CenSoc-Numident and CenSoc-DMF Weights Manual

For CenSoc Data Version 3.0, October 2023 Release*

Maria Osborne †

October 21, 2023

Summary

This technical report describes the creation of statistical weights for version 3.0 CenSoc-Numident and CenSoc-DMF mortality datasets using vital statistics data from the National Center for Health Statistics. We describe the structure of CenSoc data, National Center for Health Statistics mortality data, and motivating issues for constructing these weights. We detail the weighting procedure for different groups of records in CenSoc data, and conclude by demonstrating the effect of weights on mortality estimation using a regression framework.

*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 U.S. Census. We distribute three core data sets: 1) The Berkeley Unified Numident Mortality Database (BUNMD), a cleaned and harmonized version of SSA Numident records for deceased individuals published by the National Archives (NARA), 2) the CenSoc-Numident, consisting of men and women in the BUNMD linked to the 1940 Census, and 3) the CenSoc-DMF, consisting of men in the SSA Death Master File (DMF) linked to the 1940 Census. Linked CenSoc datasets best cover age 65+ mortality for the years of 1988-2005 (CenSoc-Numident) or 1975-2005 (CenSoc-DMF).

This report describes the creation of weights for the CenSoc-Numident version 3.0 and the CenSoc-DMF version 3.0. These datasets comprise links between mortality records and 1940 Census records according to a record linkage algorithm. However, only a minority of death records and census records are linkable to each other, and the linking process is not designed to create a sample representative of the American populace. Further, publicly available SSA mortality may itself not represent all deaths to Americans. To combat potential biases arising from these issues, we create person-level weights using vital statistics data from the National Center for Health Statistics (NCHS). Using comprehensive microdata from death certificates published by the NCHS, we create CenSoc person weights using year of death, age at death in years, sex, race, and location of birth. In broad strokes, the process for weighting CenSoc records is as follows:

1. Calculate inverse probability weights using NCHS population counts for the majority of records.

2. Create alternate weights for cases where weighting directly to NCHS counts of death is inappropriate or impossible.

3. Adjust weights to address extreme weights and non-coverage of some population subgroups.

The remainder of this report is organized as such: in Section 2, I further describe CenSoc data and NCHS population data. Section 3 discusses some known problems with the CenSoc data that we seek to address by weighting, such time trends and systemic racial disparities in mortality coverage. In Section 4 I discuss the weighting methodology in detail. In Section 5 I demonstrate use of the weights in a regression framework. Section 6 summarizes the findings of this report and concludes with considerations for researchers.

2 Data

2.1 CenSoc Data

We create weights for the CenSoc-Numident and CenSoc-DMF, which consist of Social Security mortality records linked to the 1940 Census. To appear in a linked CenSoc data set, a person must be enumerated in the 1940 Census and die with a social security number (SSN). This includes SSN holders who die outside the United States, if the death is reported to a U.S. embassy or consulate. CenSoc mortality coverage is best at ages 65 and above, the age at which most individuals born before 1940 became eligible for Social Security retirement benefits. The high coverage period for each linked data set is 1975-2005 for the CenSoc-DMF and 1988-2005 for the CenSoc-Numident.

CenSoc data links to 1940 Census data published by IPUMS-USA (Ruggles et al., 2021). Available 1940 Census data consist only of people and dwellings enumerated within the 48 contiguous United States and the District of Columbia. The 1940 Census predates statehood for Alaska and Hawaii, and while separate territorial censuses were conducted, these have not been harmonized and published by IPUMS. Therefore, people born in Alaska and Hawaii are only present in CenSoc data sets if they moved to the contiguous United States before the enumeration of the 1940 Census. This is also true for individuals born in other U.S. territories. Available place-of-death data, while incomplete, suggests that coverage of deaths occurring in current U.S. territories (Guam, Puerto Rico, American Samoa, the Northern Mariana Islands, and the U.S. Virgin Islands) is relatively high in the BUNMD. However, very few individuals who died in territories appear in the linked CenSoc-Numident, as such people must have been present in the contiguous U.S. in 1940 to be included in linked datasets. Place-of-death coverage in the DMF is presumably similar, but no data are available.

2.2 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 U.S. citizens/residents who do not hold a SSN. [1]

2.3 Differences between CenSoc and NCHS data

2.3.1 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] 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). The Genadek and Finlay (2021) report suggests that for years post 1997, several to tens of thousands of deaths to U.S. 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 U.S. territory such as Puerto Rico also represent a group present in CenSoc data but not CDC data. Available place-of-death data suggests that this group of decedents is very small, however, with fewer than 200 such people in each year of CenSoc-Numident data.

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

[1] Deaths occurring in U.S. 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 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)

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.

2.3.2 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 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 Known Issues with CenSoc data

As shown in Figure 1, CenSoc data sets contain far fewer deaths than occur in the United States each year. CenSoc linking methodology uses only a few fields (name, approximate age, and birthplace if available), and lacks unique identifiers such as social security number or Protected Identification Keys across data sets. As such, match rates are relatively low. About 20% of the men in the SSA DMF 1975-2005 can be linked to the Census using name and birth year (Breen and Osborne, 2022). For the CenSoc-Numident, which additionally uses birthplace as a matching, field, the match rate rises from around 10% of BUNMD records in older cohorts to 30% in younger cohorts. Though these rates are low, we purposefully employ a conservative matching algorithm that discards multiple

[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.

