

# Health effects of cousin marriage: Evidence from US genealogical records

Sam Hwang, Deaglan Jakob, Munir Squires

October 14, 2023

## **Abstract**

Cousin marriage rates are high in many countries today. We provide the first estimate of the effect of such marriages on the life expectancy of offspring. By studying couples married over a century ago, we observe their offspring across the lifespan. Using US genealogical data to identify children whose parents were first cousins, we compare their years of life to the offspring of their parents' siblings. We find that marrying a cousin leads to more than a three-year reduction in offspring life expectancy. This effect is strikingly stable across time, despite large changes in life expectancy and economic environment.

# 1 Introduction

The taboo against cousin marriage in the US and Europe is linked to the belief that children from these unions are likely to have genetic health problems. In many societies today, however, marriages between first or second cousins are common. A much-cited estimate is that these couples and their offspring make up 10% of the world’s population (Bittles and Black, 2010). In some countries the rate is much higher: about 50% of marriages in Pakistan are between first cousins.<sup>1</sup>

Biologists and medical researchers have converged on the conclusion that the effect of cousin marriage on offspring health is real but modest in magnitude (Bittles and Neel, 1994; Bittles and Black, 2010). According to an influential report by the US National Society of Genetic Counsellors, “There is a great deal of stigma associated with cousin unions in the United States and Canada that has little biological basis” (Bennett et al., 2002). The report concludes that the risks are smaller than assumed and do not justify any additional genetic testing.<sup>2</sup>

This paper documents health effects that are substantially larger than existing studies have found. This is in part because our data allow us to observe mortality beyond childhood, throughout the lifespan (Aizer et al., 2016). Doing so requires studying couples who were married a hundred or more years ago as well as their offspring. Since this data is not available in countries with high rates of cousin marriage today, we turn to historical US genealogical records to fill this gap. Our data allows us to directly identify first cousin marriages and measure years of life lived by the offspring of these and other marriages.

We first show a strikingly stable reduction in life expectancy for offspring of first-cousin

---

<sup>1</sup>The share of ever-married women ages 15-49 who report marrying a first cousin in the Pakistan DHS survey was 49.6% in the most recent 2017-18 round. Younger women report higher rates of cousin marriage, suggesting the practice is not in decline.

<sup>2</sup>This report was widely cited in debates on cousin marriage sparked by members of the UK parliament (Paul and Spencer, 2008).

marriages. This gap is consistent across birth cohorts from 1750-1900, and throughout the distribution of parental longevity.

To determine whether this difference reflects the causal effect of cousin marriage, we compare the children of married cousins to the children of the *siblings* of these married cousins. This empirical approach has the advantage of controlling for a wide range of potential genetic, economic and cultural sources of selection into cousin marriage. While this adjusts for unobserved characteristics shared by sets of siblings, differences between siblings who marry a first cousin and those who do not may still lead to selection bias. If siblings in worse health are more likely to marry a cousin, their children may live shorter lives for reasons unrelated to consanguinity. We test this by comparing the longevity of individuals who marry a cousin to their siblings. Reassuringly, we find no difference: within sets of siblings, marrying a cousin is not correlated with longevity-relevant characteristics.

Our main result is that the lives of children with first-cousin parents are three years shorter than the children of their parents' siblings. This is a five percent decrease from an average life expectancy of just under 60 years. Notably, we find that mortality effects accrue throughout adulthood.

A limitation of our genealogical data is that about two thirds of infant deaths are missing. Since rates of infant deaths are higher for married cousins, we show that this leads us to substantially *underestimate* the associated decline in life expectancy. Accounting for these missing infant deaths increases our estimate of the effect of cousin marriage to four years reduction in life expectancy.

Our results are robust to including county-by-decade of birth fixed effects to account for potential differences in health stock by location. We also show that a survival model produces the same results as our baseline model, though it does not allow us to implement our identification strategy.

These health effects are almost certainly genetic in nature. We find no evidence of substantial behavioral effects, either from a socio-economic decline for married cousins (Ghosh et al., 2023), or from changes in family size or in maternal age at birth. In support of a genetic channel, we find that life expectancy decreases less for more distantly related spouses (e.g., second cousins).

Finally, we provide evidence that our results are informative of the health costs of cousin marriage in countries where the practice is still common today.

This paper contributes to a literature in economics on the determinants of health, which emphasizes the importance of health stock on economic outcomes (see for example Currie et al., 2009; Strauss and Thomas, 2007, and citations therein). We contribute to this literature by documenting the health costs of a practice that, while now rare in developed countries, is widely practiced in many societies. We also advance the multi-disciplinary literature on the health effects of cousin marriage by providing the first estimate of its effects on cohort life expectancy.<sup>3</sup> Our results suggest the health costs are larger than previously estimated. We can do this because we study a population-scale set of offspring born sufficiently long ago that we can track their mortality rate throughout the lifespan.<sup>4</sup> A second contribution to this literature is our method of addressing selection into cousin marriage by restricting comparisons to close relatives who share genetic and other unobserved characteristics.<sup>5</sup> The only other paper in this literature to address selection

---

<sup>3</sup>The only somewhat similar study correlates estimates of consanguinity with life expectancy using cross-sectional country-level data (Saadat, 2011).

<sup>4</sup>Existing studies focus on infant or child mortality, which is more easily measured. There is significantly more disagreement on the effect of cousin marriage on adult health outcomes. For example, while some studies find a negative effect on adult health outcomes (Rudan et al., 2003a,b; Bener et al., 2007; Liede et al., 2002; Gilani et al., 2006), others find no effect (Bener et al., 2009; McWhirter et al., 2012; Bener et al., 2010; Denic et al., 2007). Like us, Helgason et al. (2008) and Kaplanis et al. (2018) use large-scale historical genealogical data to identify cousin marriages, though neither studies its effects on life expectancy.

<sup>5</sup>Our approach is similar to the sibling comparisons that are commonly used to address selection challenges (Abramitzky et al., 2012; Collins and Wanamaker, 2014; Ward, 2022; Lu and Vogl, 2023; Kreisman and Smith, 2023). One distinction is that we compare the *offspring* of siblings, rather than the siblings themselves.

on unobservables is Mobarak et al. (2019). They use an instrumental variables strategy that relies on variation in the availability of marriageable opposite-sex cousins. They find modest but noisy increases in under-5 mortality and prevalence of genetic diseases amongst the children of Bangladeshi and Pakistani respondents.

This paper also contributes to an economics literature on the role of culture in shaping marriage and family decisions (Fernández, 2011; Giuliano and Nunn, 2021). While there is evidence that cousin marriage plays a functional role (for example by managing inheritance (Bahrami-Rad, 2021) or by making commitments between the groom and bride’s families more credible (Mobarak et al., 2013)), anthropologists have long emphasized how deeply cultural the practice of cousin marriage is. Ghosh et al. (2023), for example, show that surname-specific rates of cousin marriage in the US are highly persistent over time. In that vein, our results add to the literature on the health and economic costs of cultural practices (Lowes and Montero, 2021; Corno et al., 2020; Atkin, 2016; Almond and Mazumder, 2011).

## 2 Genealogical data

Our measures of longevity and family ties come from FamilySearch, a genealogical website where users can view historical records and enter information about their ancestors. A main feature of the site is a set of public genealogical profiles of historical individuals, creating large, interlinked family trees.<sup>6</sup> The dataset we use in this paper consists of 40 million of these linked individual profiles. Using these, we trace the genealogies of individuals to identify cousin marriages and study their effect on longevity.

Our sample was obtained by first collecting all US marriage records up to the mid-nineteenth century available on FamilySearch. Using the profiles of these spouses, we

---

<sup>6</sup>See Hwang and Squires (2023), Price et al. (2021) and Blanc (2023) for evidence on the quality of this type of genealogical data. See Appendix Figure A.1 for a sample genealogical profile.

expanded our sample both horizontally (siblings and siblings-in-law), and vertically (parents and children). Our resulting sample is about half the size of the early-mid nineteenth century US population (Appendix Figure A.2). See Appendix B for more information on how this dataset was collected, and evidence on its representativeness.<sup>7</sup>

This section describes two key features of this dataset. First, for about a quarter of our sample we can have sufficient vertical genealogical links to determine whether their parents were first cousins. Second, these profiles include birth and death years to measure longevity, which we use as a summary measure of health. However our data come with potential challenges: infant deaths are underreported, and vital dates are sometimes recorded with error. We address these below.

To identify cousin marriages, we need to observe all four grandparents of both spouses. For the married couples in our data for whom we have all grandparents, we determine whether spouses are cousins by checking for overlap in the two sets of grandparents. The vast majority of spouses in our data have either zero (97.3%) or two (2.5%) overlapping grandparents.<sup>8</sup> Spouses with two grandparents in common are first cousins, while those with no matching grandparents are not (first) cousins. Appendix Figure A.3 illustrates the family tree of two spouses who are first cousins.

To evaluate how marrying a cousin affects the health of offspring, we use as an outcome their years of life ('longevity'). Since genealogical profiles do not contain direct information on health, such as diseases, disability, or cause of death, we treat longevity as a proxy for overall, life-time health. Measuring longevity simply requires us to take the difference

---

<sup>7</sup>As documented elsewhere, US genealogical records underrepresent individuals for whom few records exist (Hwang and Squires, 2023; Price et al., 2021). This is particularly true for enslaved individuals, for whom it is challenging to document genealogical links across generations.

<sup>8</sup>We omit the 0.2% of couples with one, three or four matching grandparents from our analysis for simplicity and because of insufficient statistical power. Having one matching grandparent would mean the spouses are half-first-cousins, and having three or four implies they are half or full siblings, or double first cousins.

between individuals’ birth and death years, which are available for about three-fourths of our sample.

The sample we use for analysis meets the following two criteria: non-missing data on all eight great-grandparents (their mother’s and father’s four grandparents), and non-missing birth and death years. We also drop individuals with missing data or errors in their genealogical profiles, and restrict our sample to those born between 1750 and 1920.<sup>9</sup> Of the 40 million individuals in our dataset, the 6.6 million who meet these criteria form our ‘analysis sample’ (Appendix Table A.1).

A major challenge with using genealogical records to estimate longevity is that offspring who die young are often missing. As we discuss in section 4.2, infant (age zero) mortality in our data is indeed much lower than existing estimates. Since cousin marriages have higher rates of infant deaths, we show that this under-reporting attenuates our estimates.

A second weakness of our dataset is that vital years may be recorded with error. Indeed, we observe heaping in death years, which unlike birth years cannot be obtained from census records. This is likely because, in the absence of alternative records, genealogical researchers use the last census year where an individual was observed as their year of death. After presenting our main results, we show that these errors do not bias our estimates.

Descriptive statistics of our analysis sample are presented in Appendix Table A.2. We split the sample between children of married cousins (2.5% of the sample) and non-cousins (97.5%) to illustrate basic differences between these two groups.<sup>10</sup> Most notably, the average longevity of offspring of cousin marriages is 55 years, relative to 58 years for the offspring

---

<sup>9</sup>Dropping individuals born after 1920 limits potential selection out of sample by individuals who are still alive. We drop individuals with missing sex or maternal age at birth (used as controls), and those whose longevity is negative or above the 99th percentile (98 years old). Individuals with impossible familial links (e.g., children being their own parents) are dropped from the sample. We discuss these sample construction procedures in Appendix Section B. To provide a consistent sample across analyses, we also drop the 0.1 million singleton observations that get dropped from our fixed effects estimation.

<sup>10</sup>See Appendix Figure A.4 for an illustration of how the rate of first-cousin marriage changes over time in our data.

of non-cousins, a three-year difference. Parental longevity, defined as the average of the mother and father’s longevity, is also shorter. Married cousins live just over half a year shorter lives. If individuals who marry their cousins are themselves in worse health, their offspring’s health cannot easily be compared to that of the population at large.

Before presenting our empirical design, which aims to address potential selection into cousin marriage, we show in Figure 1 that this longevity gap for offspring is strikingly stable. Panel (a) shows how offspring life expectancy changes with parental longevity. As expected, the offspring of longer-lived parents also live longer (Black et al., 2023). Differences in parental longevity presumably reflect a combination of genetic, socio-economic, and geographic differences between families. The figure highlights that the difference in life expectancy for offspring of married cousins is large and stable across the parental longevity distribution.

