CenSoc-DMF Weights: Technical Documentation

For CenSoc-DMF Data Version 4.0, April 2025 Release*

Maria Osborne †

April 9, 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 from years 2006-2023 and weighting the linked data through 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 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 subscription to 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.

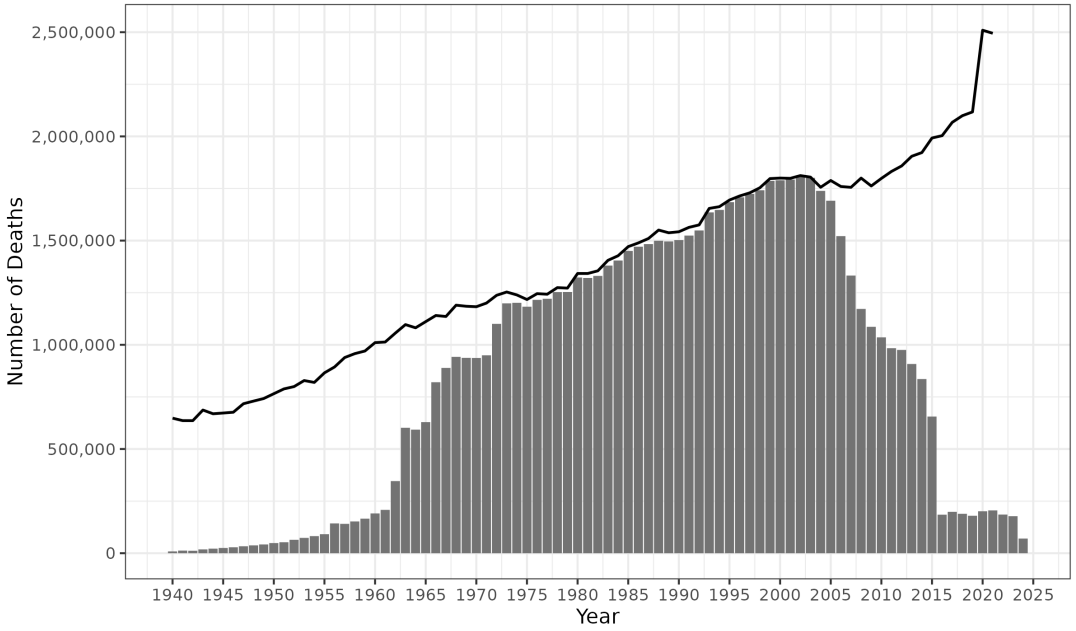

Figure 1: Yearly recorded deaths at aged 65+ in the DMF (bars) compared to the Human Mortality Database (lines). The DMF is nearly complete for years 1975-2005, but far less so outside this period.

The DMF is also incomplete before 1975, possibly because deaths were less commonly reported to the SSA or because they occurred prior to the DMF's creation in 1962. The major gaps after 2005 are due to the exclusion of certain death records from the public version of the DMF, rather than deaths going unreported to the SSA. While the SSA has a complete version of the DMF, only some deaths are included in the publicly accessible version. This raises significant concerns about potential selection in the DMF after 2005, especially as little is known about why certain death records 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 this possible selectivity or exactly which records are released to the public.

To create this version of the CenSoc-DMF, we first link the 1940 Census to the DMF (years 1975-2023) using a conservative variant of the ABE record linkage algorithm. We do not find strong evidence that the decline in DMF completion introduces severe selection on characteristics related to SES 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. We weight CenSoc records up through the year 2020, as this is the last year that we have NCHS mortality data for, though data through 2023 are linked. There are two basic steps in the weighting process:

1. Calculate weights using logistic regression and raking for people born in the contiguous United States and who died 1979-2020. This covers the majority of the CenSoc-DMF.

2. Create alternate weights for cases where weighting directly to NCHS counts of death is inappropriate or impossible. This includes people born outside the US.

The remainder of this report is organized as such: in Section 2, I further describe DMF data, the linked CenSoc-DMF dataset, and NCHS data used for weighting. 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 responsibly for shaping the DMF after 2005. As shown in Figure 2, there is variation in completeness of the DMF at a state level from about 2005-2015. Here, we use the first three letters of decedents' SSN to determine what state their 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 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.