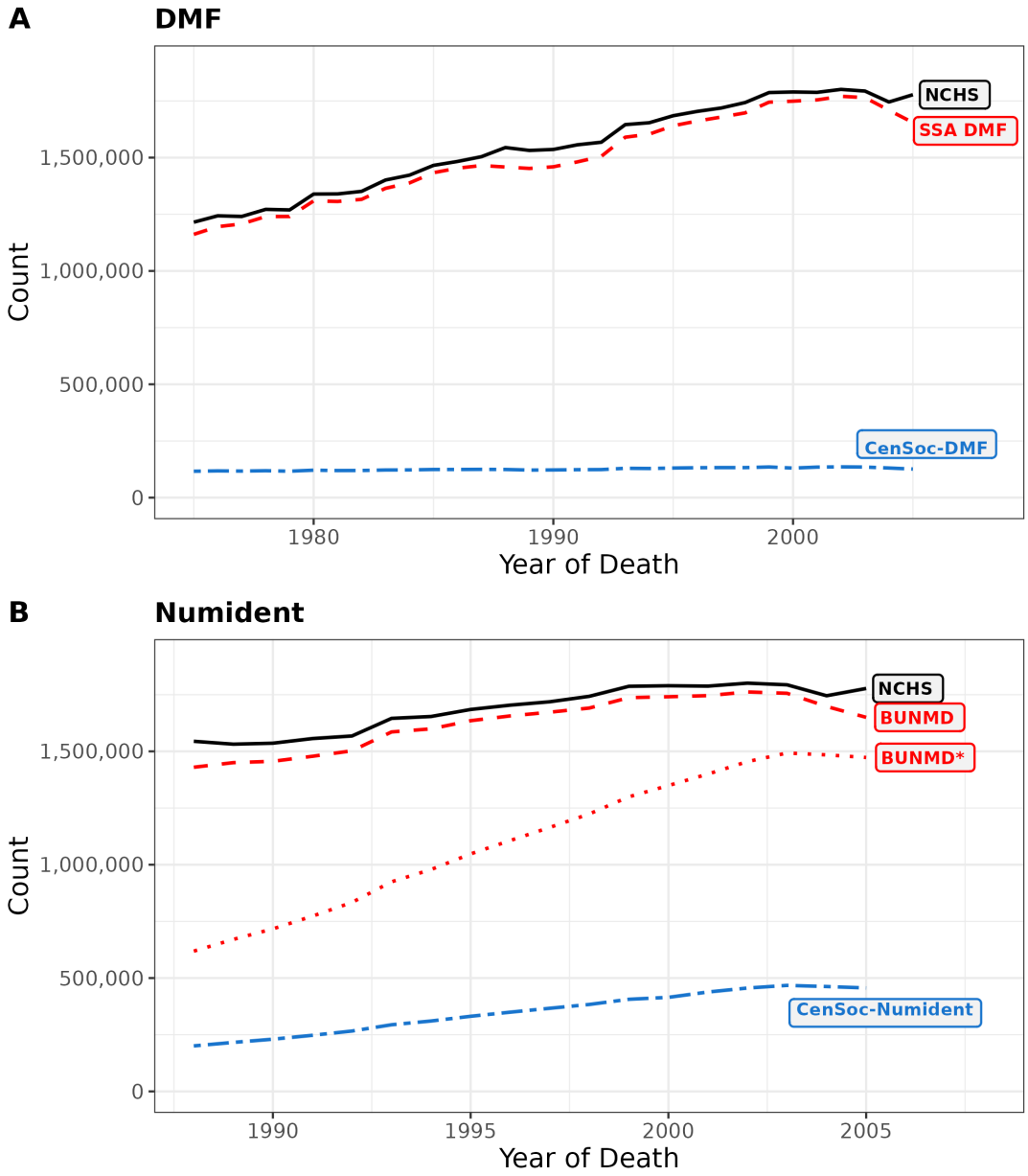

Figure 1: **Death Counts for Ages 65-100 by Year.** Panel (A) shows yearly deaths counts according to the NCHS, the public SSA DMF, and the linked CenSoc-DMF file for years 1975-2005. Panel (B) shows yearly death counts according to the NCHS, the BUNMD, and the CenSoc-Numident (linked version of the BUNMD) for the years 1988-2005. SSA DMF and BUNMD death counts are similar to but slightly lower than NCHS counts over these periods. The dotted line labeled **BUNMD*** in Panel (B) shows the number of records in the BUNMD with a known place of birth, which is used as a matching field. The number of deaths in the CenSoc-Numident increases over time, in part because of increasing birthplace availability. The CenSoc-DMF contains fewer records than the CenSoc-Numident per year because 1) the CenSoc-DMF contains only men, and 2) the lack of birthplace as a matching variable leads to lower match rates.

potential matches within set time bands. This strategy help to ameliorate type I errors (false positives) in the matching process. Such errors are difficult to quantify and can result in misleading inference, for example by inflating rates of geographic and economic mobility (Ruggles et al., 2018).

This method, however, has high potential for type II errors (false negatives). Missed

matches can stem from name spelling inconsistencies, name changes, misreported age or birthplace information, etc. Some potential matches are discarded because multiple individuals with the same name, birthplace, or age cannot be distinguished from each other. Such errors may result in a sample that is not representative of the linkable population. Indeed, Bailey et al. (2020) find that no common record linkage algorithm consistently produces representative samples.

Breen and Osborne (2022) explore these issues as they pertain to CenSoc data in greater detail. Some findings include the fact that Black Americans are underrepresented in the linked CenSoc-Numident, using either the BUNMD or 1940 Census as a benchmark. In the BUNMD, about 10.6% of decedents aged 65-100 who were born in the United States and declared a race on their first social security application classified themselves as Black. Among the same group of decedents in the linked CenSoc-Numident, only 6.3% were classified as Black on their first social security application. While the 1940 Census is an imperfect benchmark due to mortality occurring between 1940 and CenSoc coverage periods, evidence suggests that people of low socioeconomic status and people born in certain regions such as the Southern United States (Alabama, Arkansas, Louisiana, etc.) are also underrepresented in CenSoc data.

While such representation issues can arise from the matching process, they can also be related to gaps in SSA mortality records. Deaths may go unrecorded, for example, due to lack of participation in the social security program, failure to report a death to the SSA, or exclusion from the public version of mortality files. While the DMF has been relatively complete for age 65+ deaths since 1973, there is a small drop in coverage in the late 1980's that may be related to SSA computer database errors or a downturn in death reporting to the SSA by funeral directors (Hill and Rosenwaike, 2001). Huntington et al. (2013) report that among deaths to persons with a social security number in Ohio in 2003, Blacks were less likely than Whites to be included in the DMF. These issues are of greatest concern if they are related to age of the deceased. If the death age distribution of a certain group is not accurately represented, this could lead to misleading inferences when analyzing the data.

3.1 Time Trends in CenSoc Coverage

For the CenSoc-Numident, using birthplace as a matching field leads to a linkage rate that increases over time, as this information becomes more commonly available in social security Numident records. As shown in Figure 1, SSA death counts for people aged 65-100 generally track with NCHS death count. However, while NCHS death counts increase from 2004 to 2005, there is a notable downward tick in deaths in the BUNMD and SSA DMF in the same year, suggesting that the public SSA mortality files may be less complete in 2005. Death counts in the linked CenSoc-Numident and CenSoc-DMF

data also fall slightly in 2005.

This decline in death counts is unequally distributed across groups in the linked CenSoc-Numident, which can be seen in Figure 2. Of all birth states, South Carolina-born individuals experience the single largest proportional drop in deaths from 2004-2005. Other birth states, including Minnesota and New Hampshire, experience substantial declines over both 2003-2004 and 2004-2005, with death counts returning to levels seen in the 1980s. The same pattern appears in the CenSoc DMF.

Although place-of-death information is unavailable in the CenSoc-DMF and incomplete for the CenSoc-Numident, available data suggests that observed declines in deaths to people born in certain states are due to underreporting of death *occurrences* in those states (see Figure A.1). It is common for people in the U.S. to die within the state of their birth (e.g., among Minnesota-born people aged 65+ who died in the year 2000, 56% died in Minnesota, according to MCOD data). This suggests that some state-level SSA mortality records available to the public are highly incomplete for 2005, though the reasons for this are unknown.

