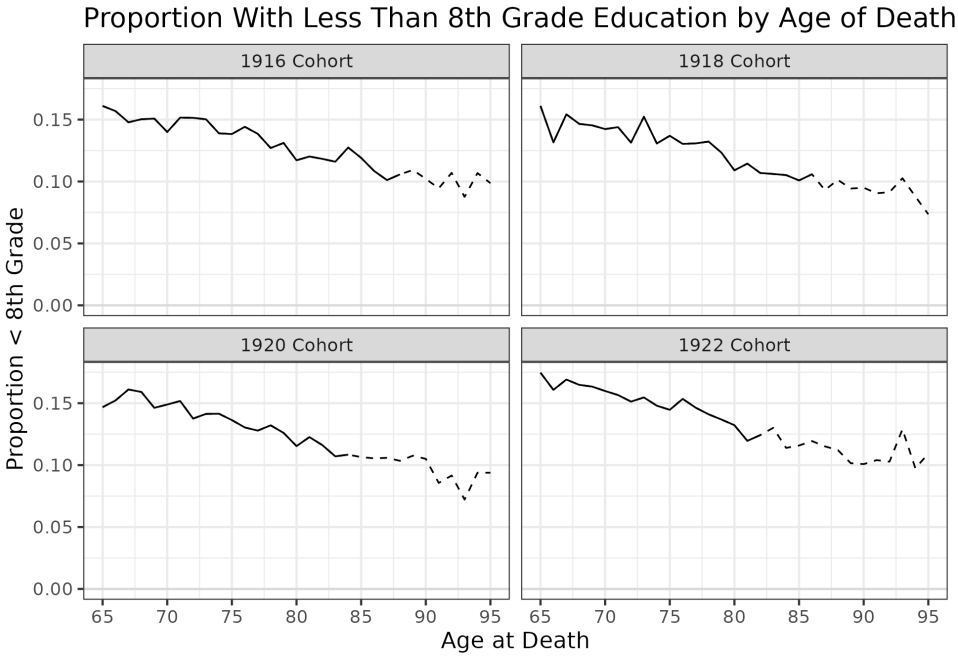

Figure 6: Proportion of decedents at each age who have less than an eighth grade education for different birth cohorts. Solid lines indicate deaths prior to 2005, and dashed lines after 2005.

4 Weighting Method

4.1 Outline

We weight records for decedents aged who died age 65-100 in the years 1975-2020, from birth cohorts 1938 and earlier. Multiple Cause of Death data from NCHS, a comprehensive source of mortality data based on virtually all death certificates filed in the United States, are used for weighting. (National Center for Health Statistics, 2016, 2023). Weights are designed based on year of death, age at death, race (Black, White, or Other), and state of birth. This procedure is outlined as follows:

1. **Calculate standard weights:** For records belonging to people who die between the ages of 65-100, in the years 1979-2020, and who were born in the contiguous United States including the District of Columbia, we compute a base weight using logistic regression for each year of data. Then, base weights are calibrating to align with population marginal totals using raking.

2. **Calculate out-of-period weights**. Multiple Cause of Death data do not contain birthplace information for the years 1975-1978. For people who die in these years, we create weights by extrapolating regression coefficients using a smoothing spline.

3. **Weight non-US birthplaces**. People born outside the United States (including Alaska, Hawaii, current US territories, and foreign nations), are only present in CenSoc data if they moved to the contiguous US before census day in 1940. NCHS data include immigrants who entered the country after 1940 and all deaths from Alaska and Hawaii, so the NCHS population does not align with the CenSoc population in these cases. People born outside the contiguous US or whose birthplace is unknown are assigned weights based on averages of the US-born population.

As a prerequisite to weighting, we must "match" the linked CenSoc-DMF with the NCHS data to assess which deaths out of the population are represented in the CenSoc-DMF. Since there is no way to do with shared unique identifiers such as SSN, this is done based on strata (unique combinations of of year, age, race, and birthplace). For example, the table below shows a made-up example of the NCHS population data table, where each row is a deceased person. We define the variable *matched*=1 if a person of the same strata appears in the CenSoc-DMF, and *matched*=0 if not:

row	death_year	death_age	race	sex	birthplace	matched
1	2016	85	White	M	Idaho	1
2	2016	85	White	M	Idaho	1
3	2016	85	White	M	Idaho	0
4	2016	65	Black	M	Missouri	0

In this example, there are three people in the blue strata in the NCHS mortality data (*death_year*=2016, *death_age*=85, *race*=white, *birthplace*=Idaho). Two people in the CenSoc-DMF match this strata, so two rows are assigned *matched*=1 and the remaining row is assigned a 0. Some strata will not be represented in the CenSoc-DMF at all, like the yellow strata above (*death_year*=2016, *death_age*=65, *race*=black, *birthplace*=Missouri). If there is one person in this strata in the population but no-one of matching strata CenSoc-DMF, then this row is assigned *matched*=0. The *matched* field is then used as the dependent variable in the following regressions.