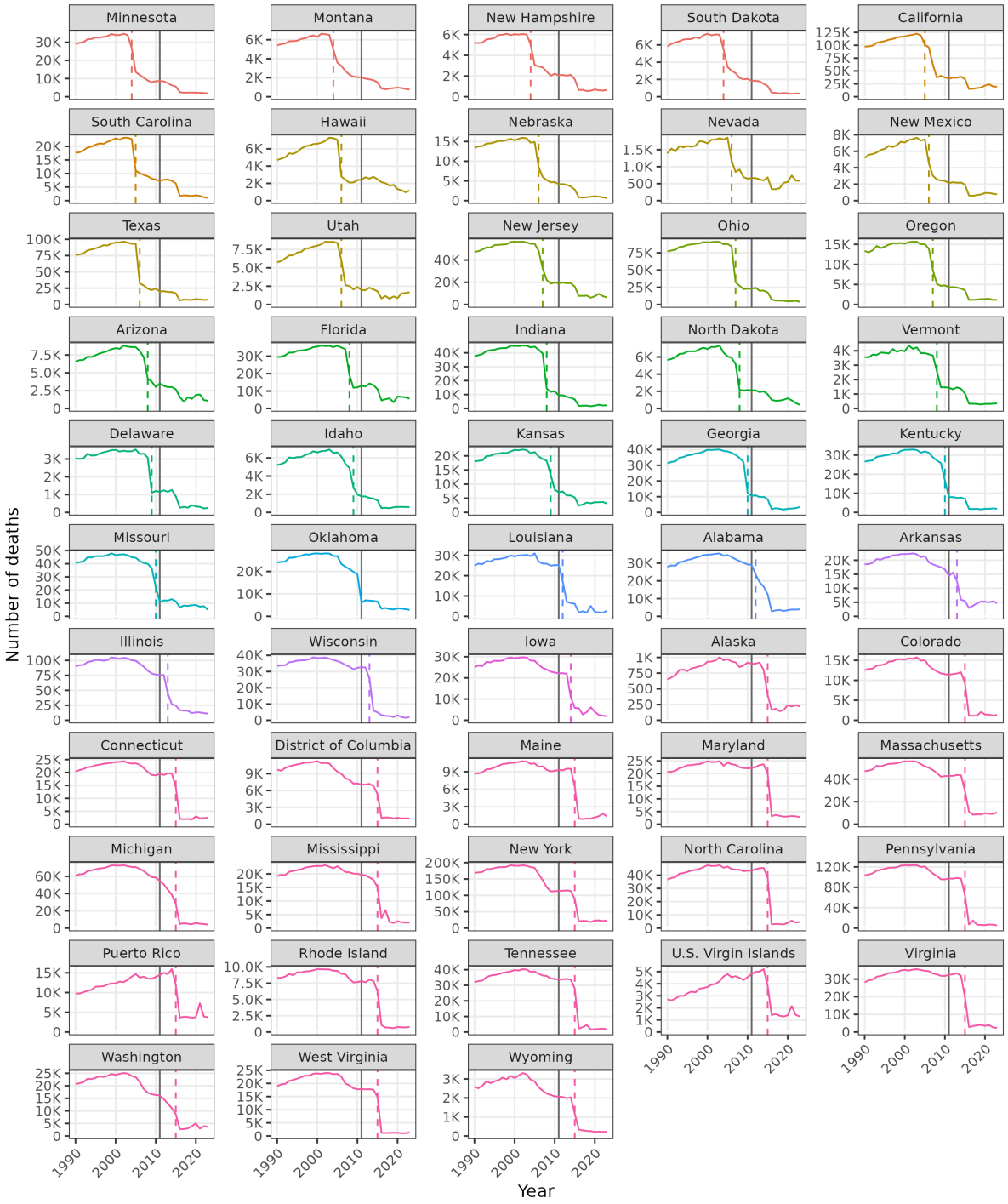

Figure 2: Number of deaths per state, based on state of SSN assignment. Dashed lines indicate the first year where the state experienced a significant year-to-year (>15%) decline in death records. Top left states (such as Minnesota) experience this drop early (around 2004-2005) compared to states in 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.

2.2 Census Data

We link the DMF to full-count 1940 Census data published by IPUMS-USA (Ruggles et al., 2021). To appear in a CenSoc dataset, it is therefore necessary that one be enumerated in the April 1, 1940 Census. Available 1940 Census data consist only of people and dwellings enumerated within the 48 contiguous United States and the District of Columbia, as full-count territorial census microdata from Alaska, Hawaii, Puerto Rico, etc., are not available. People born in US territories are functionally equivalent to immigrants, as they people must have migrated to the contiguous United States before the 1940 Census in order to appear in the linked CenSoc data.

2.3 NCHS Multiple-Cause-of-Death Data

For comprehensive population mortality data, we use Multiple Cause-of-Death (MCOD) data from National Vital Statistics System of the NCHS, a unit of the Centers for Disease Control and Prevention (CDC). MCOD data containing age of death, sex, race, and birthplace information from the years 1979-2004 is public and sourced from the National Bureau of Economic Research's public use data archive (National Center for Health Statistics, 2016). After 2004, because of suppression of birthplace information in public files due to privacy policy changes, we use restricted mortality data files only accessible to approved researchers (National Center for Health Statistics, 2023). Birthplace is not available in any MCOD data from 1968-1978.

MCOD files consist of microdata compiled from death certificates by state vital statistics offices. These data cover nearly all deaths occurring within the United States. This includes deaths to foreign nationals, visa holders not approved to work in the United States, and US citizens/residents who do not hold a SSN. [1]

2.4 Differences between CenSoc and NCHS data

2.4.1 Universes of Deaths

CenSoc data links to mortality records of SSN holders who may die in any location, while CDC mortality data include all deaths within the United States regardless of SSN possession. These two universes of deaths broadly overlap, but are not equivalent. Prior to 1997, CDC death counts exceeded Census Numident death counts, but Census Numident Counts are now consistently higher than CDC death counts (Genadek and Finlay, 2021).[2]

[1] Deaths occurring in US territories are published separately, but are only available from 1994 onward. We do not make use of these data for weighting.