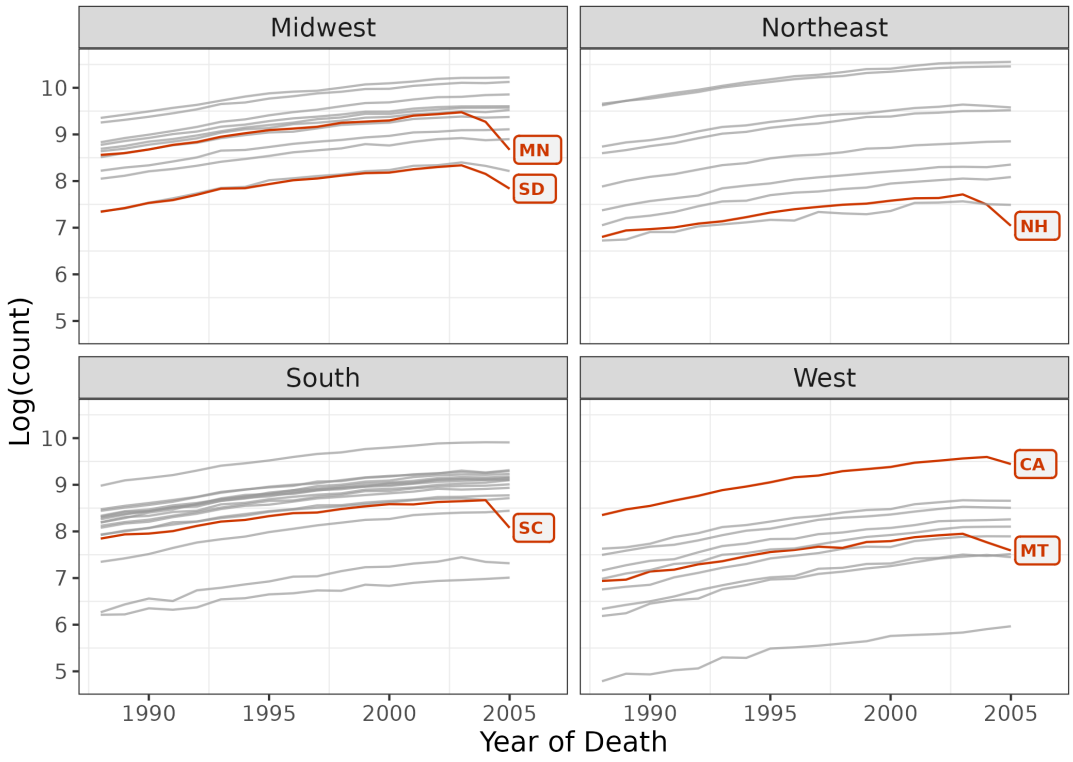

Figure 2: **CenSoc-Numident Counts of Deaths by Birth State.** The plot shows yearly logged counts of deaths occurring by state of birth (including the 48 contiguous U.S. states and the District of Columbia., sorted by IPUMS census region). Highlighted states show substantial and sudden declines in deaths counts in 2004 and/or 2005. These include Minnesota, South Carolina, and New Hampshire.

4 Weighting Method

4.1 Outline

We use poststratification to weight all records of decedents aged 65-100 in the high coverage periods of each datasets: 1979-2005 (CenSoc-DMF) or 1988-2005 (CenSoc-Numident). Where possible, we construct basic inverse probability weights. If this is not possible or appropriate, we assign records a weight using an average from other records. We then adjust weights slightly to address extreme weights and correct for non-coverage. The general process for calculating weights is as follows:

1. **Calculate standard weights:** For records belonging to people who die between the ages of 65-100, in the years 1988-2005 (CenSoc-Numident) or 1979-2005 (CenSoc-DMF), and who were born in the contiguous United States including the District of Columbia, we calculate a basic inverse probability weight.

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 borrow weights from other years.

3. **Weight non-American birthplaces**. People born outside the United States (including Alaska, Hawaii, current U.S. territories, and foreign nations), are only present in CenSoc data if they moved to the contiguous U.S. before census day in 1940. NCHS totals of these groups contain immigrants from all years and non-SSN holders, making them inappropriate to use for calculating inverse probability weights. These records are assigned a mean weight of American-born decedents of the same age, year, sex, and race.

4. **Weight all remaining records** A small number of CenSoc records cannot be directly weighted because of missing birthplace or other data issues. Such records are assigned a mean weight, or in rare cases given a weight of 1.

5. **Adjust weights** Weights are trimmed to a minimum of 1 and a maximum of 5 times the mean unadjusted weight of American-born persons, in order to eliminate small numbers of extremely high weights. For the American-born population dying 1979-2005, weights are adjusted using raking ratio estimation so that weighted marginal totals align with population marginal totals by group (age, year, race, sex, and birthplace). This process of calibrating weights to align with population marginal totals helps compensate for non-coverage of some strata in CenSoc data.

4.2 Standard Inverse Probability Weights

For a large portion of both CenSoc data sets, records are weighted directly up to population totals from NCHS. For each dataset, we divide records into strata cross-classified by year of death (y), age at death (a), sex (s), race (White, Black, or other) (r), and birth state (b). We assign a weight equal to the ratio of deaths in the NCHS data to the number of deaths in the CenSoc data for each given stratum:

$$W_{yasrb} = \frac{\text{number of deaths in NCHS cell } yasrb}{\text{number of deaths in CenSoc cell } yasrb}$$

4.3 Out-of-period weights

4.3.1 Deaths in 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 birthplace of decedents. Other than state-specific coverage issues in 2005, there is no evidence of strong time trends in CenSoc-DMF standard weights. Therefore, for the years 1975-1978, records in particular strata are assigned the same weight are records belonging to the same strata in 1979. For example, White men born in Iowa who die at age 80 in the year 1975 are assigned the same weight as White men born in Iowa who die at age 80 in the year 1979.

4.4 Non-American birthplaces

Standard inverse probability weights cannot be directly computed for people born outside the continuous United States (including foreign nations, Alaska, Hawaii, Puerto Rico, Guam, American Samoa, the U.S. Virgin Islands, and the Northern Mariana Islands). MCOD data does not report when decedents born in these areas entered the United States, so it is unknown which were present for the 1940 Census and thus eligible for inclusion in CenSoc data sets.

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

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

Thus all non-American born records in the same year/sex/race/age stratum are assigned the same weight, regardless of exact country or territory of birth. For example, a Black woman born in Canada or Puerto Rico dying at 75 in the year 1995 receives the average weight of all American-born Black women who die at age 75 in 1995.

4.5 All Other Records

Some records cannot be weighted using the above calculations. Such records include:

1. **People with missing birthplaces.** A very small number of CenSoc records do not have useful birthplace information recorded in the 1940 Census.

2. **People belonging to missing strata.** Some CenSoc records belong to strata that do not exist in the NCHS population, which may occur for a number of reasons. The CenSoc record could have been falsely matched, creating a combination of age/year/sex/race/birthplace that truly does not exist in the population. Incongruities in racial classification between the census and death certificates could also cause this phenomenon. Additionally, alternative weights for people dying in 1975-1978 and the foreign-born are sometimes not computable using the procedures above. For example, a stratum that exists in 1975 CenSoc data but not 1979 cannot be weighted using the procedure detailed in Section 4.3.

We first treat such problematic records like the non-American born and attempt to assign them the average weight of records with the same year/sex/race/age among the native-born. At the end of this process, a very small number of unweighted records remain. Such records almost exclusively consist of non-Black and non-White people, and in the CenSoc-DMF mainly come from the years 1975-1978. These remaining records are assigned a weight of 1. There are 157 such records in the CenSoc-Numident and 579 such records in the CenSoc-DMF.

