How similar is sibling mortality in later life? A feasibility study to examine determinants of population-level inequalities in mortality using US Census data

Jonathan Beauchamp, George Mason University
Jennifer Bernard, US Census
Ryan Fraser, University of Michigan
Amanda Kowalski, University of Michigan and NBER

While individual health is related between immediate family members, little is known about how these similarities persist into later life long after family members have stopped living together. These stylized facts, however, could have implications for population-level inequalities in later life mortality, which are observed to be substantial in the US (Chetty et al. 2016, JAMA Network). In this project, we will advance understanding of population-level inequalities in mortality in the United States by employing the US Census Bureau's Restricted Use Microdata to 1) estimate the resemblance of various types of siblings—same-sex twins, opposite-sex twins, and other sibling types—in later life mortality and 2) use these estimates to identify the relative role of genetic and childhood environmental contributions to inequalities in later life mortality. Our team has already received access to the data and has expertise in health economics, behavioral genetics, and the Census microdata. This pilot will be used to fund research assistance in FY 2026 to better construct family relationships and mortality outcomes in the US Census data and produce preliminary analyses that we will include in an NIH R01 to support our main analyses.

Pilot – Major goals

1. **Construct family relationship data from 1940 to present**. We will construct a large and diverse sample of sibling pairs using the near universe of American families for our period of study. These data will be built from a combination of Decennial Censuses from years 1940, 2000, 2010, and 2020 and the dependent fields of linked IRS 1040 forms from years 1969-1994 in five-year increments, 1995, and annually from 1998 to present. Additionally, our work requires the careful identification of different sibling types, including biological versus non biological siblings, and twin versus non-twin siblings as outlined below. For this goal, we draw from the frontier of the literature that uses these data to identify cohabitating adults and siblings from Census Personal Identification Keys complemented by probabilistic matching (Finlay et al. 2023, QJE; Stinson and Weiwu 2023, CES Technical Report).
2. **Clean mortality data and construct mortality outcomes for the entire population**. We have date of death from the Census Numident for our entire period of study, which we will clean and harmonize with our family relationship data to construct mortality outcomes for our cohort.
3. **Investigate cause-specific mortality in the Census microdata and external linkage to National Death Index data**. While we have detailed information on date of death, our present understanding is that the Census Numident does not contain cause-specific mortality for a large share of the population. We will first investigate whether there is cause-specific mortality in any of our linked administrative data, primarily those from the Centers for Medicare & Medicaid Services. We will then investigate how to link our data to population-level cause-specific mortality in the form of a proposal to the Census Bureau and National Death Index.

Work from this pilot will have several larger impacts. The data we construct will be the first step in **creating a resource available to researchers on approved projects** through the Census Bureau's Federal Statistical Research Data Centers. As part of our approved agreement with the Census Bureau, we will also **improve existing government data** by providing estimates of twin and other multiparous births and assessing the quality of the Census Person Identification Validation System for these individuals. Our work in preparing to link National Death Index data will be used to create a **proposal guide** for the process of linking external data to the Census.

Main analyses – Identification: The data constructed in this pilot will directly enable us to estimate the resemblance of siblings in later life mortality, and we can also use these estimates to identify the relative role of the genetic and environmental determinants of later life mortality by employing methods from behavioral genetics. Our identification hinges on **comparing same-sex and opposite-sex twins, netting out the effects of sex through comparisons to other sibling types**. At the core of this comparison is the fact that some same-sex twins are identical and share 100% of their genes, but all other siblings are fraternal and share only 50% on average. To illustrate our approach, imagine the comparison of an opposite-sex twin sibling pair A and B to a female twin sibling pair C and D. On average, siblings A and B will share 50% of their genes, but siblings C and D will share $(1 + p) \cdot 50\%$ of their genes, where p is the fraction of same-sex twins who are identical, which can be estimated directly as the excess share of same-sex twins beyond what is predicted by the population's sex ratio. Under several identification assumptions – namely additive genetic effects, no assortative mating, no gene-environment interactions, and equal family environment – we can interpret the greater similarity of same-sex twins as isolating their additional genetic similarity. With this condition met, we can estimate the **share of observed variance in mortality owing to genetic and environmental factors that are shared between siblings**.

Main analyses – NIH R01: The flexibility of the above framework affords us an unprecedented ability to estimate the relative role of genetic and environmental factors in observed inequalities in mortality for the near universe of the US population contained in the Census microdata. To support this effort, we will seek additional funding from an NIH R01. And to best fit the goals of the NIH, we have developed the following **specific aims** with guidance from Acting Deputy Director and Branch Chief, John Phillips, and Program Officer, David Braudt, of the Division of Behavioral and Social Research.

1. **Quantify the role of genetic and environmental factors in early versus late life mortality.** The scope of these data enables us to estimate the relative contribution of genetic and environmental factors for mortality outcomes at all stages of the life cycle. For example, greater environmental effects at later life would suggest that the consequences of inequality in shared environments are not apparent until many years after childrearing.
2. **Estimate variation in the role of genetic and environmental factors of mortality over time.** We will also use the scope of our data from 1940 to today to estimate how the relative importance of genetic and environmental factors *at a given age* has varied over time. In our framework, public health improvements over this time that affect all members of a family should reduce the relative importance of environment by reducing inequalities in shared environments. Detectable environmental effects thereby define the scope for policy impacts.
3. **Characterize the role of genetic and environmental factors of mortality by geography.** We will take advantage of the geographic diversity of our data to estimate how the relative importance of genetic and environmental factors of mortality vary regionally and by state. Our expectation is that places with higher levels of deprivation and inequality will exhibit a greater role of shared environment in mortality.

Beyond the novel results of our analyses for the above aims, this project is also significant in that it represents the **largest and most diverse application of behavioral genetics methods to date**. This diversity is likely to increase the impact of our work through several channels. First, we increase the external validity of our results relative to the behavioral genetics literature, which is restricted to smaller and more homogeneous samples than the Decennial Census. Second, previous work has hypothesized that the existing literature's lack of diversity in family environments may cause the framework above to underestimate the relative importance of shared environmental factors (Stoolmiller 1999, Psychological Bulletin). And third, this lack of diversity may also mask important heterogeneity if shared environment is more impactful in disadvantaged groups that are often excluded from previous literature (Turkheimer et al. 2003, Psychological Science). Lastly, in demonstrating what can be learned from using and improving population-level data, we set the stage for future researchers to further improve these resources.