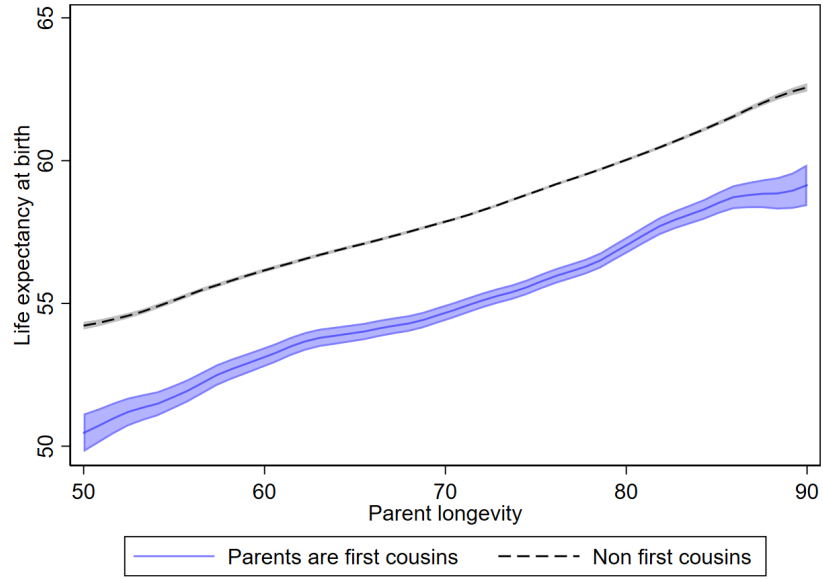
Panel (b) of Figure 1 likewise documents that this difference is stable across birth cohorts. While cohort life expectancy changed substantially in our dataset between 1750 and 1900, the gap for offspring of first-cousin parents remains stable.<sup>11</sup> Offspring of cousin marriages have consistently lower life expectancy than those born of non-cousins, despite dramatic economic change and structural transformation in the US during this period. We also find that this difference is stable across locations (Appendix Figures A.6 and A.7) despite large differences in mortality that are likely to be at least partly causal (Finkelstein et al., 2021).

---

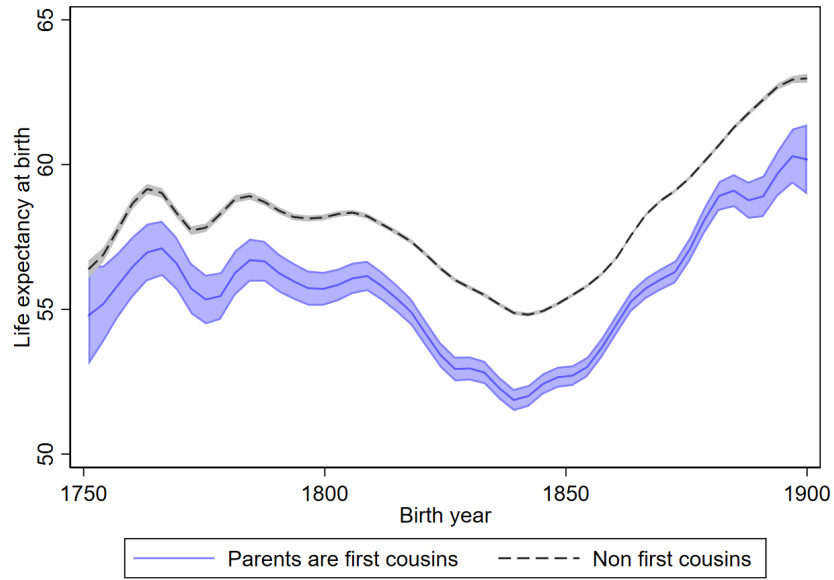
<sup>11</sup>The broad pattern of life expectancy at birth in our data is consistent with findings in Hacker (2010). Notice that the US Civil War had a large negative effect on life expectancy, especially for individuals born in the 1830s and 40s. The gap in adult life expectancy is also large and stable across birth cohorts (Appendix Figure A.5).



Figure 1: Cousin marriage and life expectancy at birth



(a) Life expectancy by parental longevity



(b) Life expectancy by birth cohort

This figure depicts the life expectancy at birth of our analysis sample of 6.5 million offspring. Children of first cousins are represented by the solid lines and children of non first cousins are represented by the dashed lines. Panel (a) is a local polynomial regression of life expectancy on parent longevity. Parent longevity is the average of the child's mother and father's longevity. Panel (b) is a local polynomial regression of life expectancy at birth on birth year.

### 3 Empirical design

#### 3.1 Regression specification

Our analysis studies children of married cousins, and compares them to the children of their parents' siblings. We estimate the effect of cousin marriage on years of life (*Longevity*) of children born of these marriages using the following empirical specification:

$$Longevity_i = \beta FirstCousinParents_i + \lambda_m Maternal_m + \lambda_p Paternal_p + \mathbf{X}_i' \boldsymbol{\delta} + \epsilon_i, \quad (1)$$

where  $i$  is an individual in our analysis sample. The treatment variable  $FirstCousinParents_i$  is equal to 1 if  $i$ 's parents are first cousins, and 0 if not. Subscripts  $m$  and  $p$  denote maternal (mother's side) and paternal (father's side) relatives of  $i$ . Specifically, each individual  $i$  shares a value of  $m$  with all children of  $i$ 's mother and maternal aunts (mother's sisters). Likewise,  $i$  shares a value of  $p$  with all children of  $i$ 's father and his brothers.  $Maternal_m$  is equal to one for all individuals with the same value of  $m$ , and zero otherwise, as with  $Paternal_p$ . Appendix Figure A.8 illustrates the relevant comparison groups for  $i$  implied by the  $Maternal_m$  and  $Paternal_p$  fixed effects. Finally,  $\mathbf{X}_i$  is a vector of individual-level controls adjusting for sex, year of birth, maternal age at birth, number of siblings, sibling sex ratio, and birth order. These are described in Appendix C. Standard errors are clustered at the level of full siblings.

The maternal and paternal fixed effects allow us to restrict comparisons to close relatives. These serve as a useful control group as they share a wide range of unobserved economic, social, and genetic characteristics. Within-family comparisons address potential concerns that arise out of non-random selection of families (sets of siblings) into higher or lower rates of cousin marriage.<sup>12</sup> Our key identifying assumption is that, *within* sets of siblings,

---

<sup>12</sup>This approach addresses another potential concern: marrying a cousin might also mean choosing a

selection into cousin marriage is independent of other traits that might affect offspring longevity. The following section tests this assumption.

### 3.2 Test of key identifying assumption

Our data allow us to directly test whether the longevity of individuals who marry a cousin differs from their siblings. To do so, we implement an empirical specification similar in principle to equation (1), but where the units of observation are parents of those in the analysis sample, and the outcome is the longevity of these parents. As direct equivalents of the maternal and paternal fixed effects, we add fixed effects for each parent’s same-sex siblings. We refer the reader to Appendix D for more details on this parent-level analysis, including the sample and the regression specification. These are, to the extent possible, direct analogues of the main analysis described above.

Column (1) of Table 1 reports that the raw difference in longevity between parents who married their cousin and those who did not is 0.6 years. Individual-level controls in column (2) reduce the estimated coefficient slightly.

This difference disappears entirely when we include sibling fixed effects in column (3) of Table 1. Adding these fixed effects means that we compare the longevity of individuals to their same-sex siblings. That is, fathers are compared to their brothers, and mothers to their sisters. Restricting to within-sibling comparisons reduces the coefficient to zero, with a reasonably precise confidence interval.

Importantly, we are not claiming that the effects reported in Table 1 are causal: indeed we are interested in selection into cousin marriage, and wish to test whether siblings who marry their cousins are in worse health. That we find no correlation with longevity after

---

spouse with traits linked to higher or lower offspring longevity. Our identification strategy deals explicitly with this concern through the combination of maternal *and* paternal fixed effects. Hence the choice of spouse also affects who one’s children will be compared to.

Table 1: Selection of parents into cousin marriage (placebo check)

	(1)	(2)	(3)
	Raw	Controls	Same-sex sibling fixed effects
<b>Parent Longevity</b>			
Married to first cousin	-0.61*** (0.09)	-0.50*** (0.09)	-0.07 (0.12)
Control mean	67.75	67.75	67.75
Observations	1,287,986	1,287,986	1,287,986
Controls	No	Yes	Yes
Same-sex sibling FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of parents. The outcome is longevity (year of death minus year of birth). The coefficients in the first column are simply the difference in means between those who marry their first cousins and those who do not. The second and third columns controls for birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in appendix C. In the third column we include wife's siblings fixed effects and husband's siblings fixed effects.

including sibling fixed effects suggests that within-sibling comparisons adequately control for the relevant confounding variables. Restricting our comparisons to the children of these same-sex siblings should therefore allow us to recover the causal effect of cousin marriage on the health of offspring.

## 4 Results

### 4.1 Main results

What effect does cousin marriage have on the life expectancy of offspring? Table 2 presents results from OLS regressions that report the difference in longevity between children of cousins and of non-cousins. Each observation is a person (‘offspring’) in our analysis sample. Column (1) of Panel A reports that individuals born of first cousins live on average about 3 fewer years. This coefficient and all others we report in this table are highly statistically significant ( $p < 0.001$ ), with standard errors clustered at the level of siblings. Adding individual controls in column (2) reduces the life expectancy gap between offspring of cousin and non-cousin parents to about 2.6 years.<sup>13</sup>

Column (3) presents our preferred estimates using maternal and paternal fixed effects to control for any factors common to the children of an individual’s aunts and uncles. These include a wide range of unobserved economic, cultural and genetic factors common to these close relatives. Including maternal and paternal fixed effects suggests that first-cousin marriage causes offspring longevity to decrease by 3.3 years. Mean life expectancy in this sample is about 60 years, so the coefficient on having first-cousin parents corresponds to a five percent decline.

To help interpret the magnitude of this result, we can compare it to the relationship

---

<sup>13</sup>We control for sex as well as quadratic terms for birth year, maternal age at birth, number of siblings, sibling sex ratio, and birth order. Appendix C describes each of these. Our results do not change if we use a full set of fixed effects instead of quadratic terms (Appendix Table A.4).

Table 2: The effect of cousin marriage on offspring longevity

	(1)	(2)	(3)
	Raw	Controls	Maternal and paternal fixed effects
<b>Panel A: Life expectancy at birth</b>			
Parents are first cousins	-3.09*** (0.09)	-2.63*** (0.09)	-3.27*** (0.33)
Control mean	58.00	58.00	58.00
Observations	6,516,999	6,516,999	6,516,999
<b>Panel B: Life expectancy at age 5</b>			
Parents are first cousins	-2.75*** (0.07)	-2.19*** (0.07)	-2.22*** (0.29)
Control mean	63.71	63.71	63.71
Observations	5,894,444	5,894,444	5,894,444
<b>Panel C: Life expectancy at age 20</b>			
Parents are first cousins	-2.39*** (0.06)	-1.86*** (0.06)	-1.78*** (0.26)
Control mean	66.73	66.73	66.73
Observations	5,550,001	5,550,001	5,550,001
Individual controls	No	Yes	Yes
Paternal FE	No	No	Yes
Maternal FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. This table shows estimates for the coefficient  $\beta$  from equation (1) estimated using OLS. The coefficients in the first column are simply the difference in means between the children of first cousins and the children of non-first cousins. The second and third columns control for birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in appendix C. In the third column we include mother's siblings fixed effects and father's siblings fixed effects as shown in Appendix Figure A.8. Standard errors are clustered at the level of the individual and their siblings.

between parent and offspring longevity we documented in panel (a) of Figure 1. A decline of 3.3 years in offspring life expectancy corresponds to a drop from the 75th to the 25th percentile of the parent longevity distribution, or from 78 years to 62.

These differences in longevity are driven by higher mortality throughout the lifespan. Table 2 also reports results on life expectancy at age 5 (Panel B) and 20 (Panel C). Coefficients remain substantial in magnitude, which suggests that changes in longevity are not just a result of differences in infant or child mortality that have been the focus of the existing literature. Conditional on living until the age of 20, offspring of first cousins live on average 1.8 fewer years than the offspring of their parents' siblings.