4.6 Adjusting Weights to Address Bias and Reduce Variance

Weighted CenSoc death totals do not equal NCHS death tallies. This is to be expected, as the foreign-born are not weighted using real population numbers. More concerning, weighted counts of American-born deaths are slightly deficient compared to NCHS counts due to age/year/sex/race/birthplace strata that exist in the population but are not captured by CenSoc data. For both the CenSoc-Numident 1988-2005 and CenSoc-DMF 1979-2005:

$$\frac{\sum (\text{Number of deaths in CenSoc cell } yasrb) \times W_{yasrb}}{\sum \text{Number of deaths in NCHS cell } yasrb} = 0.995$$

Thus, approximately 0.5% of NCHS deaths to American-born people are absent from CenSoc data when raw weights are applied. Such non-coverage errors may introduce bias. The characteristics of non-captured strata are in some cases very different from strata that are captured by CenSoc. Among all American-born decedents in NCHS data, about 9.8% of people belonging to strata represented in the CenSoc-Numident are non-White. Of strata not observed in the CenSoc-Numident, however, 73.2% of people are

non-White. People in unobserved strata also skew slightly older and more male than the general population and observable strata.

Another potential issue with the raw weights is the presence of very high weights. For instance, while the median weight of records in the CenSoc-Numident is 3.4, the maximal weight in 193.0. Such extreme weights are generally seen as undesirable, as they increase variation in weights and can lead to inflated sampling variances (Potter, 1988).

To address these issues, we trim weights to a set range, and employ raking ratio estimation (Deville and Särndal, 1989) to calibrate weighted CenSoc marginal totals by year, age, sex, race and birthplace with population marginal totals of the same control variables. When trimming weights, the minimal value is set to 1 and the maximal value is set to U, 5 times the mean unadjusted weight among the American-born for each data set. Capping weights at 5x the mean value is a simple but common strategy for trimming weights (Izrael et al., 2009). Less than 1% of raw weights are greater than U in both data sets. Raking is implemented with the R package *Sampling* (Tillé and Matei, 2021). For the American-born population dying 1979-2005, where population totals are known, records are iteratively trimmed and raked until weights are both calibrated to marginal population values and the maximum weight is within 0.0001 of U. As population marginal totals by all variables do not exist for records from 1975-1978, and for the foreign-born, such records are only adjusted by trimming.

5 Summary of Weights and Regression Examples

5.1 Weights by age, year, cohort, race, and birthplace

Figure 3 Shows mean weights for each CenSoc data set on a lexis surface. Age/year patterns of weights are not dramatic in the CenSoc-DMF, though weights are generally higher for younger ages at death. Conversely, there are strong age, year, and cohort patterns in CenSoc-Numident weights, with the highest mean weights concentrating in the late 1980s around age 80. Older ages at death receive higher weights overall, but younger ages at death are weighted more heavily within individual cohorts. This is especially apparent for birth cohorts earlier than circa 1914.

Figure 4 Shows the distributions of weights for by race of decedent in each CenSoc data sets. Weights for Black decedents are on average much higher and more variable than weights assigned to White decedents. The median weight for other races is lower than Whites in the CenSoc-DMF, but higher than Whites in the CenSoc-DMF. Additionally, the CenSoc-DMF has generally higher weights than the CenSoc-DMF. This is likely due to the overall lower match rate of the CenSoc-DMF. Birthplace is not used as a matching variable in the CenSoc-DMF, which results in more multiple potential matches that cannot be distinguished from each other must be discarded.

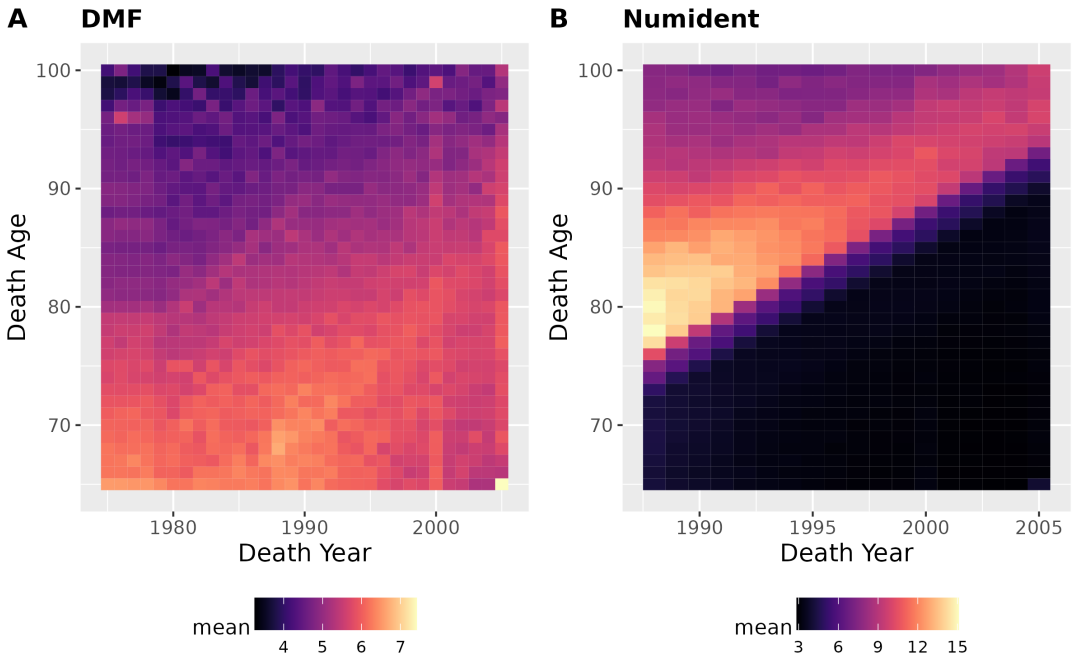

Figure 3: **Mean Weights of CenSoc Records by Age and Period.** Weights in the CenSoc-DMF (Panel A) are slightly higher for younger ages than older ages at death, both overall and within cohorts. Weights in the CenSoc-Numident (Panel B) are generally lower for younger ages than older ages. Within cohorts, however, younger ages are weighted more heavily. Weights are dramatically lower on average for cohorts born after about 1913, where individuals are much more likely to have birthplace information available in social security records.

Patterns of weight by state of birth over time reveal broad systemic differences in coverage by region of birth. CenSoc-DMF coverage (Figure 5) is generally best for people born in the Midwest (e.g., Ohio, Illinois), as indicated by relatively low average weights, and poorest for those born in the South, particularly the Deep South (e.g., Alabama, South Carolina). Weights spike for certain states in 2005. The same patterns of weights by region and in the year 2005 also present themselves in the CenSoc-Numident, shown in Figure 6. Weights also generally decrease over time corresponding to increased availability of the birthplace matching variable.