[2] The Census Numident is a version of SSA Numident records processed by the Census Bureau and only accessible to approved users within Federal Statistical Research Data Centers. The file is updated quarterly and has far more complete coverage of deaths than the public NARA Numident records used by CenSoc after 2005. The restricted Census Numident data may have slightly different coverage than

In 1975, CDC death counts exceeded Census Numident deaths counts by 8.5% ; in 2005, the Census Numident deaths exceeded CDC counts by 1.1%.

It is unclear which of these universes (deaths to SSN holders or deaths to all individuals within the United States) in larger in actuality. Further, we do not know which set of deaths is larger when limited to age 65+ mortality. However, CDC data surely omit at least some deaths of relevance to CenSoc (SSN holders who die overseas). Genadek and Finlay (2021) suggest that for years post 1997, several to tens of thousands of deaths to US citizens may be excluded from CDC counts. For earlier years, one estimate of the number of citizens dying overseas comes from Baker et al. (1992), who report that about 5000 Americans died abroad each year from the mid 1970's to mid-1980's, the majority occurring at ages 60+. In comparison, an average of 1.5 million yearly deaths at ages 60+ were reported by the CDC from 1975-1985. Although we don't know how many citizens dying abroad were American-born, the population of deaths represented in CDC data used for weighting is likely slightly smaller than the American-born population eligible for inclusion in CenSoc data sets. Persons who are born in the contiguous United States, enumerated in the 1940 Census, and then die in a US territory such as Puerto Rico also represent a group present in CenSoc data but not CDC data. Available place-of-death data from the CenSoc-Numident mortality dataset suggests that this group of decedents is extremely small.

The inclusion of certain groups in CDC data but not SSA/CenSoc data – undocumented persons, visitors, other non-citizens, and immigrants who arrived in the United States after census day in 1940 – is another problem. This issue is largely addressed by including state of birth as a weighting variable, as it is reasonable to assume that most people present in the United States without an SSN are foreign-born children and adults of working age. The presence of such decedents in MCOD data likely has a negligible effect on weights for the 65+ American-born population in CenSoc. Weighting foreign-born individuals in CenSoc is a much more problematic, as there is no way to discern year of immigration or social security participation of foreign-born decedents in MCOD data. Because of this significant mismatch between types of immigrants present in each data source, we do not directly calculate weights for foreign-born individuals in CenSoc.

2.4.2 Race categories

For CenSoc data, we use race as reported on the 1940 Census to determine racial classification. On the 1940 Census, race categories included: White, Negro (Black), Indian (now called American Indian or Alaska Native), Chinese, Japanese, Filipino, Hindu, and Korean. In MCOD data, race is reported on death certificates. Racial classification schemes vary over time, and generally include a more expansive list of categories than the 1940

the NARA Numident data prior 2005 as well. For more on the NARA Numident and creation of the BUNMD, refer to Breen and Goldstein (2022)

Census.[3] Because of these incongruities, as well as very small counts of deaths for some races, we reduce race to three mutually exclusive categories: Black, White, and Other.

We do not consider Hispanic or Latine ethnicity for the purposes of weighting, as Hispanic ethnicity/origin was not directly collected on the 1940 Census. Additionally, Hispanic status is reported inconsistently in MCOD data due to differing standards of measurement and collection across states and time periods. Not all states reported Hispanic origin to the CDC before 1997, and Hispanic origin data from certain states after 1997 are sometimes not published by the NCHS due to incompleteness.

[3]Beginning in 2003, some states began allowing more than one race to be reported on death certificates. NCHS "bridges" multiple-race responses to single-race categories based on information on races, Hispanic origin, sex, and age of decedents. We use these single-race bridged categories for weighting. More information on the single-race imputation process is available in NCHS documentation.

3 Linkage of the DMF to the 1940 Census

We link the DMF to 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.

The overall census match rate is about 10%. I.e., 10% of men born after 1900 are linked the DMF, with some variation by cohort. Note that this rate is unadjusted for mortality; people observed in the 1940 Census but who did not die between 1975 and 2023 are impossible to link. Due to this, the census linkage rate peaks for birth cohorts circa 1910-1920, whose members are likely to die in the observable window.