Figure 2 more flexibly documents how survival rates differ for offspring of first cousins across the lifespan. Panel (a) shows raw offspring survival rates. It suggests that the gap in survival grows gradually over the lifespan, rather than being concentrated in one period such as infancy. Panel (b) shows regression coefficients and 95% confidence intervals that compare survival rates for offspring of cousin and non-cousin marriages. It shows, for example, that having first-cousin parents leads to a two percentage point lower probability of surviving past age five.<sup>14</sup> This second panel confirms that the pattern observed in Panel (a) holds after controlling for selection. Indeed we see a consistently increasing gap between the survival rates of offspring from cousin and non-cousin marriages.

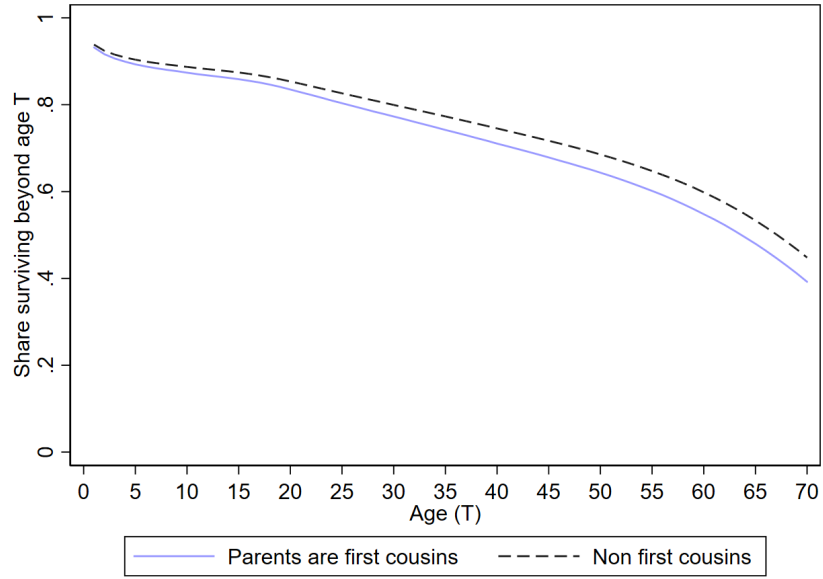
## 4.2 Underestimation of effect size due to missing infant deaths

Genealogical data such as ours are likely to underreport infant deaths since births were not consistently recorded in the US in the nineteenth century. As a benchmark, we compare

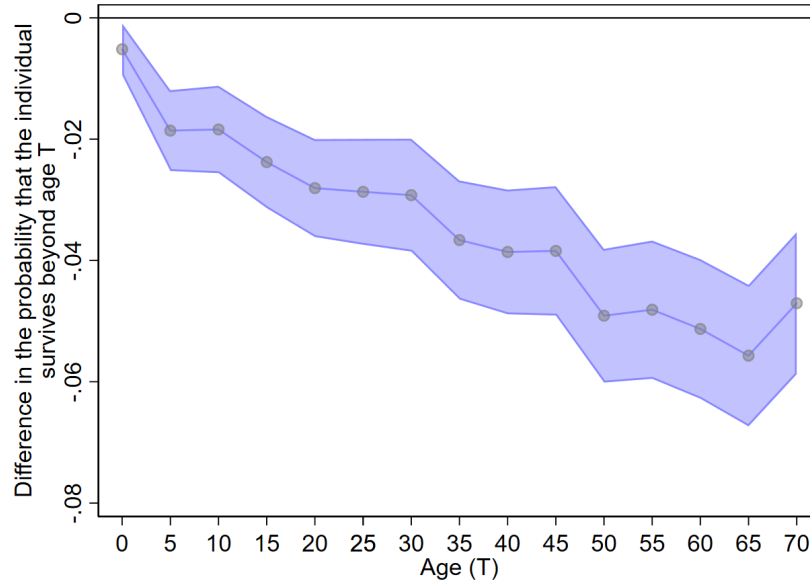
---

<sup>14</sup>Each estimate in panel (b) is from a regression where the outcome is whether an individual survived to a given age. Otherwise, these regressions are identical to equation (1), and include maternal and paternal fixed effects.

Figure 2: Mean and estimated survival probabilities



(a) Average survival rates



(b) Estimated effect of cousin marriage on survival rates

This figure depicts the probabilities of survival for our sample of 6.5 million children. Panel (a) shows average survival rates (without controls) for individuals whose parents are first cousins (solid line) and not first cousins (dashed line). Panel (b) shows estimates analogous to the coefficient  $\beta$  from equation (1) estimated using OLS and corresponding 95 % confidence intervals. Estimates include all of the controls and fixed effects used in column (3) of table 2. For each estimate, the outcome variable is equal to 1 if the individual survives beyond age T, and 0 otherwise.



infant and child mortality rates in our analysis sample to estimates from Hacker (2010).<sup>15</sup> Appendix Figure A.9 shows that our genealogical records do underreport infant deaths, relative to this benchmark. The infant (age zero) mortality rate is about 5% in our data, versus 14% in Hacker (2010). Mortality rates at ages 1 onward appear to match the benchmark figures far more closely, suggesting the underreporting in our data is specific to infants who died in the year they were born.

As we discuss at length in Appendix E, our estimate of the effect of cousin marriage on life expectancy is almost certainly an underestimate of the true effect. This is true if cousin marriage leads to higher rates of infant deaths (as shown in studies where infants deaths are not likely to be underreported), and if the probability of an infant death going unrecorded is similar for offspring of cousins and non-cousins. We formally state and provide evidence to support these assumptions in Appendix E. Adjusting for missing infant deaths increases our estimate of the effect of first-cousin marriage to a 3.9 year reduction in life expectancy (Appendix Table A.11). This would imply that first-cousin marriage caused an eight percent reduction in offspring life expectancy.

### 4.3 Sensitivity and data limitations

*Errors in vital dates.* Birth and death years in our data may be recorded with error. Indeed we observe heaping of recorded deaths (but not births) in years ending in zero (Appendix Figure A.10). This is likely because in the absence of other records, users filling out these genealogical profiles sometimes rely on the last census record available for a historical person to estimate their year of death.<sup>16</sup> Dropping all individuals from our sample whose year of death ends in zero has very little effect on our estimates (Appendix Table A.3).

---

<sup>15</sup>To our knowledge these are the best available estimates of mortality rates during our period of interest, though they are extrapolations from non-representative sub-samples.

<sup>16</sup>This conjecture is strengthened by the fact that there is no heaping in 1890, the only census round between 1850 and 1940 for which the underlying micro census data is not available.

*Flexible controls.* In our baseline specification, we control for quadratic functions of birth year, maternal age at birth, number of siblings, birth order, and sibling sex ratio. We show that replacing these with fixed effects to allow for more complex non-linearities does not affect our main results (Appendix Table A.4).

*Parental longevity.* Table 1 shows that, on average, the difference in longevity of married cousins goes to zero once we add maternal and paternal fixed effects. However, this result may mask some systematic differences that average out to zero. Our results are robust to controlling for parent longevity directly, suggesting this is unlikely to be consequential (Appendix Table A.5).

*Within-county analysis.* While married cousins may not differ from their siblings in longevity, they may systematically choose to live in different locations (e.g., stay on the farm as in Ghosh et al. (2023)). We know the county of birth of about three quarters of our sample. For this subsample, we find that results are robust to controlling for county-by-decade-of-birth fixed effects (Appendix Table A.6).

*Residual variation.* Our preferred specification includes 1.6 million fixed effect groups (0.8 million each for the paternal and maternal fixed effects). How much variation do these fixed effects absorb, and does sufficient variation remain to reliably estimate our treatment effect? Appendix Figure A.11 shows a set of histograms of the treatment and outcome variables (*FirstCousinParents*, and *Longevity*). Each panel overlays the raw and residualized distributions of these variables for comparison. The residualized variables retain substantial variation. Correspondingly, standard errors in Table 2 increase with the inclusion of the maternal and paternal fixed effects, but remain modest in absolute terms.

*Survival model.* To allow for computationally feasible estimation with over a million fixed effect groups, we estimate equation (1) using OLS. However it is common in the literature on longevity and mortality to estimate survival models. This is in part because

such models can address censoring of observations for individuals who have not yet died, a problem we do not face thanks to the historical nature of our data. We estimate a Cox Proportional Hazards model in Appendix F, and find a treatment effect that differs very little from our OLS estimate.

#### 4.4 Genetic and non-genetic channels

Are these effects purely genetic? Or are there important social or economic consequences of cousin marriage the lead to offspring having shorter lives? While the medical and population genetics literatures have focused on genetic effects, recent work in economics by Ghosh et al. (2023) suggests there may be important economic consequences to cousin marriage. Notably, they find that cousin marriage leads to lower incomes and reduces rural-to-urban migration. We describe two plausible non-genetic channels through which cousin marriage could affect offspring longevity, and discuss evidence from our dataset that can speak to these. We conclude that genetic channels seem to explain most or all of the observed treatment effects.

*Socio-economic channels.* The first broad type of non-genetic channel is economic. In a dynamic, rapidly industrializing society, the cost of low geographic and occupational mobility may mean children of married cousins are raised in poorer households and with less human capital (Ghosh et al., 2023). Their shorter lives may simply be a result of their relative decline in socio-economic status.

Estimating the strength of this channel is challenging given the genealogical profiles we use in this paper do not include income, occupation or education. However, we can test this channel using a subsample of our dataset that we linked to the IPUMS 1% random samples of the 1850 to 1930 census rounds (Ruggles et al., 2023).<sup>17</sup> For this small subset of our

---

<sup>17</sup>We describe this linking procedure in Hwang and Squires (2023). For each census round we link the IPUMS 1% random sample census individuals to their FamilySearch profiles (whenever the profile exists). For individuals living with their father at the time of the census, we can assign a father’s occupation to the child, and their all their siblings (whether they appear in the 1% sample or not). The resulting linked

data we can observe father’s occupation, which we use as a proxy of socio-economic status. Appendix Table A.7 shows that results are unchanged when adding father’s occupation fixed effects.<sup>18</sup>

Further, we argue that if cousin marriage reduces mobility, the associated health costs are likely to change as the US transitions from an agrarian, mostly local and isolated economy, to one that is increasingly industrial, urban and connected through railroads. As documented in panel (b) of Figure 1, however, the difference in life expectancy associated with cousin marriage did not change appreciably over 150 years of birth cohorts who lived through dramatic economic and technological changes.

*Fertility and parental investment.* The second broad non-genetic channel is tied to childbirth and child rearing. Ghosh et al. (2023) find that cousin marriage is linked to a higher rate of child marriage. Along these lines, it may be that cousin marriage reduces maternal age at first birth, or prolong births to a more advanced maternal age. It may also increase fertility (number of offspring), perhaps reducing investment in each child.<sup>19</sup>

To account for this, our baseline regressions include controls for each child’s maternal age at birth and for family size (number of siblings). Estimates are unaffected by the addition of these controls (Appendix Table A.8).

We also test whether marrying a cousin leads to a change in maternal age at birth or in the number of offspring. Results suggest that maternal age at birth and fertility are unlikely to be important channels. We find that married cousins have *fewer* offspring than their siblings, and that maternal age is unaffected (Appendix Tables A.9 and A.10).<sup>20</sup>

---

sample drops from 6.6 to 0.15 million.

<sup>18</sup>Due to the much reduced sample, we estimate this model without maternal and paternal fixed effects.

<sup>19</sup>The literature on the effects of cousin marriage on family size is surprisingly mixed. Studies have with roughly equal frequency found both positive (Bittles et al., 2002; Hussain and Bittles, 1999; Hosseini-Chavoshi et al., 2014) and negative (Ober et al., 1999; Hussain and Bittles, 2004) effects of cousin marriage on fertility.

<sup>20</sup>This includes maternal age at first and last birth. Results from these tables use the same parent-level specification as in Table 1, described in Appendix D.