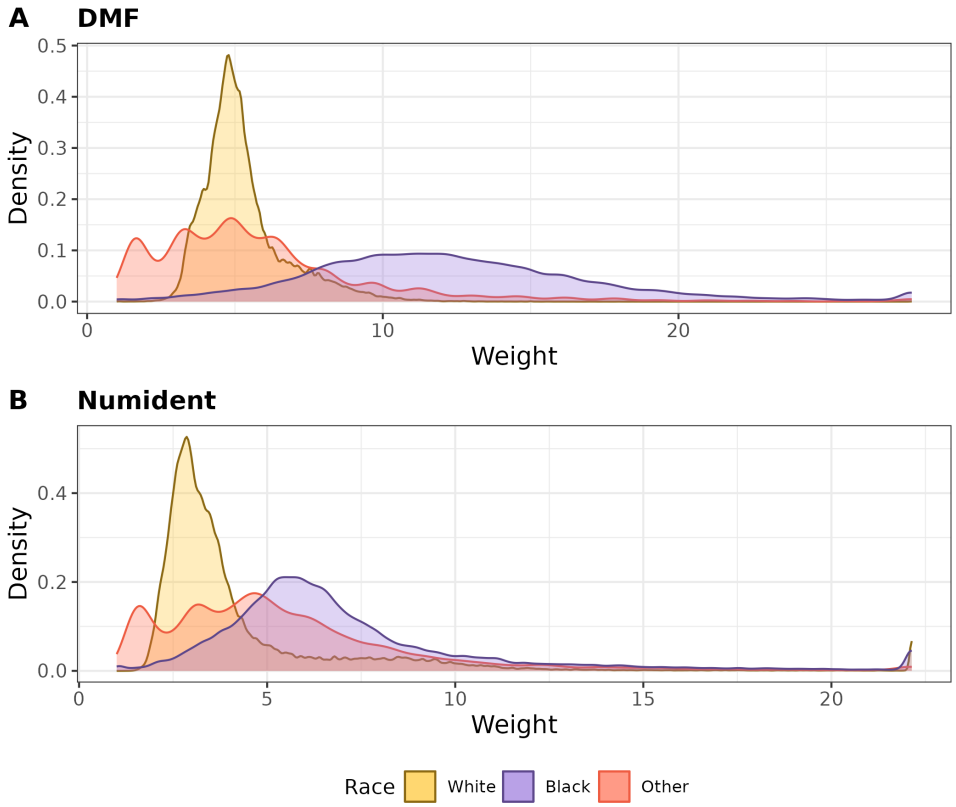

Figure 4: **Distributions of Weights by Race.** The plots show kernel density estimate plots of distribution of weights for American-born individuals. The CenSoc-DMF (Panel A) has generally higher weights than the CenSoc-DMF (Panel B). In both datasets, weights for Black decedents are on average higher and more variable than those for White decedents.

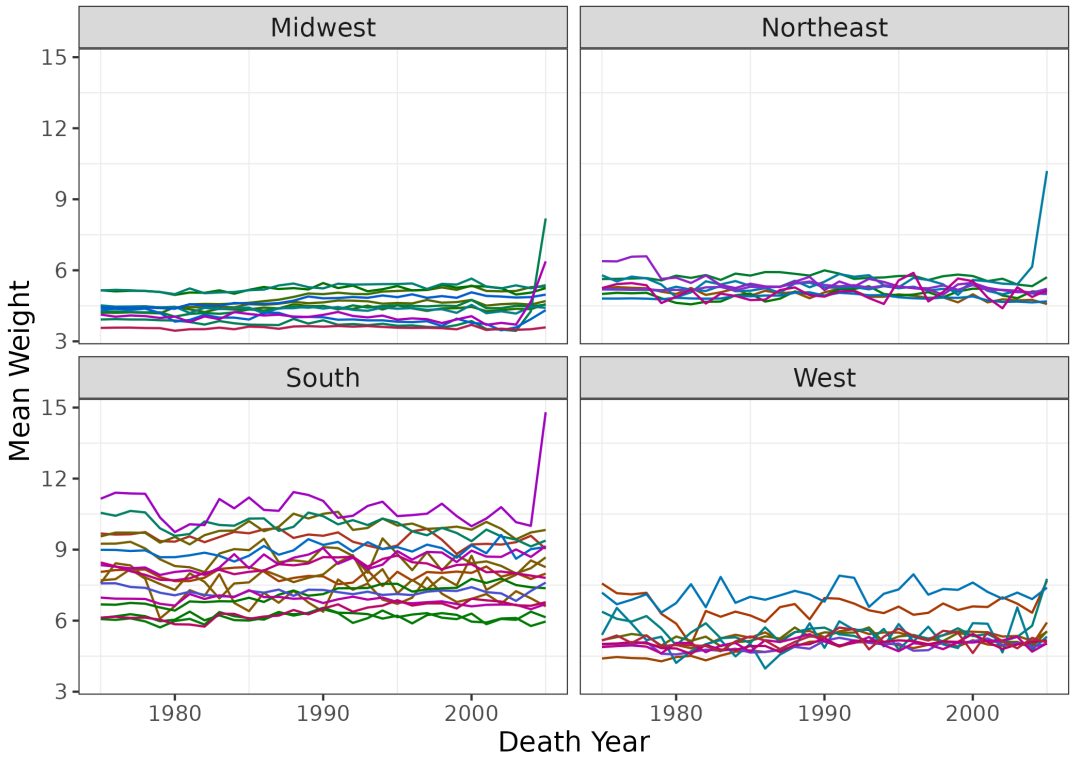

Figure 5: **CenSoc-DMF Weights by State of Birth.** The plot shows mean weights by birth state and year of death (includes the 48 contiguous states and the District of Columbia, organized by census region). Weights are highest overall for southern states of birth. Mean weights are relatively stable over time, except for some spikes in 2004 and 2005.

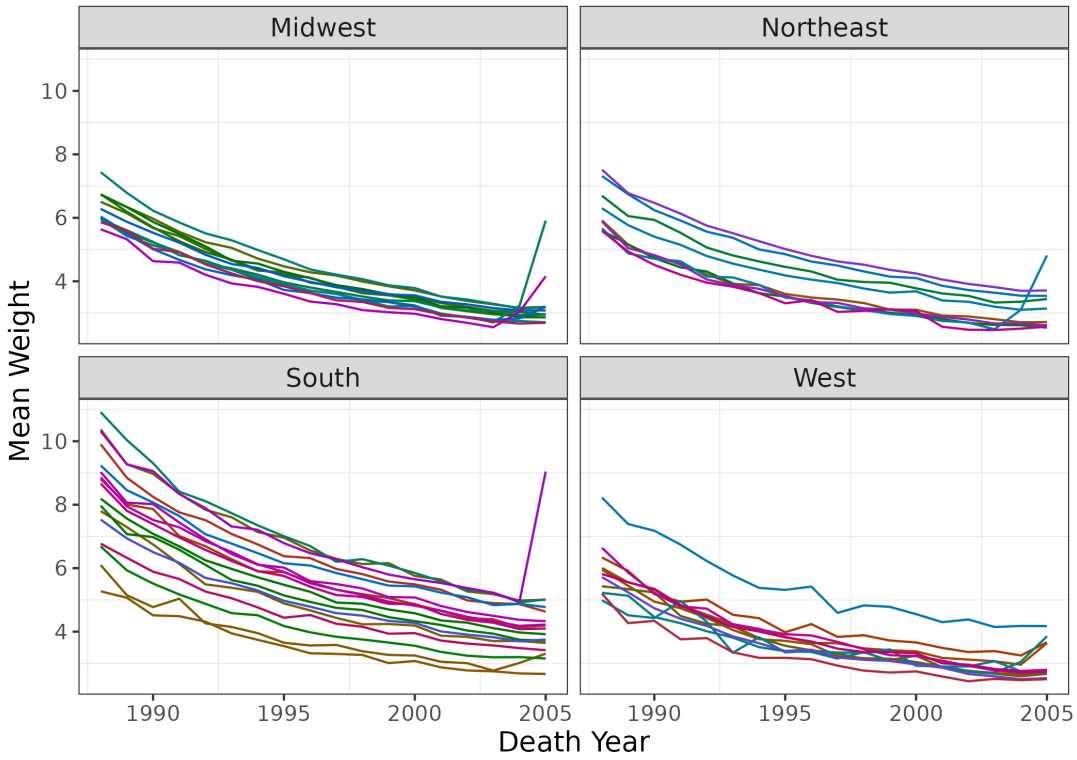

Figure 6: **CenSoc-Numident Weights by State of Birth.** The plot shows mean weights by birth state and year of death (includes the 48 contiguous states and the District of Columbia; organized by census region). Weights are highest overall for southern states of birth. Mean weights generally decline over time, except for some spikes in 2004 and 2005.