4.2 Standard Weights

For persons in the CenSoc-DMF who were born in the contiguous United States and died 1979-2020, we use logistic regression and raking (iterative proportional fitting) to create weights. The probability of any record being included in the CenSoc-DMF (p_i) is computed using a logit model for each individual year y model based on age at death, race, and birth state:

$$\text{logit}_y(p_i) = \beta_0 + \beta_1 deathAge + \beta_2 race + \beta_3 birthState$$

A base weight is then computed as:

$$\text{base weight} = \frac{1}{p_i}$$

To produce final weights, the base weights are then adjusted using raking, which ensures that the weighted marginal totals equal the population marginal totals. This procedure is also done separately for each year, so marginal totals by age, race, and birth state within each year are consistent between NCHS mortality data and weighted CenSoc-DMF data.

Weighting by individual years allows us to capture differences in inclusion probabilities over time, especially after 2005, as shown in Figure 7. Dramatic year-to-year changes in logit regression coefficients for birth state can occur after 2005, as individual state policies may have sudden impacts on state-level reporting in the DMF. Additionally, there are more gradual time trends in other variables used to weight data. For example, the odds of a Black decedent's inclusion in the CenSoc-DMF relative to a While decedent has steadily increased since about 1990. While Black decedents are less likely to appear in the CenSoc-DMF in all years, this trend suggests that the Black/White racial gap in the DMF may have narrowed over time.

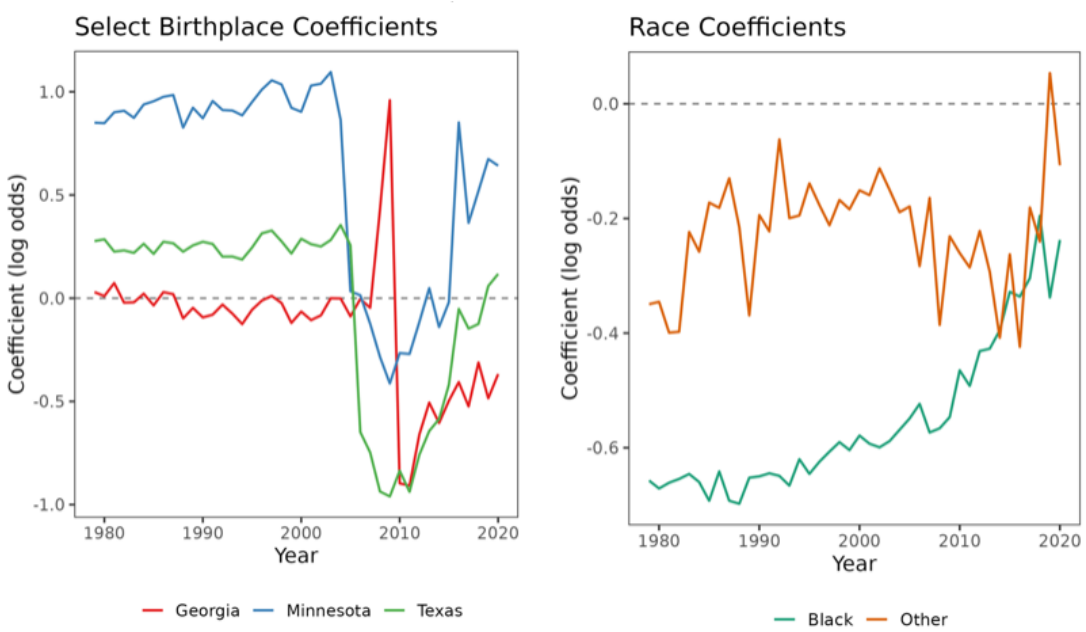

Figure 7: Logit coefficients over time for select states of birth (left panel) and race categories (right panel). A negative coefficient indicates that a group is less likely to be included in the linked CenSoc-DMF than the reference group, while positive cofficient indicates the opposite. For birth states, coefficients are relatively stable before 2005 but can change dramatically year-to-year after this point (references state = Alabama). For race coefficients, both Black and Other-race decedents are less likely to appear in the linked CenSoc-Numident than the reference category, White decedents. While the Other-race coefficient is fairly noisy, the Black coefficient trends upwards towards 0 over time.