*Genetic channel.* This leaves genetic effects as the main plausible channel for the large effects that we document. While we do not have direct evidence on congenital illnesses, for a subsample of our data we can compare the life expectancy of offspring of first cousins to offspring of more distantly related spouses: first cousins once removed and second cousins. Appendix Figure A.12 suggests that the longevity cost of cousin marriage decreases for more distantly related cousins, consistent with the smaller proportion of alleles inherited from a common source.<sup>21</sup>

## 5 External validity

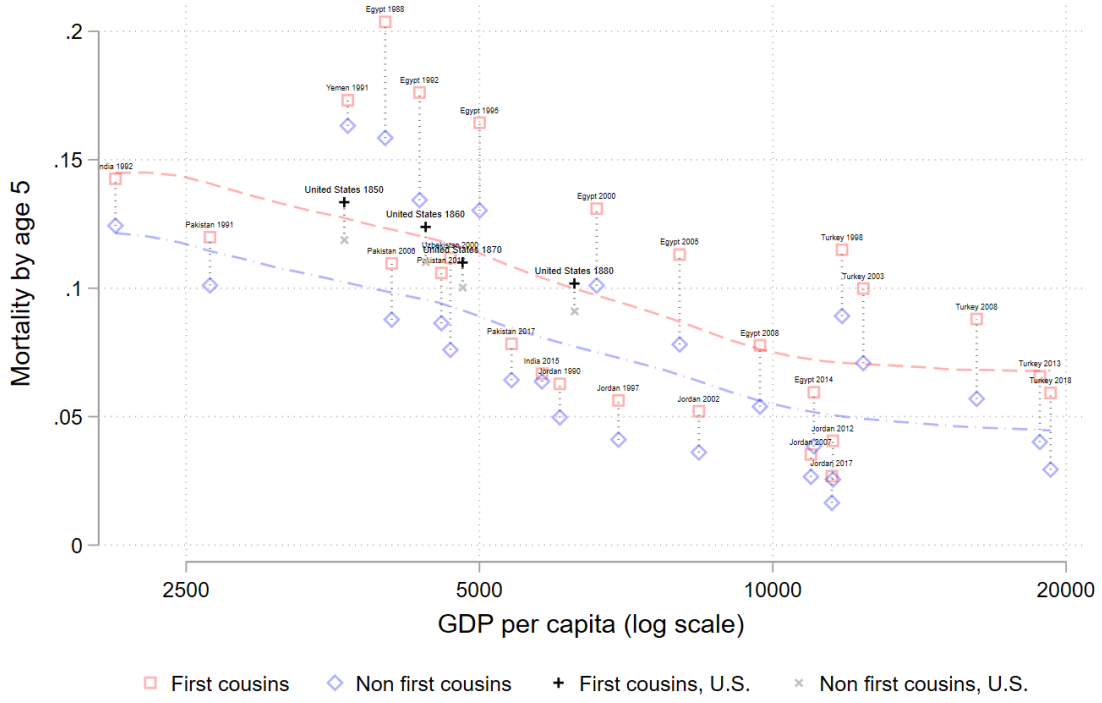
How informative are our results about the health costs of cousin marriage in countries where it is still commonly practiced today? The stability of our results across 150 years of US history seem to suggest that these effects are insensitive to the changes induced by technological progress or structural transformation. Nonetheless, medical advances in the twentieth century, including genetic counselling and testing, may have led to a reduction in the health costs of first-cousin marriage. To test this, we use data from the Demographic Health Survey (DHS). In countries with high rates of cousin marriage, the survey asks women whether their husband is their first cousin. This question was asked in 26 waves of surveys across six countries: Egypt, India, Jordan, Pakistan, Türkiye, Uzbekistan, and Yemen. For each of these surveys, we plot in Figure 3 the rate of child mortality separately for women married to a first cousin or not.<sup>22</sup>

---

<sup>21</sup>This result comes with two important caveats. First, these estimates come from the subset of our data for which we can determine whether the spouses are second cousins. For this we require all sixteen great-great-grandparents of the offspring in question. Second, they do not include the maternal and paternal fixed effects, as these absorb too much of the variation to estimate these effects. The ratio in effect sizes for first and second cousin marriages is consistent with the results in Saggar and Bittles (2008), which reviews the literature on infant and child health effects of cousin marriage.

<sup>22</sup>Child mortality is defined as a death from birth to age five. In each survey wave, the respondents are a nationally-representative sample of women ages 15-49. See figure notes for more details.

Figure 3: Cross-country child mortality differences by cousin marriage



Note: This figure describes differences in child mortality across countries and years. Our sample of countries are Egypt, India, Jordan, Pakistan, Türkiye, Uzbekistan, and Yemen. We chose these countries from the entire list of countries that ever conducted the Demographic and Health Survey (DHS) because they surveyed first cousin marriage status. The year and the number of surveys differ across countries. For each of these country-years, we map the share of children who died at or before age 5 to the GDP per capita for the corresponding country-year (Bolt and Van Zanden, 2020). The statistics for first-cousin couples are marked with squares, and those for non-first-cousin couples are marked with diamonds. Markers for the same country-year are connected with a dotted line. The dashed line is the kernel-weighted local polynomial fitted to first-cousin markers, and the dash-dot line is that for non-first-cousin markers. The crosses and the X's are markers for the United States. For each census year between 1850 and 1880, we impose on our analysis sample the same gender and age restriction as the DHS (i.e., females between the age of 15 and 49) and calculated the share of children who died at or before age 5 by each census year. We reweighted our sample to match the age distribution of the female population in each census.

In every survey wave, the child mortality rate for women married to a first cousin is higher than for non-cousins, often substantially so. Local polynomial best-fit lines trace out the relationship between child mortality and per capita income for each country at the time of the corresponding survey wave. These lines show that child mortality rates are on average about two to three percentage points higher for first-cousin spouses. Consistent with our findings using historical US data, this difference seems to be independent of income per capita and of the baseline mortality rate.

Figure 3 also includes estimates from the US in 1850, 1860, 1870 and 1880.<sup>23</sup> The corresponding US results suggest that our historical US data come from a society at a stage of development roughly comparable to many of the countries in these DHS survey waves, in terms of both income per capita and child mortality rates. While the results in Figure 3 do not address selection, they suggest that our estimate of the effect of cousin marriage on life expectancy is informative about its costs in countries with high contemporary rates of cousin marriage. Further, as higher incomes and reductions in communicable diseases increase life expectancy and reduce child mortality, cousin marriage is likely becoming an increasingly important public health concern in relative terms.

## 6 Conclusion

This paper uses forty million genealogical profiles to study the effect of first-cousin marriage on the health of offspring. Causal estimates come from comparisons between the offspring of married cousins and the offspring of their siblings. We find that cousin marriage reduces offspring longevity by three years, a reduction of about five percent. This difference is the result of increased mortality throughout the lifespan, which highlights the importance of

---

<sup>23</sup>For each of these years, we include mothers in our sample between the ages of 15-49, weighted to represent the country-wide age distribution taken from each of the respective census rounds.

studying adult health outcomes of offspring of cousin marriages. Strikingly, we also find that these effects are stable across 150 years of birth cohorts. Dramatic transformations in the US during this period implies that these effects are not very sensitive to the social or economic environment at a given time. This is consistent with data from countries with high contemporary rates of cousin marriage. At a first approximation, our findings imply that one seventh of the difference in life expectancy between the US and Pakistan is due to high rates of first-cousin marriage in the latter country.<sup>24</sup> These large effects suggest that cheaper and more widely available genetic counselling and screening may lead to substantial public health improvements.

---

<sup>24</sup>Using 2020 World Bank estimates of life expectancy of 77.3 and 66.3 years for the US and Pakistan, respectively. We assume first-cousin marriage rates of 0% and 50%, and a reduction of 3.3 years for offspring of cousin marriages.



## References

- Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012) “Europe’s tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration,” *American Economic Review*, 102 (5), 1832–1856.
- Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney (2016) “The long-run impact of cash transfers to poor families,” *American Economic Review*, 106 (4), 935–971.
- Almond, Douglas and Bhashkar Mazumder (2011) “Health capital and the prenatal environment: the effect of Ramadan observance during pregnancy,” *American Economic Journal: Applied Economics*, 3 (4), 56–85.
- Atkin, David (2016) “The caloric costs of culture: Evidence from Indian migrants,” *American Economic Review*, 106 (4), 1144–1181.
- Austin, Peter C (2009) “Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples,” *Statistics in medicine*, 28 (25), 3083–3107.
- Bahrami-Rad, Duman (2021) “Keeping it in the family: Female inheritance, inmarriage, and the status of women,” *Journal of Development Economics*, 153, 102714.
- Bener, Abdulbari, Hanadi R El Ayoubi, Lotfi Chouchane, Awab I Ali, Aisha Al-Kubaisi, Haya Al-Sulaiti, and Ahmad S Teebi (2009) “Impact of consanguinity on cancer in a highly endogamous population,” *Asian Pac J Cancer Prev*, 10 (1), 35–40.
- Bener, Abdulbari, Hanadi Rafii El Ayoubi, Awab Ibrahim Ali, Aisha Al-Kubaisi, and Haya Al-Sulaiti (2010) “Does consanguinity lead to decreased incidence of breast cancer?” *Cancer epidemiology*, 34 (4), 413–418.

- Bener, Abdulbari, Rafat Hussain, and Ahmad S Teebi (2007) “Consanguineous marriages and their effects on common adult diseases: studies from an endogamous population,” *Medical Principles and Practice*, 16 (4), 262–267.
- Bennett, Robin L, Arno G Motulsky, Alan Bittles et al. (2002) “Genetic counseling and screening of consanguineous couples and their offspring: Recommendations of the National Society of Genetic Counselors,” *Journal of genetic counseling*, 11 (2), 97–119.
- Bittles, Alan H (2012) *Consanguinity in context*, 63: Cambridge University Press.
- Bittles, Alan H and Michael L Black (2010) “Consanguinity, human evolution, and complex diseases,” *Proceedings of the National Academy of Sciences*, 107, 1779–1786.
- Bittles, Alan H, JC Grant, SG Sullivan, and R Hussain (2002) “Does inbreeding lead to decreased human fertility?” *Annals of human biology*, 29 (2), 111–130.
- Bittles, Alan H and James V Neel (1994) “The costs of human inbreeding and their implications for variations at the DNA level,” *Nature genetics*, 8 (2), 117–121.
- Black, Sandra E, Neil Duzett, Adriana Lleras-Muney, Nolan G Pope, and Joseph Price (2023) “Intergenerational Correlations in Longevity,” Working Paper 31034, National Bureau of Economic Research, 10.3386/w31034.
- Blanc, Guillaume (2023) “Crowdsourced Genealogies,” working paper, University of Manchester.
- Bolt, Jutta and Jan Luiten Van Zanden (2020) “Maddison project database, version 2020,” *Maddison Style Estimates of the Evolution of the World Economy*, <https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020?lang=en>.

- Chong, Michael, Diego Alburez-Gutierrez, Emanuele Del Fava, Monica Alexander, and Emilio Zagheni (2022) “Identifying and correcting bias in big crowd-sourced online genealogies,” Working Paper 2022-005, Max Planck Institute for Demographic Research, 10.4054/MPIDR-WP-2022-005.
- Collins, William J and Marianne H Wanamaker (2014) “Selection and economic gains in the great migration of African Americans: new evidence from linked census data,” *American Economic Journal: Applied Economics*, 6 (1), 220–252.
- Corno, Lucia, Eliana La Ferrara, and Alessandra Voena (2020) “Female Genital Cutting and the Slave Trade,” CEPR Discussion Paper DP15577, CEPR, <https://ssrn.com/abstract=3753982>.
- Correia, Sergio (2015) “Singletons, Cluster-Robust Standard Errors and Fixed Effects: A Bad Mix,” working paper, Duke University, <http://scoreia.com/research/singletons.pdf>.
- Courtault, Jean-Michel, Bertrand Crettez, and Naila Hayek (2006) “Characterization of stochastic dominance for discrete random variable,” working paper, HAL open science, <https://shs.hal.science/halshs-00446413/document>.
- Currie, J et al. (2009) “Healthy, wealthy, and wise: socioeconomic status, poor health in childhood, and human capital development.,” *Journal of Economic Literature*, 47 (1), 87–117.
- Denic, Srdjan, Chris Frampton, and M Gary Nicholls (2007) “Risk of cancer in an inbred population,” *Cancer Detection and Prevention*, 31 (4), 263–269.
- Fernández, Raquel (2011) “Does culture matter?” *Handbook of social economics*, 1, 481–510.

- Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams (2021) “Place-based drivers of mortality: Evidence from migration,” *American Economic Review*, 111 (8), 2697–2735.
- Ghosh, Arkadev, Sam Il Myoung Hwang, and Munir Squires (2023) “Economic Consequences of Kinship: Evidence From U.S. Bans on Cousin Marriage,” *The Quarterly Journal of Economics*, 138 (4), 2559–2606.
- Gibson, Campbell and Kay Jung (2002) “Historical Census Statistics on Population Totals by Race, 1790 to 1990, and by Hispanic origin, 1970 to 1990, for the United States, Regions, Divisions, and States,” Working Paper POP-WP056, US Census Bureau.
- Gilani, Ghausia Masood, Shahid Kamal, and Syed Aamir Masood Gilani (2006) “Risk factors for breast cancer for women in Punjab, Pakistan: Results from a case-control study,” *Pakistan Journal of Statistics and Operation Research*, 17–26.
- Giuliano, Paola and Nathan Nunn (2021) “Understanding cultural persistence and change,” *The Review of Economic Studies*, 88 (4), 1541–1581.
- Grant, JC and AH Bittles (1997) “The comparative role of consanguinity in infant and childhood mortality in Pakistan,” *Annals of human genetics*, 61 (2), 143–149.
- Hacker, J David (2010) “Decennial life tables for the white population of the United States, 1790–1900,” *Historical methods*, 43 (2), 45–79.
- Haines, Michael R. and Inter-university Consortium for Political and Social Research (2010) “Historical, Demographic, Economic, and Social Data: The United States, 1790-2002,” 10.3886/ICPSR02896.v3.
- Helgason, Agnar, Snæbjörn Pálsson, Daníel F Gudbjartsson, Thórdur Kristjánsson, and Kári Stefánsson (2008) “An association between the kinship and fertility of human couples,” *Science*, 319 (5864), 813–816.

- Hosseini-Chavoshi, Meimanat, Mohammad J Abbasi-Shavazi, and Alan H Bittles (2014) “Consanguineous marriage, reproductive behaviour and postnatal mortality in contemporary Iran,” *Human heredity*, 77 (1-4), 16–25.
- Hussain, R and AH Bittles (1999) “Consanguineous marriage and differentials in age at marriage, contraceptive use and fertility in Pakistan,” *Journal of Biosocial Science*, 31 (1), 121–138.
- Hussain, Rafat and Alan H Bittles (2004) “Assessment of association between consanguinity and fertility in Asian populations,” *Journal of Health, Population and Nutrition*, 1–12.
- Hwang, Sam and Munir Squires (2023) “Linked Samples and Measurement Error in Historical US Census Data,” working paper, University of British Columbia.
- Kaplanis, Joanna, Assaf Gordon, Tal Shor et al. (2018) “Quantitative analysis of population-scale family trees with millions of relatives,” *Science*, 360 (6385), 171–175.
- Kreisman, Daniel and Jonathan Smith (2023) “Distinctively Black names and educational outcomes,” *Journal of Political Economy*, 131 (4), 877–897.
- Liede, Alexander, Imtiaz A. Malik, Zeba Aziz, Patricia de los Rios, Elaine Kwan, and Steven A. Narod (2002) “Contribution of BRCA1 and BRCA2 Mutations to Breast and Ovarian Cancer in Pakistan,” *American Journal of Human Genetics*, 71 (3), 595–606.
- Long, J Scott (1997) *Regression models for categorical and limited dependent variables*: Sage Publications, Inc.
- Lowes, Sara and Eduardo Montero (2021) “The legacy of colonial medicine in Central Africa,” *American Economic Review*, 111 (4), 1284–1314.

- Lu, Frances and Tom Vogl (2023) “Intergenerational persistence in child mortality,” *American Economic Review: Insights*, 5 (1), 93–109.
- McWhirter, Rebekah E, Ruth McQuillan, Elizabeth Visser, Carl Counsell, and James F Wilson (2012) “Genome-wide homozygosity and multiple sclerosis in Orkney and Shetland Islanders,” *European Journal of Human Genetics*, 20 (2), 198–202.
- Mobarak, A Mushfiq, Theresa Chaudhry, Julia Brown et al. (2019) “Estimating the health and socioeconomic effects of cousin marriage in South Asia,” *Journal of biosocial science*, 51 (3), 418–435.
- Mobarak, Ahmed Mushfiq, Randall Kuhn, and Christina Peters (2013) “Consanguinity and other marriage market effects of a wealth shock in Bangladesh,” *Demography*, 50 (5), 1845–1871.
- Ober, Carole, Terry Hyslop, and Walter W Hauck (1999) “Inbreeding effects on fertility in humans: evidence for reproductive compensation,” *The American Journal of Human Genetics*, 64 (1), 225–231.
- Paul, Diane B and Hamish G Spencer (2008) ““It’s ok, we’re not cousins by blood”: the cousin marriage controversy in historical perspective,” *PLoS Biology*, 6 (12), e320.
- Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley (2021) “Combining family history and machine learning to link historical records: The Census Tree data set,” *Explorations in Economic History*, 80, 101391.
- Rudan, Igor, D Rudan, H Campbell et al. (2003b) “Inbreeding and risk of late onset complex disease,” *Journal of Medical Genetics*, 40 (12), 925–932.
- Rudan, Igor, Nina Smolej-Narancic, Harry Campbell, Andrew Carothers, Alan Wright,

- Branka Janicijevic, and Pavao Rudan (2003a) “Inbreeding and the genetic complexity of human hypertension,” *Genetics*, 163 (3), 1011–1021.
- Ruggles, Steven, Sarah Flood, Matthew Sobek, Danika Brockman, Grace Cooper, Stephanie Richards, and Megan Schouweiler (2023) “Version 13.0 1 percent random sample of U.S. Federal Census, 1850-1930,” <https://doi.org/10.18128/D010.V13.0>.
- Saadat, Mostafa (2011) “Association between healthy life expectancy at birth and consanguineous marriages in 63 countries,” *Journal of Biosocial Science*, 43 (4), 475–480.
- Saggar, Anand K and Alan H Bittles (2008) “Consanguinity and child health,” *Paediatrics and Child Health*, 18 (5), 244–249.
- Stelter, Robert and Diego Alburez-Gutierrez (2022) “Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics,” *Proceedings of the National Academy of Sciences*, 119 (10), e2120455119.
- Strauss, John and Duncan Thomas (2007) “Health over the life course,” *Handbook of development economics*, 4, 3375–3474.
- The Church of Jesus Christ of Latter-Day Saints (2023) “The Importance of Family,” <https://www.churchofjesuschrist.org/comeuntochrist/article/importance-of-family>, Accessed: 2023-10-09.
- US Census Bureau (2021) “Historical Population Change Data (1910-2020),” April, <https://www.census.gov/data/tables/time-series/dec/popchange-data-text.html>, Accessed: 2023-10-13.
- Ward, Zachary (2022) “Internal Migration, Education, and Intergenerational Mobility Evidence from American History,” *Journal of Human Resources*, 57 (6), 1981–2011.

## A Online Appendix: Additional Tables and Figures

Table A.1: Construction of analysis sample

	(1) Total Dropped	(2) Percent Dropped	(3) Remaining Observations
Nonmissing Sex	60,501	0.15	40,514,188
Nonmissing Birth Year	581,967	1.44	39,932,221
Nonmissing Death Year	10,903,538	27.31	29,028,683
Nonmissing Great-Grandparents	21,618,119	74.47	7,410,564
Nonmissing Maternal Age at Birth	28,666	0.39	7,331,556
Longevity between 0 and 98	50,342	0.68	7,360,222
Birth Year between 1750 and 1920	665,908	9.08	6,665,648
Singletons	148,649	2.23	6,516,999

This table shows how we create our final sample of 6.5 million children from the data we scrape from FamilySearch profiles. Each row shows the number of observations we drop after keeping only those for which a specific variable is missing. Singletons are groups with only one observation. See Correia (2015) for a more detailed description.



Table A.2: Descriptive statistics

<b>Analysis sample:</b> Individuals with non-missing great-grandparents			
	(1) Parents are first cousins	(2) Non-cousin	(3) Difference
Longevity	54.91 [27.98]	58.00 [27.86]	-3.09 (0.00)
Parent Longevity	68.95 [11.99]	69.49 [11.69]	-0.54 (0.00)
Year of Birth	1,843.22 [32.04]	1,848.56 [33.83]	-5.34 (0.00)
Mother's Age at Birth	30.19 [7.02]	29.83 [6.98]	0.36 (0.00)
Female	0.47 [0.50]	0.47 [0.50]	-0.00 (0.00)
Number of brothers	3.56 [2.06]	3.62 [2.08]	-0.06 (0.00)
Number of sisters	3.13 [1.93]	3.20 [1.97]	-0.07 (0.00)
Birth order	3.81 [2.48]	3.86 [2.51]	-0.05 (0.00)
Observations	165,773	6,351,226	6,516,999
Percent	2.54	97.46	100

Observations are at the level of children. This table shows the mean of each variable we use in our preferred specification in table 2. Column (1) shows means for children whose parents are first cousins. Column (2) shows means for children whose parents are not first cousins. Column (3) shows the difference between columns (1) and (2). Variable descriptions are in appendix C. Standard deviations are in square brackets. Standard errors are in parentheses.

Table A.3: The effect of cousin marriage on offspring longevity - dropping death years ending in 0

	(1)	(2)	(3)
	Raw	Controls	Mother and father sibling fixed effects
<b>Panel A: Offspring longevity</b>			
Parents are first cousins	-3.08*** (0.10)	-2.60*** (0.10)	-2.86*** (0.37)
Control mean	58.19	58.19	58.19
Observations	5,725,696	5,725,696	5,725,696
<b>Panel B: Conditional on surviving to age 5</b>			
Parents are first cousins	-2.73*** (0.08)	-2.16*** (0.08)	-2.12*** (0.32)
Control mean	63.98	63.98	63.98
Observations	5,168,088	5,168,088	5,168,088
<b>Panel C: Conditional on surviving to age 20</b>			
Parents are first cousins	-2.36*** (0.07)	-1.83*** (0.07)	-1.53*** (0.29)
Control mean	66.85	66.85	66.85
Observations	4,877,622	4,877,622	4,877,622
Individual controls	No	Yes	Yes
Paternal FE	No	No	Yes
Maternal FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children, keeping only those whose death years do not end in 0. This table shows estimates for the coefficient  $\beta$  from equation (1) estimated using OLS. The coefficients in the first column are simply the difference in means between the children of first cousins and the children of non-first cousins. The second and third columns include controls described in table 2 and appendix C. In the third column we include mother's siblings fixed effects and father's siblings fixed effects as shown in Appendix Figure A.8. Standard errors are clustered at the level of the individual and their siblings.

Table A.4: The effect of cousin marriage on offspring longevity - non-parametric controls (fixed effects)

	(1)	(2)	(3)
	Raw	Controls	Mother and father sibling fixed effects
<b>Panel A: Offspring longevity</b>			
Parents are first cousins	-3.09*** (0.09)	-2.57*** (0.09)	-3.29*** (0.33)
Control mean	58.00	58.00	58.00
Observations	6,516,999	6,516,999	6,516,999
<b>Panel B: Conditional on surviving to age 5</b>			
Parents are first cousins	-2.75*** (0.07)	-2.15*** (0.07)	-2.22*** (0.29)
Control mean	63.71	63.71	63.71
Observations	5,894,444	5,894,444	5,894,444
<b>Panel C: Conditional on surviving to age 20</b>			
Parents are first cousins	-2.38*** (0.06)	-1.84*** (0.06)	-1.80*** (0.26)
Control mean	66.73	66.73	66.73
Observations	5,550,001	5,550,001	5,550,001
Individual controls	No	Yes	Yes
Paternal FE	No	No	Yes
Maternal FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. This table shows estimates for the coefficient  $\beta$  from equation (1) estimated using OLS. The coefficients in the first column are simply the difference in means between the children of first cousins and the children of non-first cousins. The second and third columns include controls described in table 2 and appendix C as fixed effects instead of as quadratic terms as usual. In the third column we include mother's siblings fixed effects and father's siblings fixed effects as shown in Appendix Figure A.8. Standard errors are clustered at the level of the individual and their siblings.