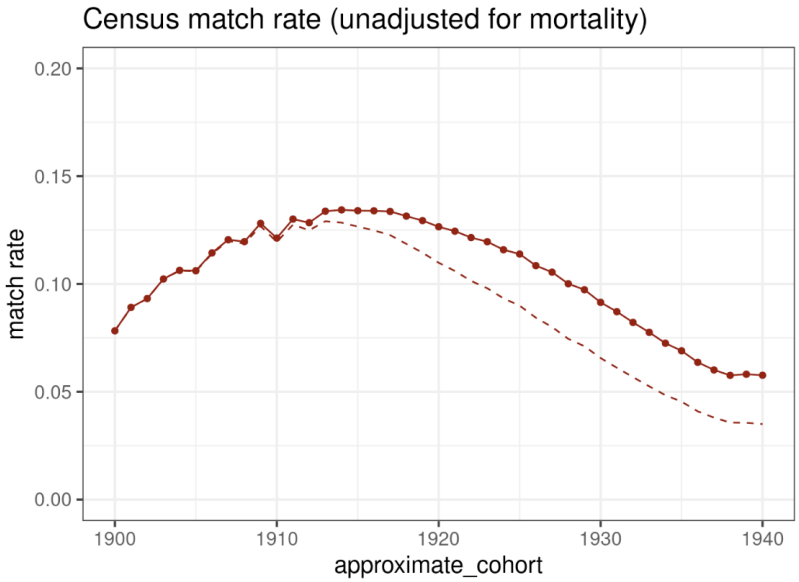

Figure 3: CenSoc-DMF Match Rate (proportion of males in the 1940 Census linked to the DMF) by cohort. The solid line indicates linkages through 2023, in comparison to linkages through 2005 as indicated with the dashed line. Linking records through 2023 primarily increases the census linkage rate for the cohorts of about 2915 and later.

3.1 Assessment of the Linked DMF

Undercoverage in the DMF after 2005 has clear impacts on the age distribution of deaths for later cohorts in the CenSoc-DMF. Below in Figure 4 for example, we show that the unweighted distribution of deaths for an earlier birth cohort (1905) and a more recent cohort (1930). The 1905 cohort, which was largely extinct by 2005, appears Gompertzian, with the modal age of death appearing around age 80. The 1930 cohort, which turned 75 in 2005, is clearly not so. The pattern of deaths after age 75 mirrors the overall decline in death converage in the DMF, rather than a realistic human mortality schedule.

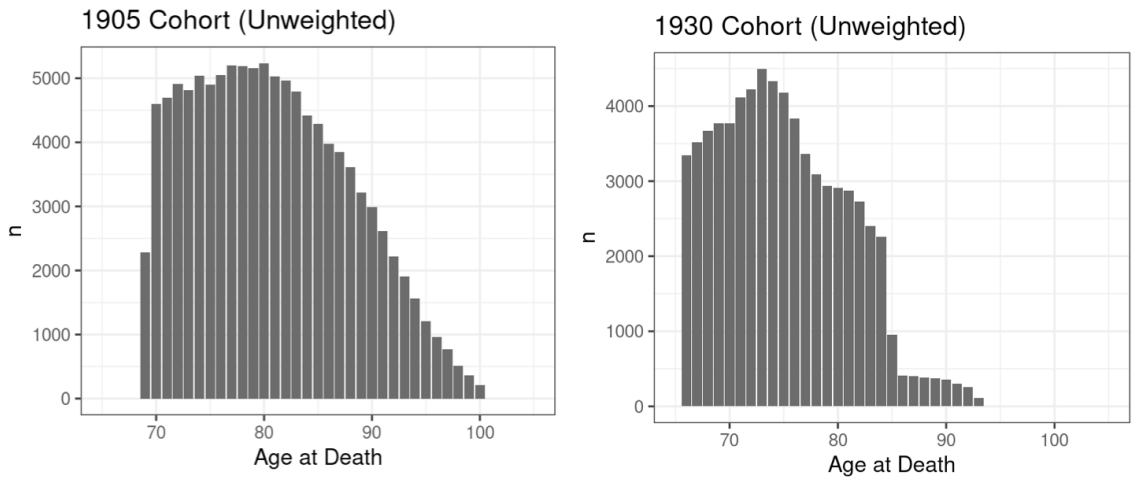

Figure 4: Unweighted cohorts of 1905 (left) and 1930 (right), demonstrating effect of DMF undercoverage on distribution of deaths for more recent cohorts.

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 we use homeownership in 1940 (whether the person lived in a home that was owned by the household head, vs. rented) as a simple measure of SES to see if the composition of deaths in the linked CenSoc-DMF meaningfully changes after 2005. In all cohorts, the trends after 2005 appear to align with pre-2005 trends, though the data can be noisier (especially at very old ages of death). We conclude that while the number of deaths in the DMF after 2005 is certainly much lower that actual deaths in the population, we do not have reason to believe that this undercoverage disproportionately changes the representation of certain groups of people based on 1940 census characteristics.

Figure 5: Proportion of people dying at each age that lived in an owned home (vs. a rented home) for three different birth cohorts. Solid lines indicate deaths prior to 2005, and dashed lines after 2005. For all cohorts, the proportion generally rises as age of death increases, implying that persons who lived in owned homes live longer than those who lived in rented homes in 1940. This trend continues into post-2005 deaths.

4 Weighting Method

4.1 Outline

We weight records decedents aged 65-100 who died in the years 1975-2020. 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**. MCOD 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.