4.3 Deaths 1975-1978

The high coverage period of the CenSoc-DMF includes deaths in the years 1975-1978, a period for which NCHS does not publish the birthplace of decedents. We generate these weights for the US-born in these years by extrapolating from the 1979-2020 logit model coefficients. This is done with a smoothing spline with low degrees of freedom to capture the general trends in coefficients. In Figure 8, the raw and smoothed coefficients for select ages at death are shown, with smoothed coefficients extrapolated back to 1975. As with standard weights, these are converted to predicted probabilities and the inverse is taken to compute a base weight. However, weights for these years are not raked, as due to lack of birthplace information in NCHS data, the population totals for US-born decedents is unknown.

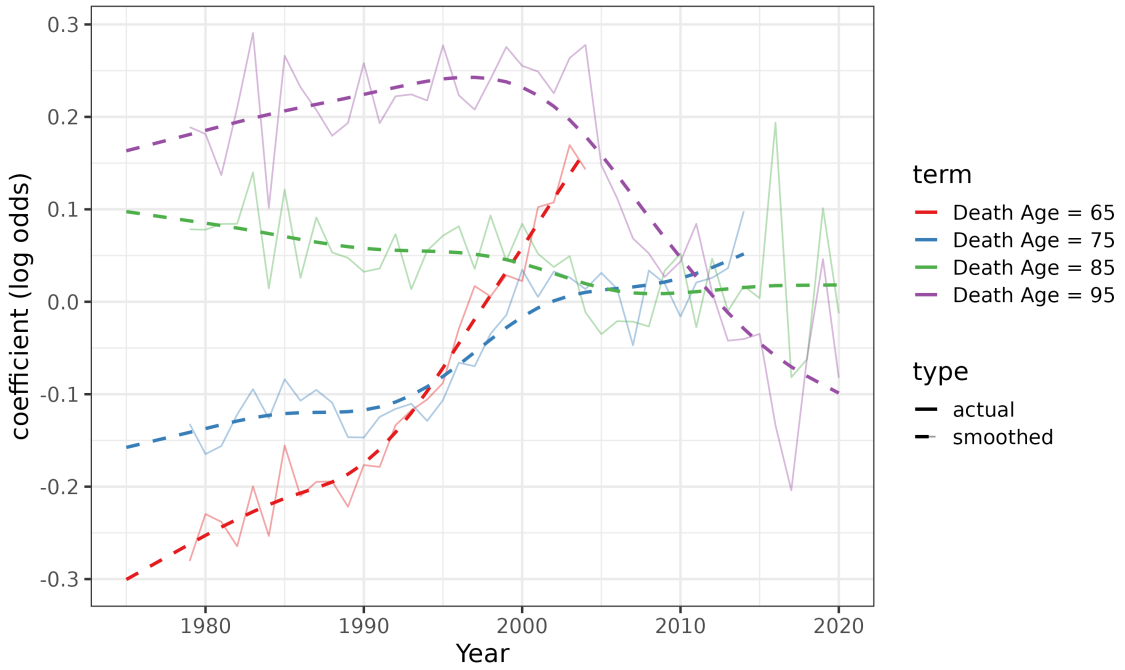

Figure 8: Actual and smoothed age coefficients of each yearly logit model. Smoothed coefficients have been extrapolated back to 1975 in order to construct weights for years 1975-1978. The reference age used for all years is 82.

4.4 Non-US birthplaces

We do not directly weight people in the CenSoc-DMF who were born outside the contiguous United States (including foreign nations, Alaska, Hawaii, Puerto Rico, Guam, American Samoa, the US Virgin Islands, and the Northern Mariana Islands). The CenSoc-DMF only includes such people if they moved to the contiguous US before census day in 1940[1]. NCHS data, conversely, include people who were born before the 1940 Census but did

[1]Alaska and Hawaii were not states in 1940. While territorial censuses were conducted, census microdata for these areas are not available from IPUMS.

not enter the country until later. Weighting this population in the CenSoc-DMF directly to NCHS would thus lead to artificially inflated weights, especially for more recent birth cohorts of migrants who are less likely to have entered the country before 1940. There are also a small number of people with unknown birthplace in the CenSoc-DMF (0.03% of records in the high-coverage window), and so cannot be directly weighted based on place of birth.