Table A.5: The effect of cousin marriage on offspring longevity - parent longevity controls

	(1)	(2)	(3)
	Raw	Controls	Mother and father sibling fixed effects
<b>Panel A: Offspring longevity</b>			
Parents are first cousins	-3.21*** (0.10)	-2.66*** (0.10)	-3.83*** (0.51)
Control mean	58.01	58.01	58.01
Observations	5,979,422	5,979,422	5,979,422
<b>Panel B: Conditional on surviving to age 5</b>			
Parents are first cousins	-2.81*** (0.08)	-2.18*** (0.08)	-2.33*** (0.44)
Control mean	63.72	63.72	63.72
Observations	5,409,263	5,409,263	5,409,263
<b>Panel C: Conditional on surviving to age 20</b>			
Parents are first cousins	-2.43*** (0.07)	-1.86*** (0.06)	-1.54*** (0.40)
Control mean	66.76	66.76	66.76
Observations	5,092,075	5,092,075	5,092,075
Individual controls	No	Yes	Yes
Parent longevity controls	No	Yes	Yes
Paternal FE	No	No	Yes
Maternal FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. We drop all observations for which we are unable to assign a value for their parents' longevity. This table shows estimates for the coefficient  $\beta$  from equation (1) estimated using OLS. The coefficients in the first column are simply the difference in means between the children of first cousins and the children of non-first cousins. The second and third columns control for birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, birth order, and mother and father's longevity. These are described in appendix C. In the third column we include mother's siblings fixed effects and father's siblings fixed effects as shown in Appendix Figure A.8. Standard errors are clustered at the level of the individual and their siblings.

Table A.6: The effect of cousin marriage on offspring longevity - county-decade FE

	(1)	(2)	(3)
	Raw	Controls	Mother and father sibling fixed effects
<b>Panel A: Offspring longevity</b>			
Parents are first cousins	-3.11*** (0.10)	-3.17*** (0.10)	-3.14*** (0.42)
Control mean	58.38	58.38	58.38
Observations	5,053,778	5,053,778	5,053,778
<b>Panel B: Conditional on surviving to age 5</b>			
Parents are first cousins	-2.24*** (0.08)	-2.24*** (0.08)	-2.06*** (0.36)
Control mean	64.12	64.12	64.12
Observations	4,564,513	4,564,513	4,564,513
<b>Panel C: Conditional on surviving to age 20</b>			
Parents are first cousins	-1.80*** (0.07)	-1.77*** (0.07)	-1.88*** (0.33)
Control mean	67.05	67.05	67.05
Observations	4,302,406	4,302,406	4,302,406
Individual controls	No	Yes	Yes
County by decade FE	No	Yes	Yes
Paternal FE	No	No	Yes
Maternal FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. We drop all observations without information on the child's county of birth. This table shows estimates for the coefficient  $\beta$  from equation (1) estimated using OLS. The coefficients in the first column are simply the difference in means between the children of first cousins and the children of non-first cousins. The second and third columns control for birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in appendix C. In the third column we include mother's siblings fixed effects and father's siblings fixed effects as shown in Appendix Figure A.8, and county-by-decade fixed effects. Standard errors are clustered at the level of the individual and their siblings.

Table A.7: The effect of cousin marriage on longevity, controlling for father's occupation categories

	(1)	(2)	(3)
Parents are first cousins	-3.82*** (0.60)	-3.85*** (0.60)	-3.87*** (0.60)
Control mean	59.21	59.21	59.21
Observations	138,047	138,047	138,047
Father's occupation	No	Yes	Yes
Father's age	No	No	Yes
Birth state FEs	Yes	Yes	Yes
Other controls	Yes	Yes	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. This table shows estimates for the coefficient  $\beta$  from equation (1) without parents' sibling fixed effects. In the second and third columns, we control for occupation categories of fathers. Each occupation category is defined by the first digit of the IPUMS code for the variable "occ1950". In the third column, we additionally control for the age of the father at which his occupation is measured. In all three specifications in this table, we control for the following covariates: birth state fixed effects, birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in appendix C. Standard errors are clustered at the level of the individual and their siblings. For this table, we use the subsample whose father is linked to one of the 1 percent samples of U.S. Federal Censuses from 1850 to 1930.

Table A.8: Mother's age at birth and siblings controls only

	(1) Raw	(2) Mother's age at birth controls	(3) Number of siblings controls	(4) Combined
Parents are first cousins	-3.09*** (0.09)	-3.07*** (0.09)	-3.12*** (0.09)	-3.10*** (0.09)
Control mean	58.00	58.00	58.00	58.00
Observations	6,516,999	6,516,999	6,516,999	6,516,999
Maternal age at birth controls	No	Yes	No	Yes
Number of siblings controls	No	No	Yes	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. This table shows estimates for the coefficient  $\beta$  from equation (1) estimated using OLS. The coefficients in the first column are simply the difference in means between the children of first cousins and the children of non-first cousins. The second column controls for the mother's age at birth. The third column controls for the number of brothers and sisters the individual has. The fourth column controls for mother's age at birth and the number of siblings. These are described in appendix C. Standard errors are clustered at the level of the individual and their siblings.

Table A.9: Cousin marriage and fertility

	(1)	(2)	(3)
	Raw	Controls	Same-sex sibling fixed effects
<b>Panel A: Number of children</b>			
Married to first cousin	-0.01 (0.02)	-0.06*** (0.02)	-0.19*** (0.02)
Control mean	5.25	5.25	5.25
Observations	1,287,986	1,287,986	1,287,986
<b>Panel B: Conditional on child surviving to age 5</b>			
Married to first cousin	-0.06*** (0.01)	-0.11*** (0.01)	-0.27*** (0.02)
Control mean	4.78	4.78	4.78
Observations	1,287,986	1,287,986	1,287,986
<b>Panel C: Conditional on child surviving to age 20</b>			
Married to first cousin	-0.10*** (0.01)	-0.15*** (0.01)	-0.31*** (0.02)
Control mean	4.50	4.50	4.50
Observations	1,287,986	1,287,986	1,287,986
Controls	No	Yes	Yes
Same-sex sibling FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of parents. The outcome is fertility, defined as the number of children an individual has between ages 12 and 50. Panel B counts only children surviving to age 5, and panel C counts only children surviving to age 20. The coefficients in the first column are simply the difference in means between those who marry their first cousins and those who do not. The second and third columns controls for birth year, sex, and maternal age at birth. In the third column we include wife's siblings fixed effects and husband's siblings fixed effects.



Table A.10: Cousin marriage and age at birth

	(1)	(2)	(3)
	Raw	Controls	Same-sex sibling fixed effects
<b>Panel A: Maternal age at birth (average)</b>			
Married to first cousin	0.31*** (0.03)	0.12*** (0.03)	0.07 (0.05)
Control mean	29.33	29.33	29.33
Observations	631,462	631,462	631,462
<b>Panel B: Maternal age at first birth</b>			
Married to first cousin	0.19*** (0.04)	0.10* (0.04)	0.13* (0.05)
Control mean	23.33	23.33	23.33
Observations	631,462	631,462	631,462
<b>Panel C: Maternal age at last birth</b>			
Married to first cousin	0.37*** (0.05)	0.06 (0.05)	-0.07 (0.07)
Control mean	36.02	36.02	36.02
Observations	631,423	631,423	631,423
Controls	No	Yes	Yes
Same-sex sibling FE	No	No	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of female parents (mothers). Outcomes are the ages at which the mother has children. Panel A is the average age the mother gives birth at. Panel B at which the mother gives birth for the first time, and panel C is the age at which the mother gives birth for the last time. The coefficients in the first column are simply the difference in means between those who marry their first cousins and those who do not. The second and third columns controls for birth year, sex, and maternal age at birth. In the third column we include wife's siblings fixed effects and husband's siblings fixed effects.

Table A.11: The effect of cousin marriage on longevity, with or without imputation of missing longevity

	(1) No imputation	(2) With imputation
Parents are first cousins	-3.27*** (0.33)	-3.88*** (0.35)
Control mean	58.00	47.05
Observations	6,516,999	8,167,207
Individual controls	Yes	Yes
Paternal FE	Yes	Yes
Maternal FE	Yes	Yes


\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. This table shows estimates for the coefficient  $\beta$  from equation (1) with or without imputing the missing longevity. No imputation was employed for column (1), and the sample for that column is the same as that used for our baseline estimate. For column (2), we augment our sample by including observations with missing death years. We imputed their lifespan to be zero. In both specifications in this table, we control for the parents' sibling fixed effects and the following covariates: birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in appendix C. Standard errors are clustered at the level of the individual and their siblings.

Figure A.1: Example FamilySearch Profile


## Family Members

Show All Family Members

### Spouses and Children




**William Washington Peacock**  
1822-1885 • LYSS-X9V




**Candace D. Holland**  
1825-1907 • 2459-464

**Marriage**  
1845  
Wayne, North Carolina, United States


**Children (10)**




**Jincy Caroline Peacock**  
1845-1879 • 9C63-24T




**Joseph Brantley Peacock**  
1849-1918 • K45W-RB8




**David L. Peacock**  
1850-1875 • MYRX-LMK




**William A. Peacock**  
1852-Deceased • 9C63-2HS




**John Coffee Peacock**  
1855-1902 • K8VW-5XL



**Paul C Peacock**  
1858-1862 • 9484-K8M




**Pacian Rebecca Peacock**  
1859-1930 • K85B-73Y




**Hannah Fannie M Peacock**  
1860-Deceased • KCJ7-S7D

### Parents and Siblings




**Elisha Holland**  
1764-1833 • M7HV-9G8




**Patience "Patie" Peacock**  
1788-1857 • LXMB-H4T

**Marriage**  
1806  
Wayne, North Carolina, United States


**Children (9)**




**Warren Holland**  
1805-1864 • LHJC-SL3




**Ave Nancy Holland**  
1809-1895 • K2MW-5VH




**Jincy Holland**  
1809-1895 • LCC2-1FK




**Exum Holland**  
1811-1880 • LHJC-3MM




**Jinnett Holland**  
1813-1880 • L6M5-M59



**West Holland**  
1820-1903 • KJWV-KVB



**Green Holland**  
1822-1886 • 2459-48P

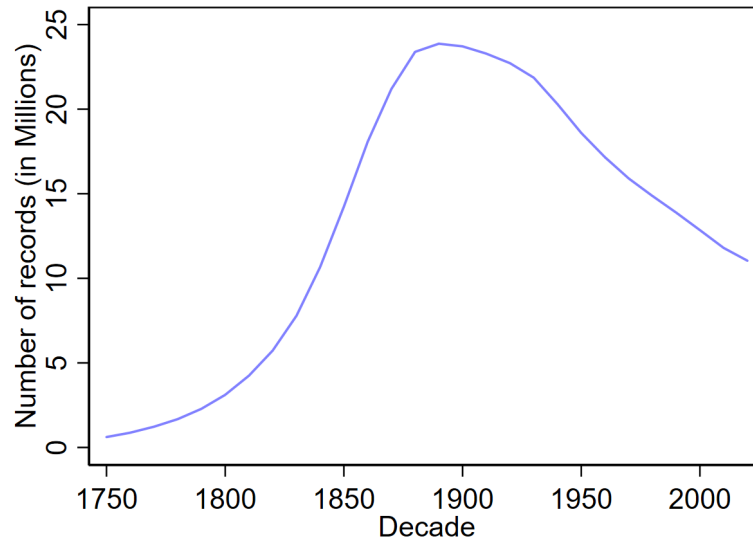


**Candace D. Holland**  
1825-1907 • 2459-464

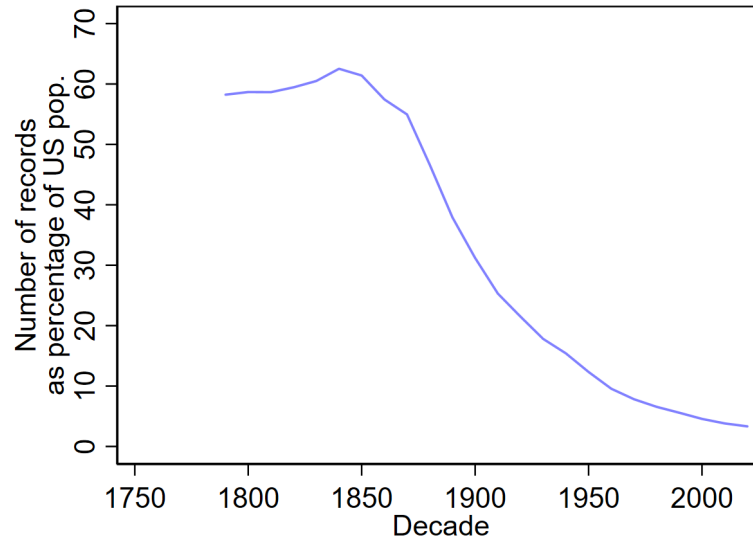
Note: This figure depicts a typical FamilySearch profile (that of Candace D. Holland). Not pictured are the individual's place of birth and date of birth.