Users should note that as an exception, we do not publish weights for the 1939 and 1940 birth cohorts, even though members of this cohort can die in the specified age and year window. This is because we only observe about 25% of the 1940 cohort (those born by April 1). Consequently, we observe very few deaths in the linked CenSoc-DMF for certain ages in each year. This artificially inflates weights for these ages, affecting weights for both the 1939 and 1940 cohorts.

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 to create weights. First, 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 iterative proportional fitting (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, as shown in Figure 6. Dramatic year-to-year changes in log odds by birth state can occur after about 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 itself may have narrowed over time.

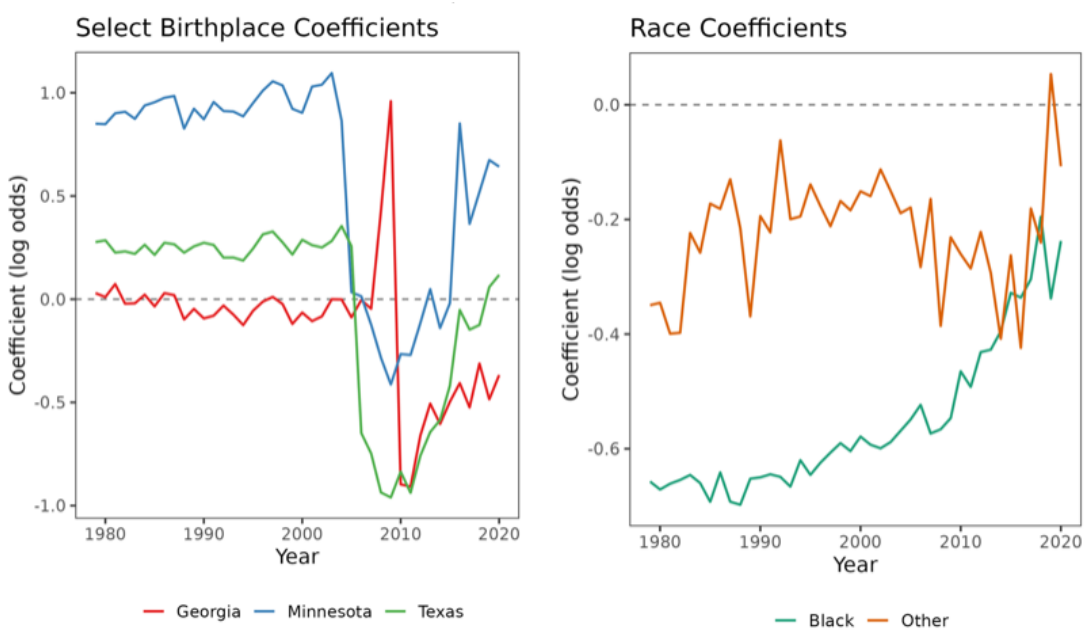

Figure 6: Logit coefficients over time for select states of birth (left panel) and race categories (right panel). A negative coefficient indicates that a group is less likely to be included in the linked CenSoc-DMF than the reference group, while positive cofficient indicates the opposite. For birth states, coefficients are relatively stable before 2005 but can change dramatically year-to-year after this point (the reference state used for this variable is 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. For example, in Figure 7, the raw and smoothed coefficients for select ages at death are shown, with smoothed coefficients extrapolated back to 1975. These predicted coefficients are then used to calcuate propensity scores for deaths in these years. 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.