5.2 Example Analyses

In this section, I demonstrate the effects of using weighted data on inference with some simple mortality models. A straightforward method of estimating mortality with CenSoc data is to use a linear regression on age at death with cohort fixed effects to account for the effects of double truncation (Breen and Goldstein, 2022). While I will make use of this regression framework, it should be noted the magnitudes of effect sizes may be attenuated due to truncation (Goldstein et al., 2023).

Table 1

	Dependent variable:			
	Age at death in years			
	(1)	(2)	(3)	(4)
	Numident 1988-2005 (unweighted)	Numident 1988-2005 (weighted)	DMF 1975-2005 (unweighted)	DMF 1975-2005 (weighted)
Race = Black	-0.655^{***}	-0.880^{***}	-0.888^{***}	-1.468^{***}
	(0.021)	(0.016)	(0.032)	(0.021)
Race = East Asian	0.639^{***}	0.912^{***}	1.536^{***}	1.953^{***}
	(0.112)	(0.097)	(0.143)	(0.141)
Race = Native American	-0.508^{***}	-0.237^{**}	-1.052^{***}	-0.493^{***}
	(0.107)	(0.093)	(0.147)	(0.137)
Constant	85.501^{***}	84.802^{***}	78.523^{***}	78.123^{***}
	(0.036)	(0.018)	(0.020)	(0.020)
Observations	1,255,438	1,255,438	1,266,539	1,266,539
R^2	0.170	0.201	0.018	0.018
Adjusted R^2	0.170	0.201	0.018	0.018

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: **Relationship between Race and Mortality.** These models compare the effect of race on longevity after age 65 (reference category = White) between weighted and unweighted CenSoc-Numident and CenSoc-DMF data. White, Black, Native American, and East Asian (Chinese and Japanese) Men born in the contiguous United States 1910-1920 are included. All models include cohort fixed effects. Use of weights implies more severe mortality disadvantage for Blacks relative to Whites, particularly with CenSoc-DMF data (-0.89 years compared to -1.47 years). Point estimates for other races moved in the positive direction.

As a first example, Table 2 shows the estimated effect of race on longevity of American-born men using weighted and unweighted CenSoc data. In this instance, using the weighted data implies a larger longevity disadvantage for Black men relative to White

men. The increase in magnitude of this estimate is especially apparent in CenSoc-DMF data. For Asians and Native Americans, both groups classified as "other" races for weighting purposes, weights pulls point estimates in the positive direction. Table A.4 in the appendix includes the same analysis with HMD weights (used for previous releases of CenSoc), showing that the weights discussed in the paper have a much more dramatic impact on these estimates than previous versions of CenSoc weights that did not account for systemic racial disparities in mortality coverage. Table A.2 in the appendix contains the same regression, but also includes men born in Alaska and Hawaii. While it is more appropriate to treat these areas as territories than states, researchers may not make this distinction. Indeed, inclusion of these very small groups yields highly comparable results and is largely inconsequential.

Table 4 shows this regression at the state-of-birth level. For women in both birth states presented, weighting increases the negative effect of being Black on longevity. For women born in New York, the effect of being Black is only statistically significant is weights are used. The impact of weights on non-Black racial groups is less consistent.

In some cases, use of weights may be consequential to inference. In Table 5, the unweighted regression of state of birth on longevity has counterintuitive results. It shows that women born in Minnesota have *decreased* longevity compared to women born in Alabama, despite Alabama consistently ranking far lower than Minnesota in terms of life expectancy (Montez et al., 2020). While most measures of state-level morality are based on place of residence rather than birth, many individuals remain in the same area throughout their life, and evidence suggests that state of birth is associated with risk of dementia, Alzheimer's disease, stroke, and cardiovascular disease related mortality regardless of adult residence (Xu et al., 2021; Glymour et al., 2011). The weighted regression yields results more aligned with expected geographic patterns in mortality.

These examples estimate the effect of variables that are used to create weights, but the impact of weights may be more subtle when estimating the effects of non-weighting variables. Table 7 shows the effect of educational attainment on longevity using unweighted and weighted CenSoc-Numident data. Relative to the reference category of middle school completion, both unweighted and weighted regressions show that more educated people live longer, as expected. The use of weights has a noticeable but small effect on the point estimates: the lowest educational attainment categories (no education and less than middle school) face a slightly more severe mortality penalty, and the effects of highest attainment (some college and college completion) are slightly attenuated. However, the sign and statistical significance of estimates is unaffected.

Finally, for an example that uses non-American born records, refer to Table A.6, which compares longevity of immigrants by country of origin relative to American-born whites in the CenSoc-DMF. In this case, weights have relatively little substantive impact on estimates.

Table 3

	Dependent variable:	
	Age at death in years	
	(1)	(2)
	Unweighted Numident	Weighted Numident
Panel A: Alabama		
Race = Black	-0.451^{***}	-0.529^{***}
	(0.084)	(0.076)
Race = East Asian	-0.598	-0.135
	(2.171)	(2.663)
Race = Native American	2.814	3.366
	(3.433)	(3.889)
Constant	89.506^{***}	89.155^{***}
	(0.148)	(0.133)
Observations	19,315	19,315
R^2	0.180	0.176
Adjusted R^2	0.179	0.175
Panel B: New York		
Race = Black	-0.094	-0.462^{**}
	(0.201)	(0.192)
Race = East Asian	3.313^{***}	3.307^{*}
	(1.224)	(1.876)
Race = Native American	0.263	-0.482
	(1.358)	(1.884)
Constant	89.653^{***}	89.260^{***}
	(0.087)	(0.069)
Observations	75,338	75,338
R^2	0.189	0.176
Adjusted R^2	0.189	0.176
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4: **Relationship between Race and Longevity Within Birth States.** These model compare the effect of race on longevity after age 65 (reference category = White) between weighted and unweighted CenSoc-Numident data among selected states of birth. Women from birth cohorts 1905-1915 are included. All models include cohort fixed effects. The magnitude of point estimates for Black women increased in both states when weights were used. For women born in New York, the coefficient is only statistically significant with weighed data. The impact of weights on estimates for other races is less consistent.

	Dependent variable:	
	Age at Death in years	
	(1)	(2)
	Unweighted	Weighted
BPL = Massachusetts	−0.035	0.200***
	(0.045)	(0.042)
BPL = Michigan	0.001	0.224***
	(0.044)	(0.041)
BPL = Minnesota	−0.131***	0.431***
	(0.048)	(0.046)
BPL = New Hampshire	−0.360***	0.178*
	(0.085)	(0.091)
Constant	82.188***	81.771***
	(0.048)	(0.044)
Observations	117,326	117,326
R^2	0.092	0.087
Adjusted R^2	0.092	0.087
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 5: **Relationship Between State of Birth and Mortality.** This model compares the effect of race on longevity after age 65 (reference category = Alabama) between weighted and unweighted CenSoc-Numident data. Women born from 1915-1920 are included. All models include cohort fixed effects. The unweighted model has counterintuitive results, showing that the longevity of women born in high life-expectancy states in the Northeast and Midwest is similar to or even worse than women born in Alabama. The weighted regression instead shows that women born in the selected Northeast/Midwest states live longer than women born in Alabama. This example shows that some models may yield misleading inferences without weights.