For all such individuals, we assign weights based on age, year, and race using the native-born Americans as a standard. For each person with a non-US birthplace and year of death y_i, race r_i and age at death a_i,

$$W_{y_i r_i a_i} = \text{mean}(W_{y_i r_i a_i}) \text{ among US-born}$$

Thus all non-American born records in the same year/sex/age stratum are assigned the same weight, regardless of exact country or territory of birth. For example, a White man born in Canada dying at 75 in the year 1995 receives the average weight of all US-born White men who die at age 75 in 1995. For very rare cases where this procedure cannot be applied because there is no analogous strata of the same year, age, and race in the US-born population, decedents are assigned the mean weight of that year for persons of the same race.

5 Analysis of the Weighted CenSoc-DMF

5.1 Weights by age and year

Figure 9 Shows mean weights for each CenSoc data set on lexis surfaces. This has been split up into three different period (1975-2005, 2006-2015, and 2016-2020), as the range of weights assigned in each period is very different. In the 1975-2005 period, when the DMF is a nearly-complete source of deaths, weights are relatively low. There are noticeable age patterns in weights, with younger ages at death generally weighted more heavily than older ages at death, though overall variance is also low in this period. During the 2006-2015 period, average weights rise by year as undercoverage in the DMF increases, but age appears to be less important. After 2016, weights by single age and year are high and there is more variation in weights, making any patterns difficult to discern.