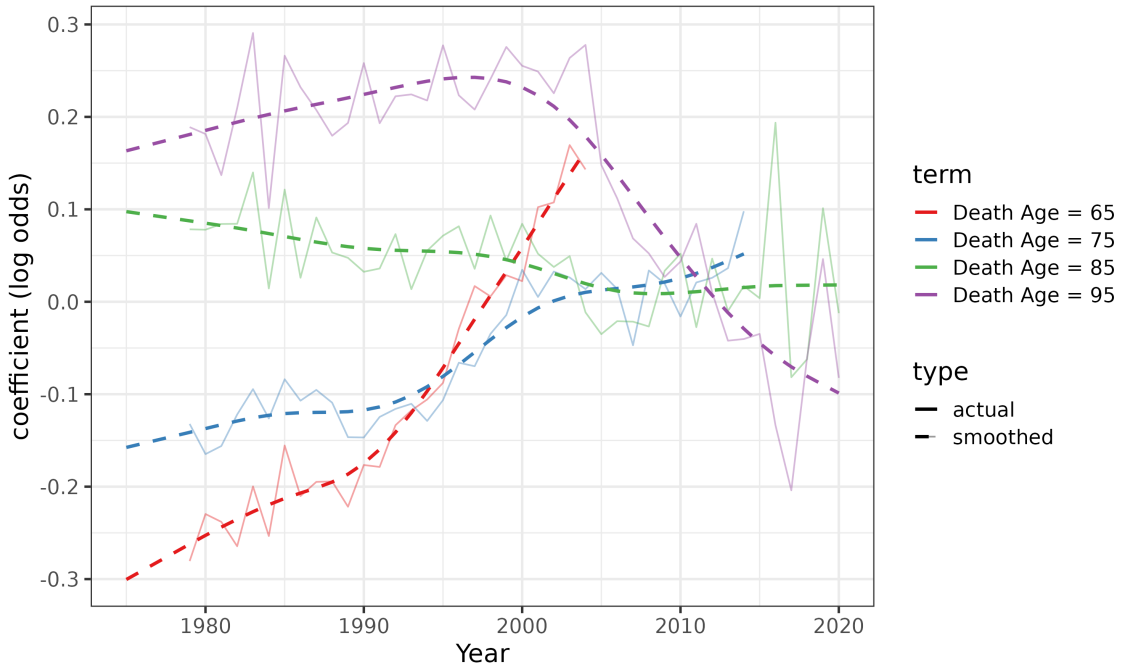

Figure 7: Actual and smoothed age coefficients of each yearly logit model. Smoothed coefficients have been extrapolated back to 1975 in order to construct weights for years 1975-1978. The reference age used for all years is 82.

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). This is because in the CenSoc-DMF, such people must necessarily have moved to the contiguous US before the 1940 Census in order to be observed. NCHS data, conversely, includes people who entered the country after 1940. Weighting this population directly to NCHS would lead to artificially inflated weights, especially for more recent birth cohorts of migrants who are unlikely 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 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 Summary of Weights and Regression Examples

5.1 Weights by age and year

Figure 8 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 subtle but noticeable age patterns in weights, with younger ages at death generally weighted more heavily than older ages at death. During the 2006-2015 period, average weights rise by year as undercoverage in the DMF increases, but age becomes less important. After 2016, weights by single age and year are high and relatively noisy, making any patterns difficult to discern.

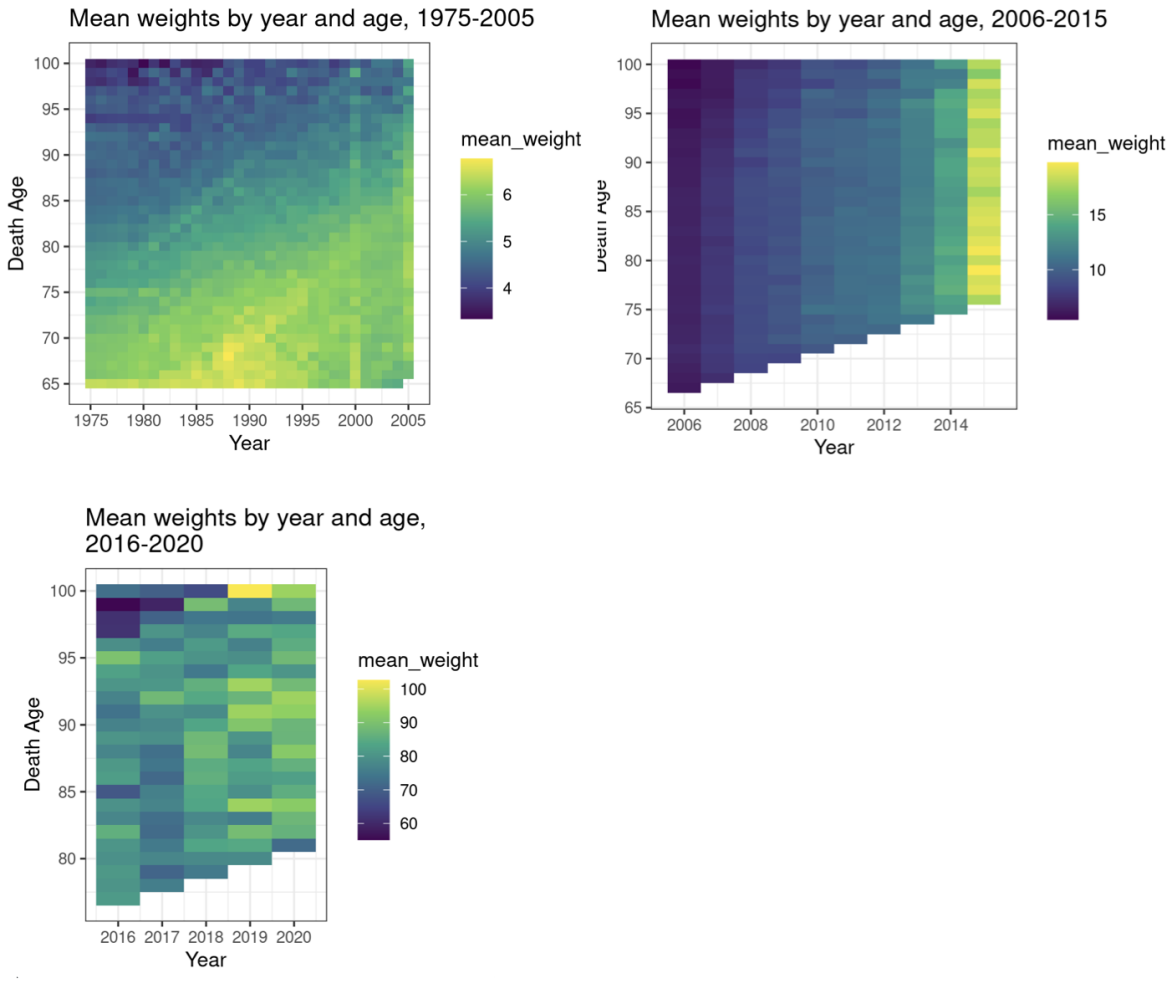

Figure 8: Mean weights by combination of age and year of death. Periods 1975-2005, 2006-2015, and 2016-2020 are plotted separately, as the range of mean weights differs substantially between these three periods. Weights are highest after 2015, the period where DMF coverage is lowest.