Table 6

	Dependent variable:	
	Age at death in years	
	(1)	(2)
	Unweighted	Weighted
Educ = None	-0.617^{***}	-0.794^{***}
	(0.084)	(0.070)
Educ = Less than Middle School	-0.296^{***}	-0.356^{***}
	(0.024)	(0.021)
Educ = Some High School	0.037^{*}	0.067^{***}
	(0.021)	(0.020)
Educ = High School	0.404^{***}	0.411^{***}
	(0.020)	(0.020)
Educ = Some College	0.572^{***}	0.522^{***}
	(0.028)	(0.027)
Educ = College or Higher	0.897^{***}	0.892^{***}
	(0.026)	(0.026)
Constant	88.750^{***}	88.224^{***}
	(0.053)	(0.034)
Observations	528,969	528,969
R^2	0.157	0.208
Adjusted R^2	0.157	0.208
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 7: **Relationship Between Educational Attainment and Longevity.** These model compare the effect of education on longevity after age 65 (reference category = Middle School Completion) between weighted and unweighted CenSoc-Numident. American-born men from the cohorts 1905-1915 are included. All models included cohort fixed effects. The effect of weights on coefficients is relatively small. The coefficients for high levels of education are slightly attenuated, and those for the lowest educational attainment are larger in magnitude.

6 Conclusions

This report describes the creation of weights for linked CenSoc datasets using NCHS population figures by age, year, sex, race, and birthplace. The weighting process involves calculating inverse probability weights by population strata, creating alternative weights where necessary, and performing some minor adjustments of weights to increase accuracy and usability. We report the following:

1. There are overall strong time trends in mortality coverage in the CenSoc-Numident. While such trends are generally absent from the CenSoc-DMF, coverage of certain states is notably poor in 2004/2005 in both datasets.

2. Age and cohort patterns differ between datasets. In the CenSoc-DMF, older ages are better captured than younger ages, and vice versa for the CenSoc-Numident. Within cohorts, younger ages generally receive higher weights than older ages at death, particular for pre-1914 cohorts in the CenSoc-Numident.

3. Coverage of persons born in the South, particularly the Deep South, is worse compared to other regions. Coverage of non-white decedents is worse than that of white decedents.

4. Use of weights is unlikely to drastically change statistical inferences in most cases. In the previous examples comparing unweighted and weighted estimates using OLS, weights generally affected the magnitude of regression coefficients, but not their sign or statistical significance. However, it is possible to design models in which failing to consider weights may lead to misleading conclusions, such as the previous example showing unexpected mortality differentials by state of birth. The effect of weights on estimates was most apparent when estimating the effect of a weighting variable itself (race, state of birth), and relatively muted when estimating the effect of a non-weighting variable (education).

6.1 Caveats and Considerations for Researchers

Users of the CenSoc-Numident and CenSoc-DMF datasets should be aware of the following issues when utilizing weights. First, weights for persons born outside the contiguous United States should be used with caution, as true population totals are unknown. Technically, this applies to individuals born in Alaska and Hawaii, though inclusion of such individuals with the American-born is unlikely to have consequential effects on analyses. Similarly, weights for the years 1975-1978 cannot be constructed from actual NCHS data and are less informed than weights from other years.

Additionally, we have not accounted for all forms of bias in the calculation of these weights. A portion of NCHS mortality records lack birthplace information (outside of

years 1975-1978 when birthplace is universally unavailable). About 0.56% of decedents aged 65-100 dying 1979-2005 are missing birthplace information. These records are excluded from the population used to construct weights, which may lead to slight undercounts of some population groups. Missing birthplaces are especially prevalent in NCHS data years 1989 and 1990.

References

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

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.

Breen, C. F. and Osborne, M. (2022). An assessment of censoc match quality. Working paper.

Deville, J.-C. and Särndal, C.-E. (1989). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

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

Glymour, M. M., Kosheleva, A., Wadley, V. G., Weiss, C., and Manly, J. J. (2011). The geographic distribution of dementia mortality: Elevated mortality rates for black and white americans by place of birth. *Alzheimer Disease & Associated Disorders*, 25(3):196–202.

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).

Hill, M. R. and Rosenwaike, I. (2001). The social security administration's death master file: the completeness of death reporting at older ages. *Social Security Bullatin*, 64(1).

Huntington, J. T., Butterfield, M., Fisher, J., Torrent, D., and Bloomston, M. (2013). The social security death index (ssdi) most accurately reflects true survival for older oncology patients. *Journal of Cancer Research*, 3(5):518–522.

Izrael, D., Battaglia, M. P., and Frankel, M. R. (2009). Extreme survey weight adjustment as a component of sample balancing (a.k.a.raking). Working Paper 247-2009, SAS Global Forum.

Montez, J. K., Beckfield, J., Cooney, J. K., Grumbach, J. M., Hayward, M. D., Koytak, H. Z., Woolf, S. H., and Zajacova, A. (2020). US state policies, politics, and life expectancy. *The Milbank Quarterly*, 98(3):668–699.

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.

Potter, F. (1988). Survey of procedures to control extreme sampling weights. *American Statistical Association 1988 Proceedings: Survey Research Methods Section*.

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.

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

Tillé, Y. and Matei, A. (2021). *sampling: Survey Sampling*. R package version 2.9.

Xu, W., Topping, M., and Fletcher, J. (2021). State of birth and cardiovascular disease mortality: Multilevel analyses of the national longitudinal mortality study. *SSM-Population Health*, 15.