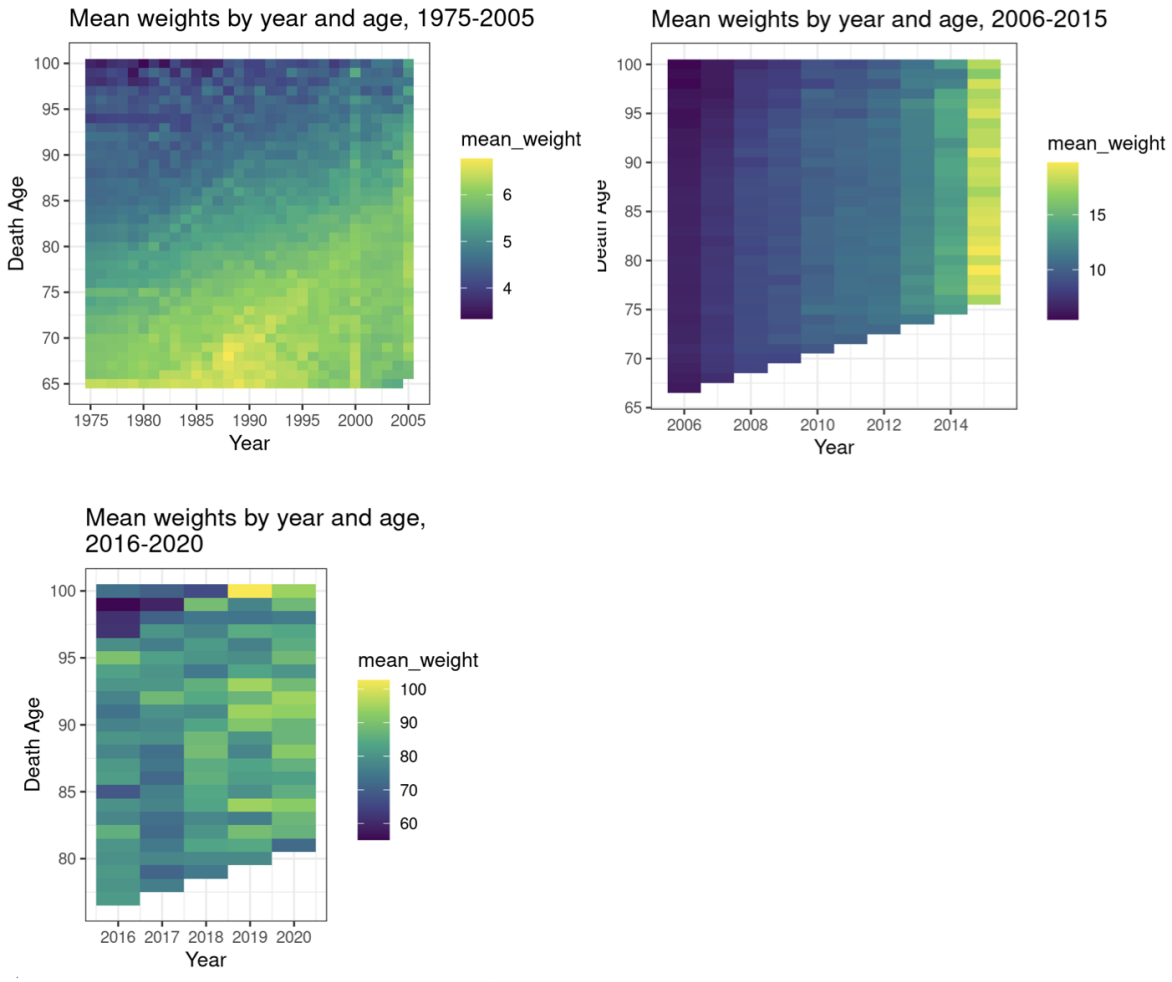

Figure 9: Mean weights by combination of age and year of death. Periods 1975-2005, 2006-2015, and 2016-2020 are plotted separately, as the range of mean weights differs substantially between these three periods. Weights are highest after 2015, the period where DMF coverage is lowest.

5.2 Impact of weights on representativeness

In the following tables, we compare the characteristics of the 1940 Census population to the subset of the population that was linked to a death record and included in the CenSoc-DMF. The proportional composition of both weighted and unweighted CenSoc-DMF data are included. Certain populations, such as Black people, are underrepresented in the CenSoc-DMF. In addition to lower linkage rates among such groups, these differences may be caused by greater mortality among Black men prior to the observation window of the CenSoc-DMF. In general, more socioeconomically disadvantaged groups are underrepresented in the DMF, but weighting the data usually brings the composition of the CenSoc-DMF into closer alignment with the 1940 Census.

Table 1 shows the composition of the 1940 Census vs the linked CenSoc-DMF low-coverage period (2005 onward). This table includes only those men who were aged 2-16 at the time of the Census, who are more likely to die in the low-coverage period of the DMF than men born in earlier cohorts. This table shows that there are some compositional differences in the general census population compared to the subset where a link to the CenSoc-DMF was established. The CenSoc-DMF under represents Black men, men who lived in rented homes, as well as those who lived in the South Atlantic and East South Central regions (these include states such as Florida, Maryland, Alabama, and Tennessee). Conversely, Whites and Middle Atlantic/East North Central regions(including New York, Pennsylvania, and Illinois), are overrepresented in the CenSoc-DMF. In all cases, weights bring proportional representation more closely into alignment with the 1940 Census.

	Count		Proportion (%)			Difference (%)
	CenSoc-DMF	1940 Census	CenSoc-DMF (unweighted)	CenSoc-DMF (weighted)	1940 Census	Weighted DMF - Census
Race						
Black	25,656	1,929,746	5.0	8.3	11.2	-2.9
Other	2,186	92,324	0.4	0.7	0.5	0.2
White	480,916	15,152,691	94.5	91.0	88.2	2.8
Home Ownership						
Live in owned home	223,788	6,803,653	44.0	43.1	39.6	3.5
Live in rented home	284,970	10,371,108	56.0	56.9	60.4	-3.5
Rural						
Rural	252,660	9,013,916	49.7	51.8	52.5	-0.7
Urban	256,098	8,160,845	50.3	48.2	47.5	0.7
1940 Region						
New England	41,869	1,005,443	8.2	6.2	5.9	0.3
Middle Atlantic	126,099	3,218,348	24.8	19.7	18.7	1.0
East North Central	116,384	3,240,518	22.9	19.6	18.9	0.7
West North Central	55,280	1,740,013	10.9	10.9	10.1	0.8
South Atlantic	57,641	2,685,190	11.3	13.9	15.6	-1.7
East South Central	35,052	1,716,644	6.9	9.0	10.0	-1.0
West South Central	37,142	1,949,799	7.3	10.8	11.4	-0.6
Mountain	15,470	599,037	3.0	3.8	3.5	0.3
Pacific	23,821	1,019,769	4.7	6.1	5.9	0.2

Table 1: Comparing the composition of all men aged 2-16 in the 1940 Census with men aged 2-16 who were linked to the low-coverage period (post-2005) of the DMF. The rightmost column is the proportion of people in the weighted CenSoc data with a characteristic minus the proportion of people in the 1940 census with that characteristic A negative difference indicates that a group is underrepresented in the CenSoc-DMF compared after weighting. For these characteristics, weights serve to align the CenSoc-DMF population more closely with the 1940 Census. Compositional differences may remain either because of systemic errors in linkage/weighting, or because of differential mortality by group prior to 2005.