5.2 Example Analyses

In this sections, I explore the effects of using weights with a few simple OLS regressions. In Figure 9, 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 Minnesota starts to decline earlier than Alabama, which makes it appear as though people born in Minnesota are usually die at relatively young ages. Using weights corrects for this issue of differing state-level coverage over time in the DMF.

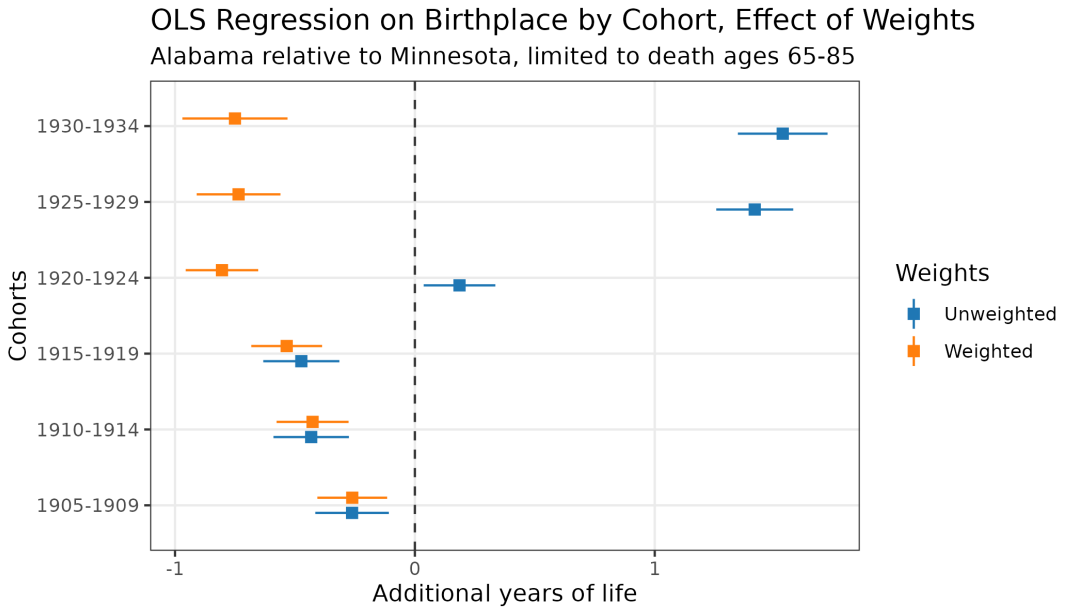

Figure 9: Effect of using weights in OLS regression on state of birth. Without weights, estimates for younger cohorts are unexpected (implying people from Alabama live for more than a year longer on average than those from Minnesota), due to differing coverage of these states in the DMF after 2005. Using weights corrects for this.

The above is an 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. 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.

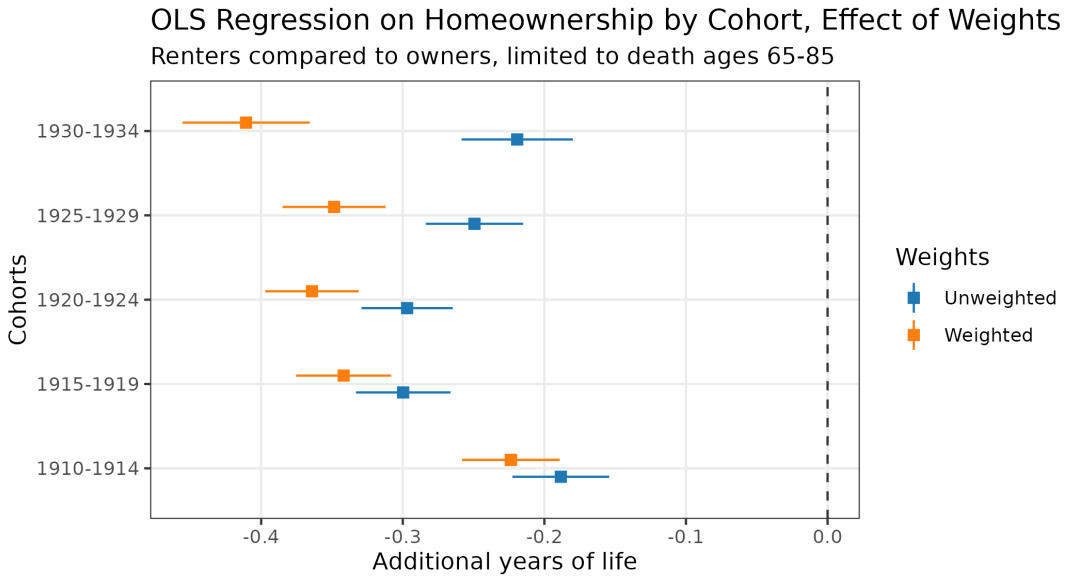

Figure 10: Effect of using weights in OLS regression on homeownership. Weights do affect the magnitude of the estimated effect, especially for later cohorts, but do not change the direction of the estimates.