43

Figure A.2: Record coverage



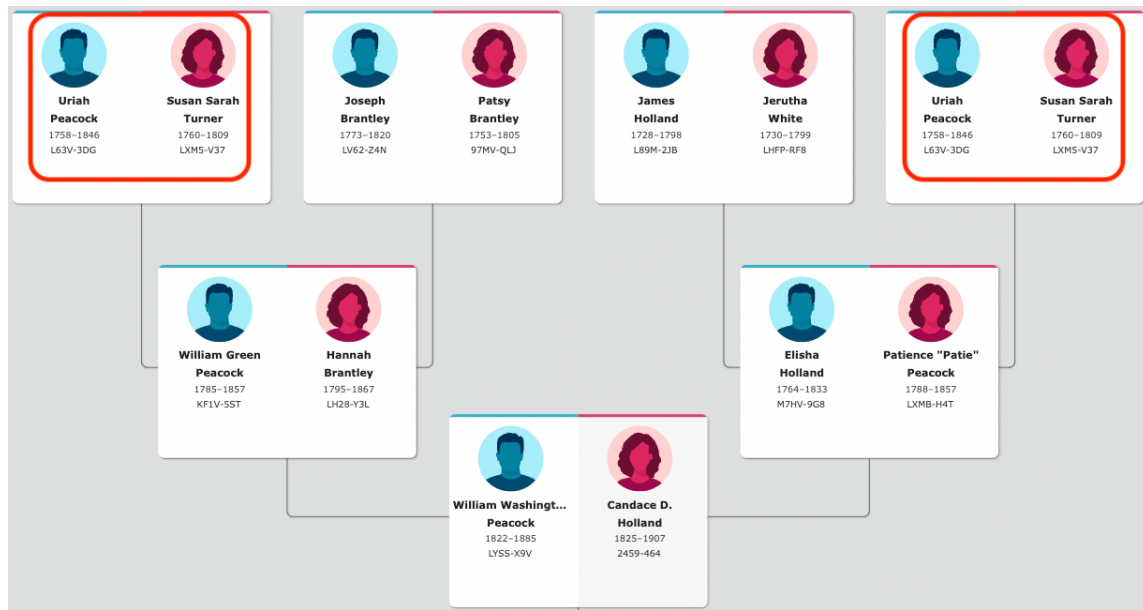
(a) Number of individuals in our dataset alive per decade



(b) Number of individuals alive as percentage of US population

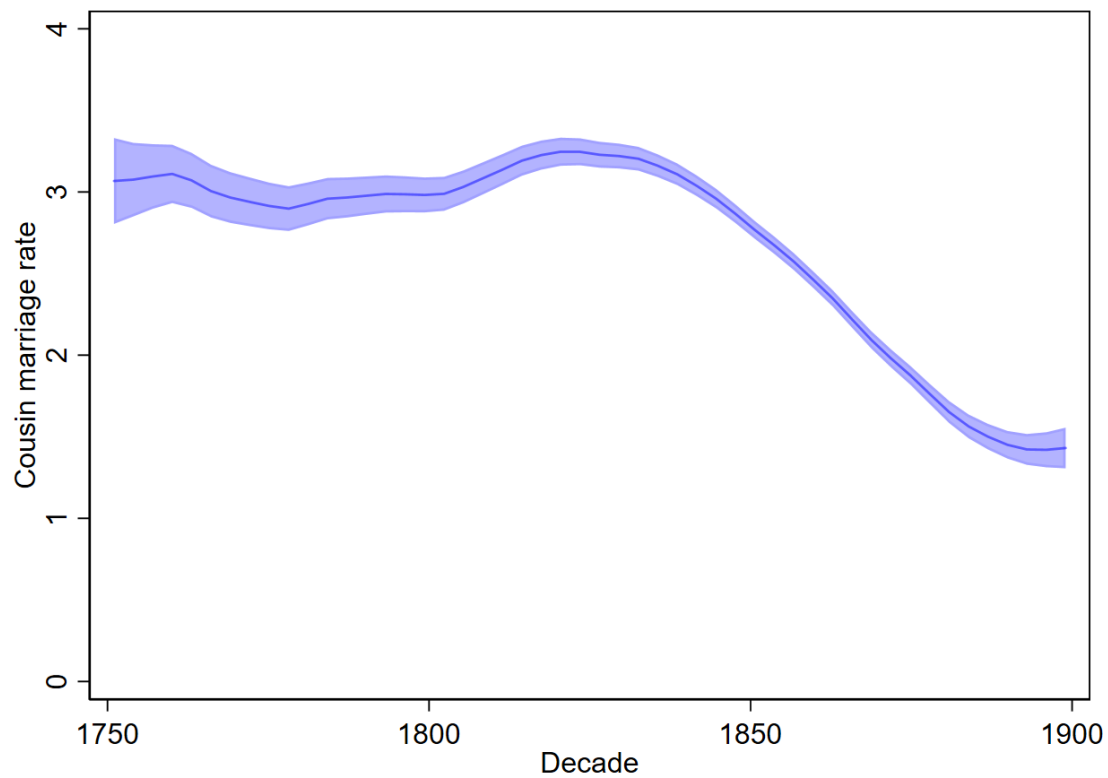
Note: Panel (a) shows the total number of records in our full dataset of 40 million individuals. An individual is counted if they were alive at any point in a given decade. Panel (b) shows these records as a percentage of the US population at the time. US population estimates come from US Census Bureau (2021) for years 2000-2020 and Gibson and Jung (2002) for all other years.

Figure A.3: Genealogical profile of first cousin spouses



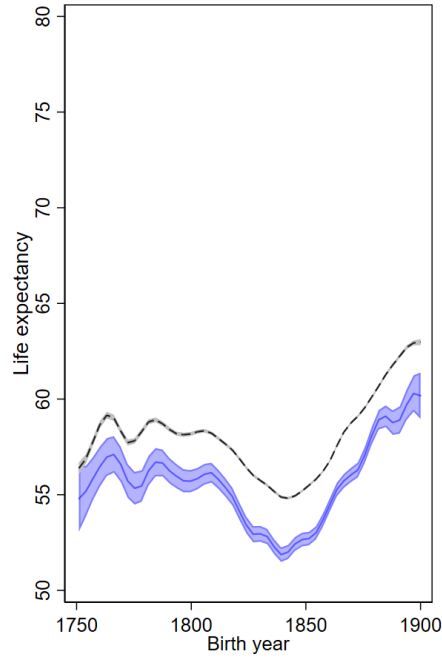
Note: This figure taken from FamilySearch shows the parents and grandparents of spouses (William and Candace) whose names and vital dates are in the bottom row. The husband's father and the wife's mother are siblings. This can be seen by observing the overlapping set of grandparents in the top row of profiles, highlighted in red.

Figure A.4: Cousin marriage rates over time

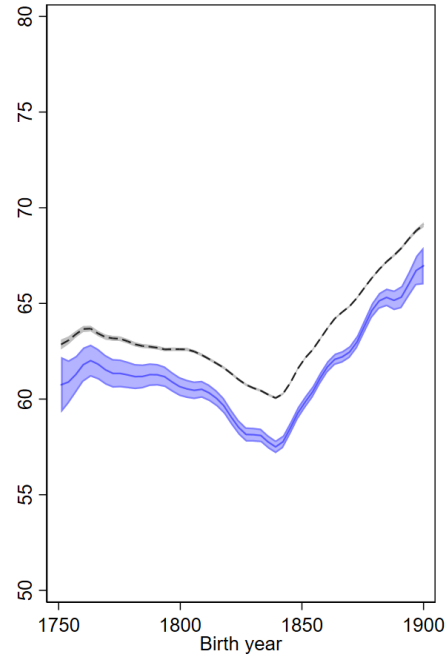


This figure depicts the share of marriages in our analysis sample of 6.5 million children that are between first cousins. As a proxy for year of marriage this figure uses the year of birth of the first child born of a given union. The rate is computed by taking the number of first-born children with first-cousin parents in a given decade divided by the total number of first-born children born in that decade.

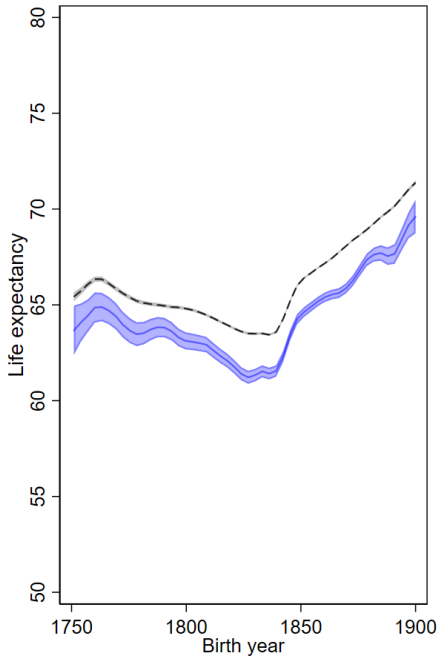
Figure A.5: Life expectancy by birth cohort (at birth and at age 5, 20, and 60)