5.3 Example Analyses

In this sections, I explore the effects of using weights with a few simple OLS regressions. In Figure 10, the relationship between state of birth and longevity after age 65 is plotted. Coefficients indicate effect on longevity (in years) of being born in Alabama, relative to being born in Minnesota. For earlier cohorts, those born in Alabama live slightly shorter lives than those born in Minnesota, and estimates are relatively consistent between weighted and unweighted data. For later cohorts, if weights are not used, people born in Alabama appear to have a major longevity advantage over those born in Minnesota. This is due to the fact that death coverage in the DMF starts to decline earlier for Minnesota than it does for Alabama, which makes it appear like people born in Minnesota die at disproportionately younger ages. Using weights corrects for this issue of differing state-level coverage over time in the DMF.

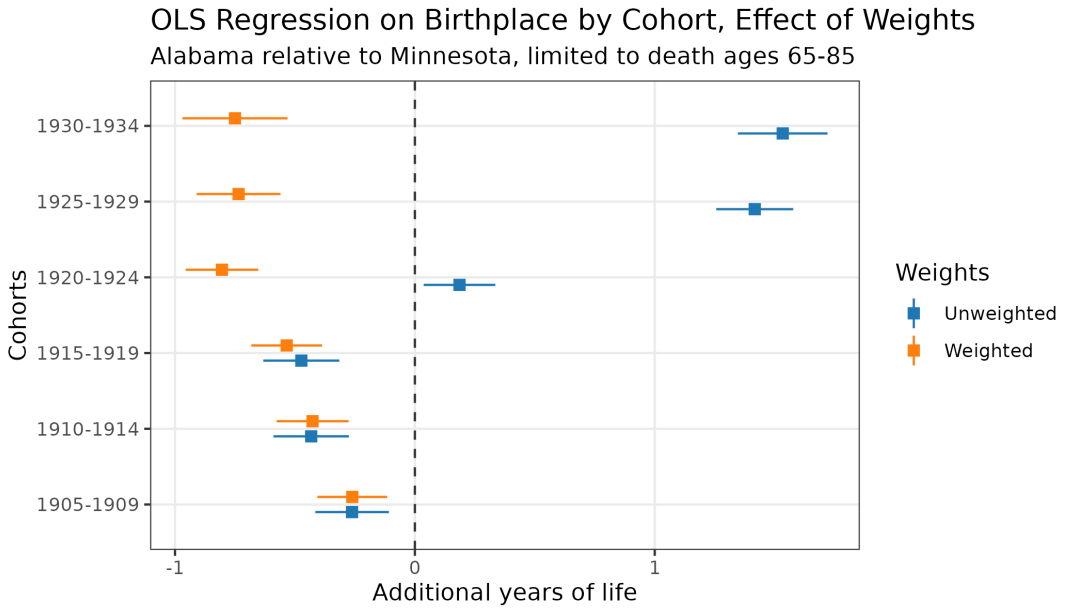

Figure 10: Effect of using weights in OLS regression on state of birth. Without weights, estimates for younger cohorts are unexpected (implying people from Alabama live for more than a year longer on average than those from Minnesota), due to differing coverage of these states in the DMF after 2005. Using weights corrects for this.

The above is a more extreme example, as state-level death reporting in the DMF is highly variables from 2005-2015 and thus can directly impact geographic mortality comparisons to a high degree. For most analyses within an OLS framework, weights are less likely to change inference from models. As another example, we analyze the relationship between homeownership and longevity for different groups of cohorts (Figure 11). In this case, persons who live in rented homes live shorter lives on average than those in owned houses, regardless of cohort. Weights have some impact on the magnitude, but not the direction, of the estimated effect.