A Additional Figures and Tables

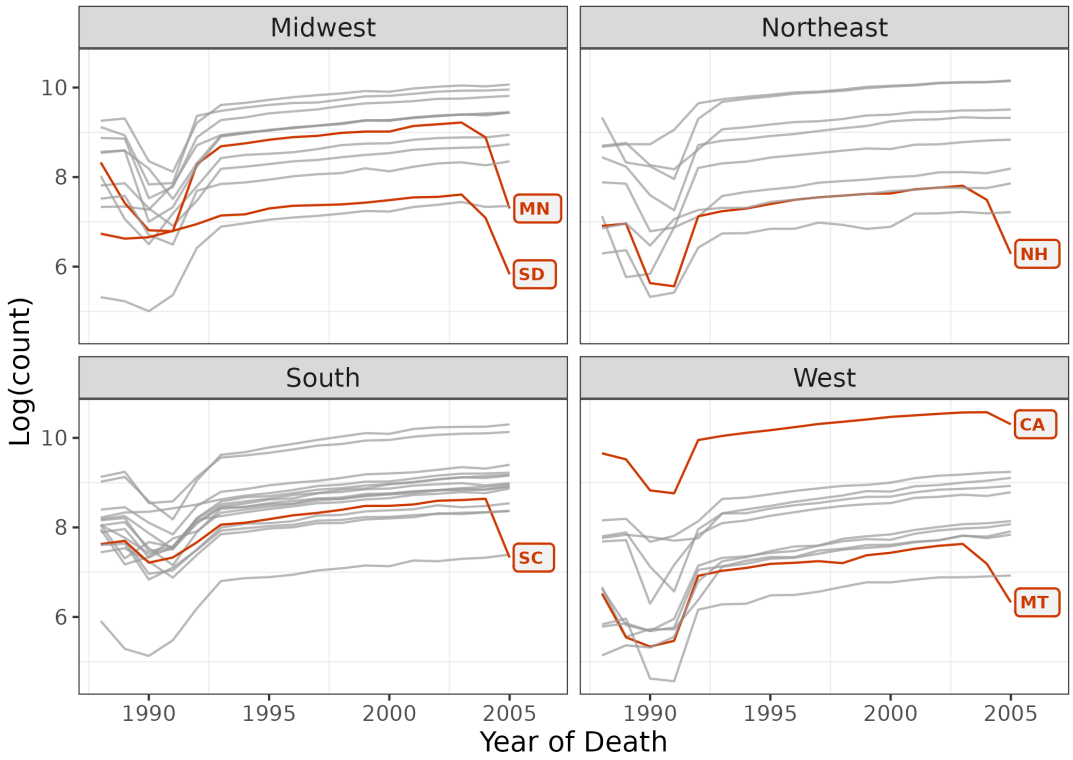

Figure A.1: **CenSoc-Numident Counts of Deaths by Death State.** These plots show yearly logged counts of deaths occurring in each state (includes 48 contiguous states sorted by census region). Highlighted states, including Minnesota and South Carolina, show substantial declines in death occurrences in 2004 and/or 2005. Counts of deaths to people born in these states generally show the same sudden declines. These patterns indicate that public Social Security mortality records may be highly incomplete for certain states in these years

Note: State of death is inferred from last ZIP code of residence on file with the Social Security Administration. ZIP codes are commonly missing from 1989-1992 data, contributing to deflated deaths counts in those years. From 1993 onward, less than 4% of records have missing or invalid ZIP codes. However, ZIP code missingness may vary across states and thus distort true trends in CenSoc-Numident coverage by death state.

Table A.1

	Dependent variable:			
	Age at death in years			
	(1)	(2)	(3)	(4)
	Numident	Numident	DMF	DMF
	1988-2005	1988-2005	1975-2005	1975-2005
	(unweighted)	(weighted)	(unweighted)	(weighted)
Race = Black	-0.656^{***}	-0.880^{***}	-0.889^{***}	-1.469^{***}
	(0.021)	(0.016)	(0.032)	(0.021)
Race = East Asian	0.623^{***}	0.908^{***}	1.557^{***}	1.950^{***}
	(0.109)	(0.096)	(0.139)	(0.138)
Race = Native American	-0.508^{***}	-0.237^{**}	-1.056^{***}	-0.495^{***}
	(0.106)	(0.093)	(0.147)	(0.137)
Constant	85.501^{***}	84.802^{***}	78.524^{***}	78.124^{***}
	(0.035)	(0.018)	(0.020)	(0.020)
Observations	1,256,065	1,256,065	1,267,146	1,267,146
R^2	0.171	0.201	0.018	0.018
Adjusted R^2	0.170	0.201	0.018	0.018
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A.2: **Relationship between race and longevity after age 65, including Alaska-born and Hawaii-born decedents** This model compares the effect of race on longevity after age 65 (reference category = White) between weighted and unweighted Numident and DMF. White, Black, Native American, and East Asian (Chinese and Japanese) men born in the contiguous United States, Alaska, and Hawaii 1910-1920 are included. All models include cohort fixed effects. The results are extremely similar to models excluding Alaska-born and Hawaii-born men. While it is not exactly appropriate to treat persons born in Alaska and Hawaii as"Native Born" due to the nature of CenSoc data, in practice it is likely inconsequential to categorize them as such.

Table A.3

	Dependent variable:					
	Age at Death in years					
	(1)	(2)	(3)	(4)	(5)	(6)
	Numident	Numident	Numident	DMF	DMF	DMF
	1988-2005	1988-2005	1988-2005	1975-2005	1975-2005	1975-2005
	Unweighted	HMD weights	NCHS weights	Unweighted	HMD weights	NCHS weights
Race = Black	-0.656^{***}	-0.634^{***}	-0.880^{***}	-0.889^{***}	-0.881^{***}	-1.469^{***}
	(0.021)	(0.020)	(0.016)	(0.032)	(0.032)	(0.021)
Race = East Asian	0.623^{***}	0.629^{***}	0.908^{***}	1.557^{***}	1.541^{***}	1.950^{***}
	(0.109)	(0.113)	(0.096)	(0.139)	(0.139)	(0.138)
Race = Native American	-0.508^{***}	-0.524^{***}	-0.237^{**}	-1.056^{***}	-1.040^{***}	-0.495^{***}
	(0.106)	(0.101)	(0.093)	(0.147)	(0.146)	(0.137)
Constant	85.501^{***}	84.466^{***}	84.802^{***}	78.524^{***}	78.187^{***}	78.124^{***}
	(0.035)	(0.017)	(0.018)	(0.020)	(0.020)	(0.020)
Observations	1,256,065	1,256,065	1,256,065	1,267,146	1,267,146	1,267,146
R^2	0.171	0.198	0.201	0.018	0.017	0.018
Adjusted R^2	0.170	0.198	0.201	0.018	0.017	0.018
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.4: **Relationship between race and mortality, comparison of different weights.** These models compare the effect of race on longevity after age 65 (reference category = White) between unweighted data, CenSoc Version 2.1 weights (calculated using HMD data) and CenSoc Version 3.0 weights (calculated using NCHS data and described in this paper). White, Black, Native American, and East Asian (Chinese and Japanese) men born in the contiguous United States 1910-1920 are included. All models include cohort fixed effects. Estimates using HMD weights are relatively similar to unweighted estimates. Estimates with NCHS weights, which actively account for systemic racial disparities in mortality coverage, are markedly different from either unweighted estimates or those produced with HMD weights. Notably, they imply a larger mortality disadvantage for Black men.

Table A.5

	Dependent variable:	
	Age at death in years	
	(1)	(2)
	Unweighted	Weighted
BPL = Canada	-0.468^{***}	-0.675^{***}
	(0.062)	(0.050)
BPL = Cuba	0.178	0.620^{*}
	(0.358)	(0.338)
BPL = England	0.211^{**}	0.316^{***}
	(0.105)	(0.104)
BPL = Germany	4.536^{***}	4.241^{***}
	(0.522)	(0.425)
BPL = Japan	0.395	0.706
	(0.571)	(0.608)
BPL = Mexico	0.638^{***}	0.687^{***}
	(0.111)	(0.101)
BPL = Puerto Rico	0.271	0.123
	(0.207)	(0.184)
Constant	85.480^{***}	84.782^{***}
	(0.036)	(0.018)
Observations	1,202,939	1,202,939
R^2	0.169	0.198
Adjusted R^2	0.169	0.198
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A.6: **Longevity of Non-US-born Men** These models show the relationship between country/territory of origin and longevity after age 65 for (reference category = US-born White) using unweighted and weighted CenSoc-Numident data. Men born 1910-1920 are included. All models have cohort fixed effects. Weights have varying effects on point estimates, but generally do not impact statistical inference. One exception is that Cuban-born men have a mortality advantage, statistically significant at the p = 0.1 level, when weights are utilized.