6 Conclusions

This report details the creation of weights of the linked CenSoc-DMF dataset, years 1975-2023, and weighting of data through the year 2020. 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 much variation in when major declines begin to occur. This can lead to unexpected results if not accounted for, such as life expectancy over 65 for people born in Alabama appearing much longer than that for people born in Minnesota. After 2015, coverage in the DMF appears to be universally low.

3. The probabilities of being included in the linked CenSoc-DMf do change over time with respect to race and age, but more gradually. 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 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 dimenstions of undercoverage and how it changes over time.

6.1 Considerations and Recommendations for Researchers

The CenSoc-DMF links men born prior by April 1, 1940 to the Death Master File, 1975-2023. We create weights for the data for decedents who: 1) died in the years 1975-2020, 2) Died at ages 65-100, and 3) 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 11. Members of these cohorts were observed at ages 10-40 in the 1940 Census, and are likely to die age 65+ in the window of 1975-2020. Cohorts older than about 1920 are extinct or nearly-extinct by 2020, reducing the potential issues caused by working with truncated data (see Goldstein et al. (2023) for a further explanation of how truncated data may bias results).

Researchers should be aware for the following caveats:

1. Researchers should be aware that not all CenSoc data are weighted. The CenSoc-DMF is only weighted up through the year 2020 and birth cohort of 1938. Unweighted years, cohorts, and ages should be used with caution.

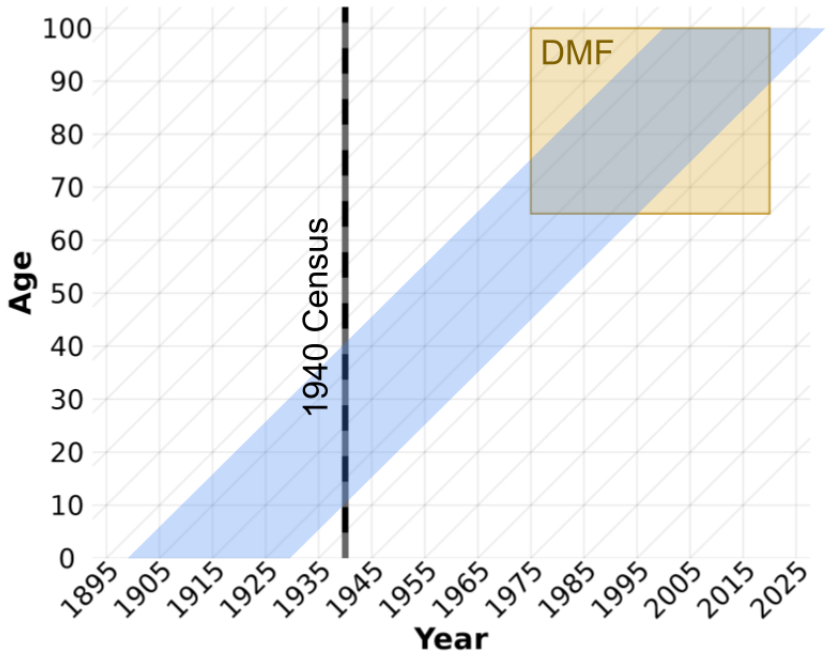

Figure 11: Cohorts of 1900-1930 shown on a Lexis surface with the CenSoc-DMF, which links data from the 1940 Census to DMF mortality data 1975-2023.

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. For the years 1975-2005, the CenSoc-DMF represents around a 10% sample of all deaths at ages 65+, but this falls to about 1% after 2015.

3. In general, we recommend exercising 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. Data for these cohorts are also truncated at lower ages of death, which may cause issues if attempting to fit a curve to the distribution of deaths.

References

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

Baker, T. D., Hargarten, S. W., and Guptill, K. S. (1992). The uncounted dead – American civilians dying overseas. *Public Health Reports*, 107(2):155–159.

Breen, C. F. and Goldstein, J. R. (2022). Berkeley unified numident mortality database: Public administrative records for individual-level mortality research. *Demogaphic Research*, 47(5):111–142.

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.

Genadek, K. R. and Finlay, K. (2021). Measuring all-cause mortality with the census numident file. Working Paper 2021-3, US Census Bureau.

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.

Ruggles, S., Fitch, C. A., Goeken, R., Hacker, J. D., Nelson, M. A., Roberts, E., Schouweiler, M., and Sobek, M. (2021). IPUMS Ancestry full count data version 3.0. dataset.

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.