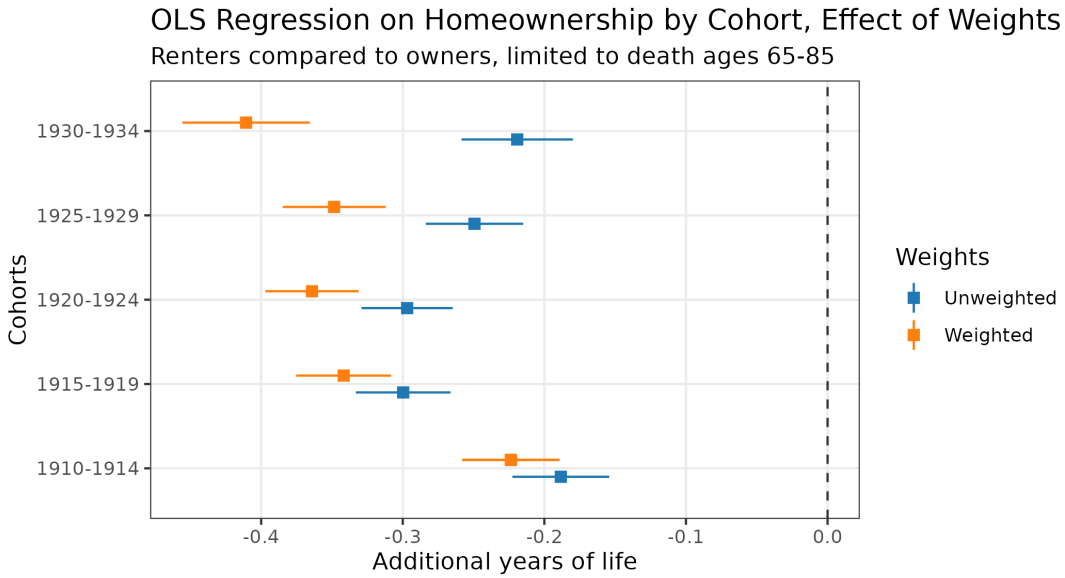

Figure 11: Effect of using weights in OLS regression on homeownership. Weights do affect the magnitude of the estimated effect, especially for later cohorts, but do not change the direction of the estimates.

6 Conclusions

This report details the nature of the Death Master File in recent years, our creation of the linked CenSoc-DMF dataset 1975-2020, and methodology for weighting the linked data. We summarize our process and findings as follows:

1. Undercoverage in the DMF is significant after 2005, and especially after 2015.

2. Geographic coverage varies in the DMF in consequential ways. This is true even before 2005, but especially after 2005 when inclusion probabilities by state of birth can change suddenly and drastically. State-level coverage appears to drop off sometime between 2005 and 2015 for most states, with variation in the timing of major declines. This can lead to unexpected results if not accounted for, such as life expectancy over 65 for people born in Alabama appearing to be much longer than that for people born in Minnesota. After 2015, coverage in the DMF appears to be universally low and less dependent on state.

3. The probabilities of being included in the linked CenSoc-DMF do change over time with respect to race and age, but more gradually than those for state. For example, Black decedents, while still less likely to be captured in the linked CenSoc-DMF than White decedents in all years, become increasingly likely to appear in the CenSoc-DMF relative to Whites over time.

4. We weight to population mortality data from the NCHS by birth state, race, and age of death by each year by raking, helping to correct for these dimensions of undercoverage and how it changes over time.

6.1 Usage and Recommendations for Researchers

The CenSoc-DMF links men born by April 1, 1940 to the Death Master File, 1975-2020. The dataset contains 7 fields, indicating date of birth, date of death, age of death, HISTID, and weight. The HISTID variable is a unique identifier that allows users to link the CenSoc-DMF to 1940 Census extracts from IPUMS. The table below shows a sample line of the CenSoc-DMF. This example is the record for Marlon Brando, the famous American actor, who was born in April of 1924 and died at age 80 in July of 2004:

HISTID	byear	bmonth	dyear	dmonth	death_age	weight
7AD219A2-4821-4A94-880E-BAEC19FEFF0F	1924	4	2004	7	80	4.717

We create weights for the data for decedents who: 1) died at ages 65-100, and 2) were born before 1939. We recommend focusing analyses on the cohorts of about 1900 - 1930, which are shown below on a lexis surface in Figure 12. Members of these cohorts

were observed at ages 10-40 in the 1940 Census, and are likely to die at age 65+ in the window of 1975-2020. Cohorts older than about 1920 are extinct or nearly-extinct by 2020, possibly ameliorating the potential issues caused by working with truncated data, such as attenuation bias that can increase with the severity of truncation (see Goldstein et al. (2023)).

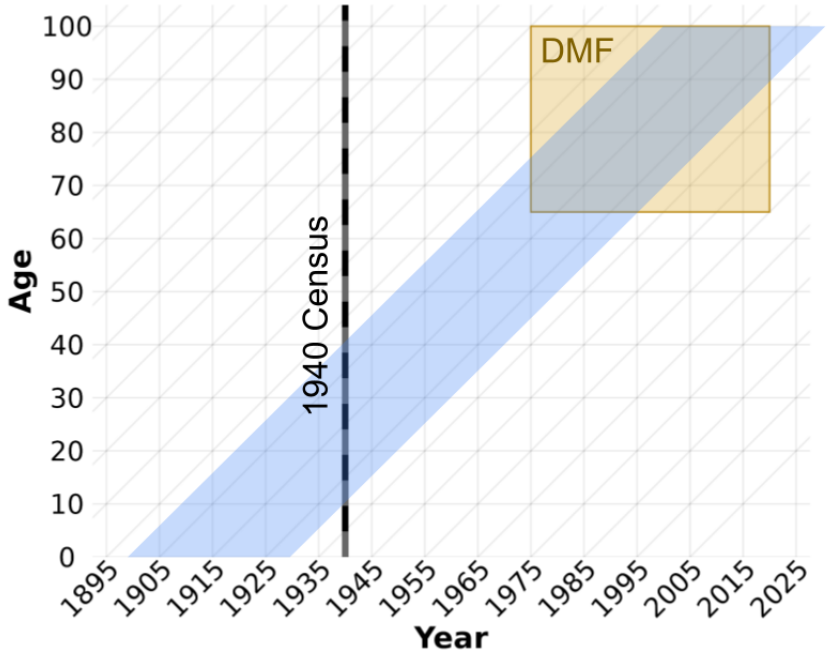

Figure 12: Cohorts of 1900-1930 shown on a Lexis surface with the CenSoc-DMF, which links data from the 1940 Census to DMF mortality data 1975-2020.

Users of the dataset should be aware for the following limitations:

1. Only birth cohorts through 1938 and ages 65-100 are weighted. Unweighted cohorts and ages are available in the data but should be used with caution.

2. Due to the paucity of deaths in the DMF after 2005, and especially after 2015, smaller groups may not be well represented in mortality data after these years.

3. In general, we recommend exercising some caution when working with cohorts later than about 1930. Undercoverage in the DMF affects later cohorts more severely, as members of these cohorts are more likely to survive past 2005, and uncorrected artifacts relating to undercoverage may remain in the data. Data for these cohorts are also truncated at lower ages of death, which may cause issues with model fit.

References

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

Da Graca, B., Filardo, G., and Nicewander, D. (2013). Consequences for Healthcare Quality and Research of the Exclusion of Records From the Death Master File. *Circulation: Cardiovascular Quality and Outcomes*, 6(1):124–128.

Goldstein, J. R., Osborne, M., Atherwood, S., and Breen, C. F. (2023). Mortality modeling of partially observed cohorts using administrative death records. *Population Research and Policy Review*, 42(36).

Levin, M. A., Lin, H.-M., Prabhakar, G., McCormick, P. J., and Egorova, N. N. (2019). Alive or dead: Validity of the Social Security Administration Death Master File after 2011. *Health Services Research*, 54(1):24–33. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6773.13069.

National Center for Health Statistics (2016). Multiple Cause of Death Data, 1979-2004.

National Center for Health Statistics (2023). Detailed mortality – limited geography (states only) (2005-2021), as compiled from data provided by the 57 vital statistics jurisdictions through the vital statistics cooperative program.

National Technical Information Service (n.d.). Change in Public Death Master File Records.

Navar, A. M., Peterson, E. D., Steen, D. L., Wojdyla, D. M., Sanchez, R. J., Khan, I., Song, X., Gold, M. E., and Pencina, M. J. (2019). Evaluation of Mortality Data From the Social Security Administration Death Master File for Clinical Research. *JAMA Cardiology*, 4(4):375–379.

North Carolina Medical Board (2020). State rolls out electronic death registration system.

Social Security Administration (2012). Fact Sheet: Change to the Public Death Master File (DMF).

Social Security Administration Office of the Inspector General (2017). State Use of Electronic Death Registration Reporting. Technical Report A-09-15-50023.