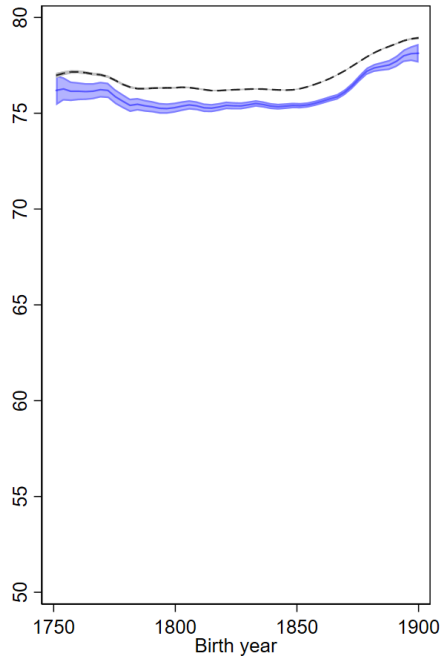
(a) At birth



(b) Conditional on surviving to age 5



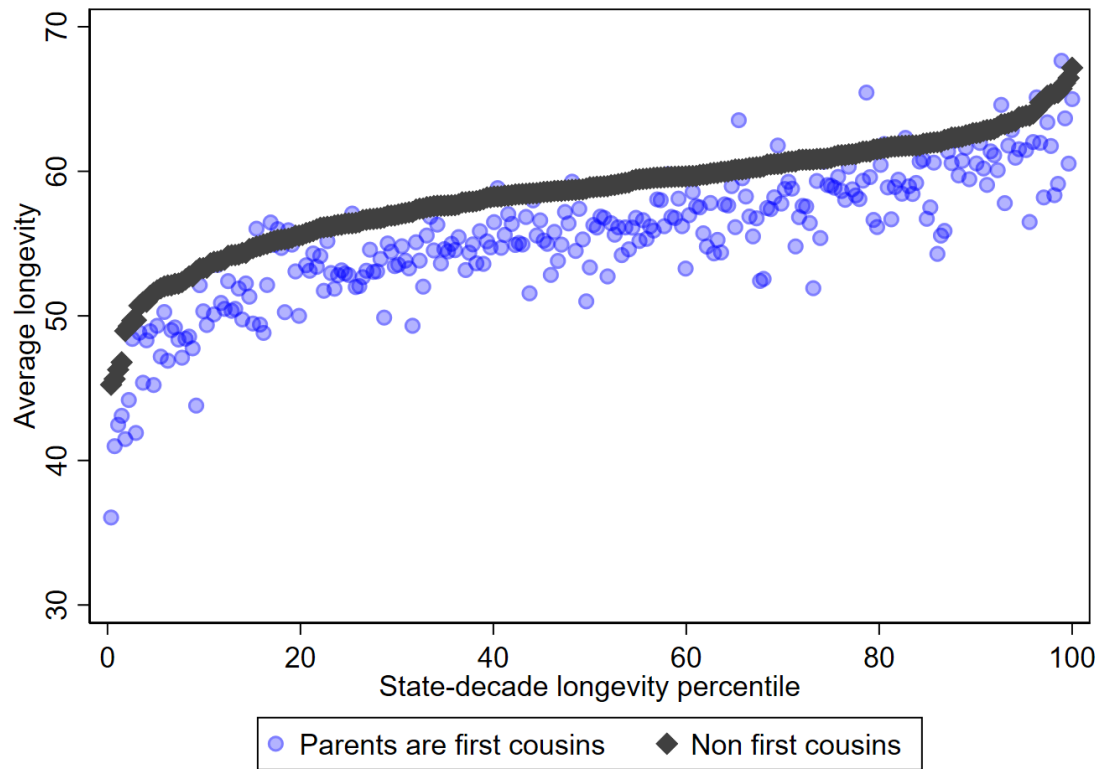
(c) Conditional on surviving to age 20



(d) Conditional on surviving to age 60

This figure depicts the life expectancy of our analysis sample of 6.5 million children conditional on surviving to a specified age. Children of first cousins are represented by the blue lines and children of non first cousins are represented by the gray lines. Panel (a) is a local polynomial regression of life expectancy at birth on birth year. Panels (b), (c), and (d) are local polynomial regressions of life expectancy at age 5, 20, and 60, respectively, on birth year.

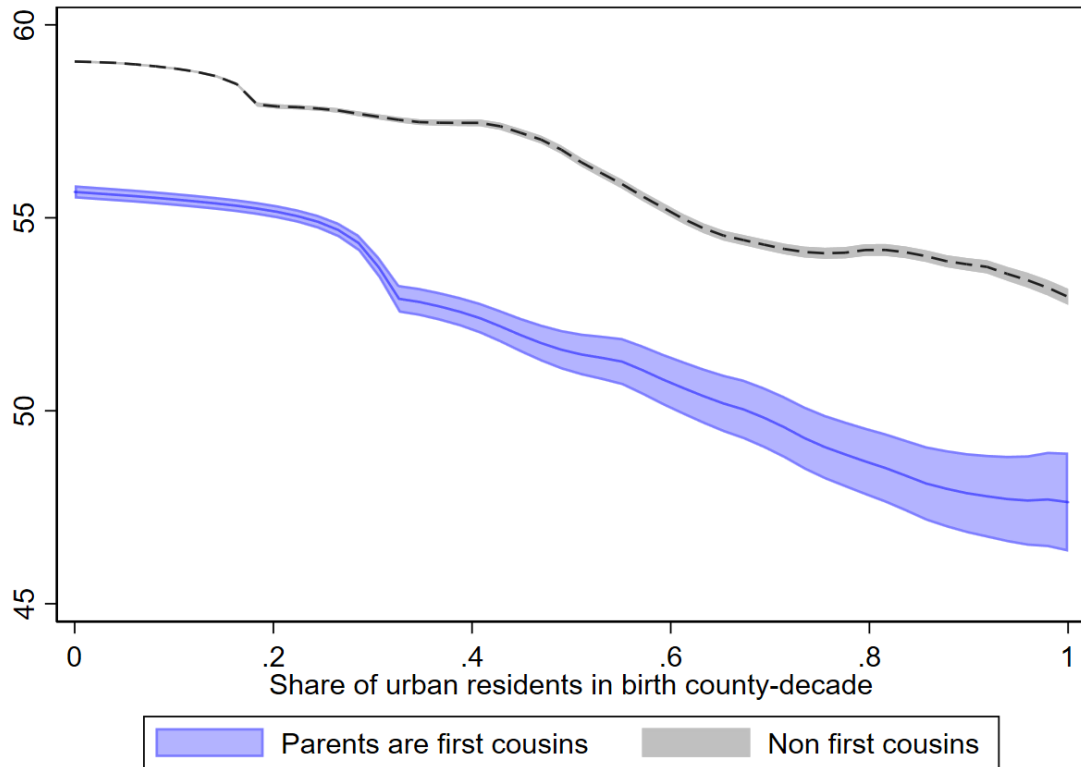
Figure A.6: Life expectancy by state-decade of birth



This figure depicts the average longevity by state of birth and decade (without controls) for 5.9 million children in our analysis sample for which state of birth is available. Each point is a state-decade pair. Data are sorted by state-decade longevity for individuals whose parents are not first cousins.

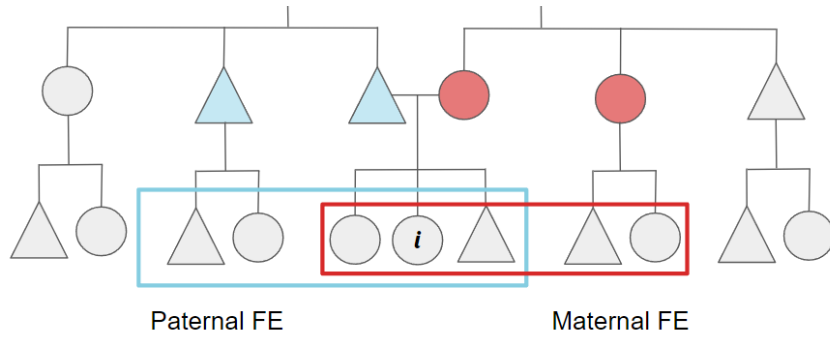


Figure A.7: Life expectancy and share of urban residents in birth county-decade



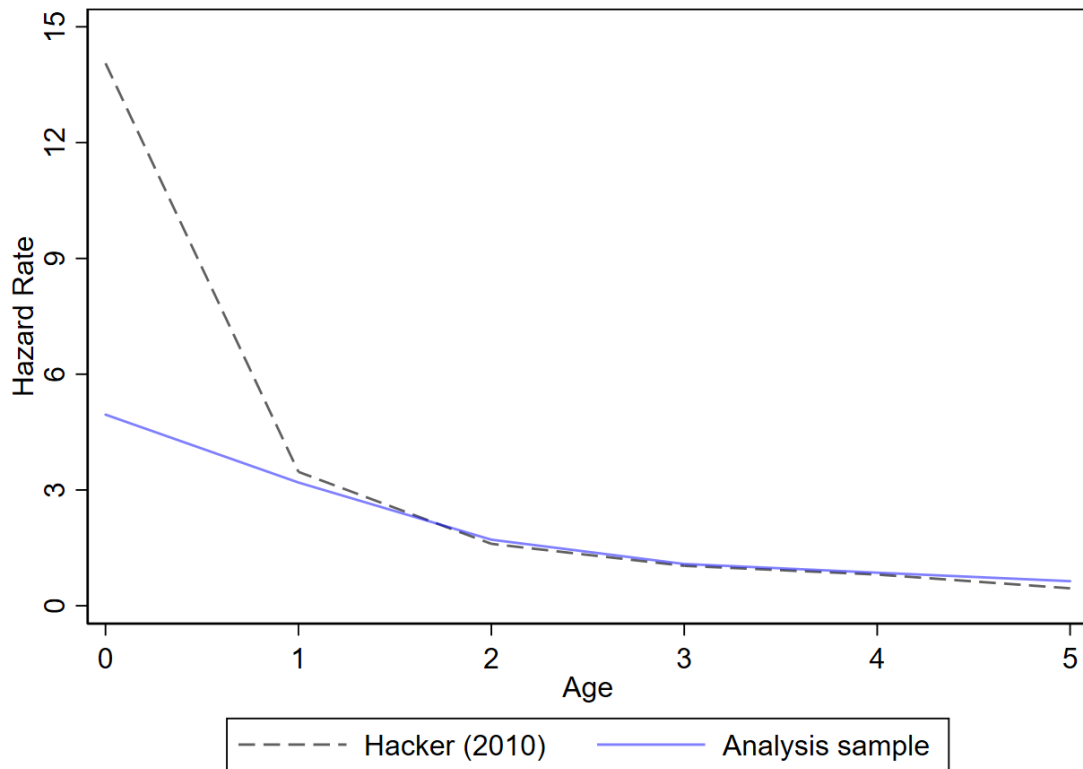
Note: This figure describes the correlation between life expectancy and the share of urban residents in one's birth county-decade. The sample for this figure consists of people in our analysis sample whose birth county is observed ( $N = 4,605,226$ ). The two curves shown in the figure are the kernel-weighted local polynomial fitted to the data, and the shaded areas are the 95th percentile confidence interval. The data on county-decade-level shares of urban residents come from Haines and Inter-university Consortium for Political and Social Research (2010).

Figure A.8: Empirical design



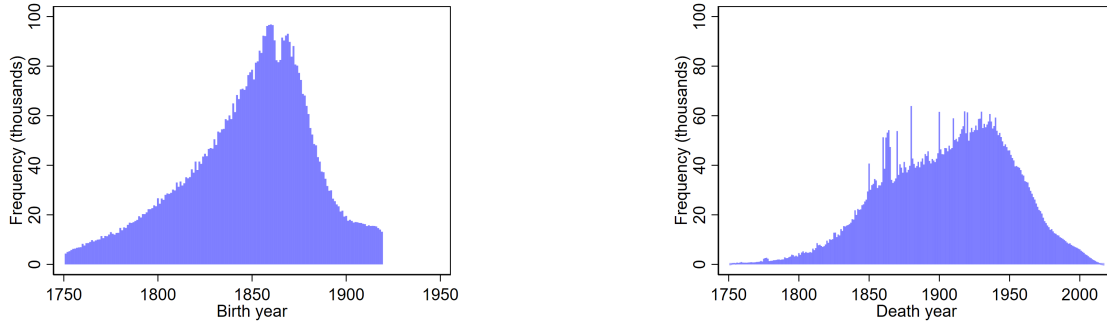
Notes: This figure visualizes the *Maternal* and *Paternal* fixed effects through two generations of related males (triangles) and females (circles). The bottom row represents the ‘offspring’ of married cousins or non-cousins, and represent the observations in our analysis. The blue and red rectangles represent the maternal and paternal fixed effects that apply to an individual  $i$ . These include the maternal and paternal (parallel) cousins of that focal individual  $i$ , corresponding to the children of their mother’s sisters (red) and their father’s brothers (blue).

Figure A.9: Mortality rates, ages 0-5

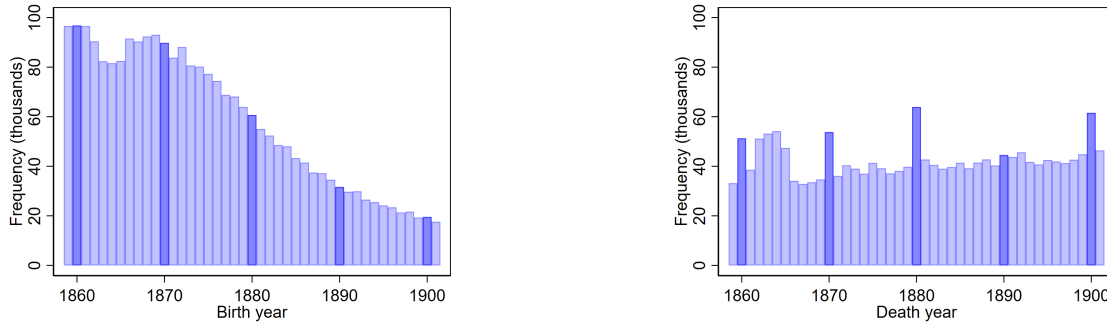


This figure depicts hazard rates for the male children in our data who died between 1880-1889. We define hazard as the percentage of individuals who die at a given age, conditional on surviving to that age. Historical longevity estimates depicted by the dashed line are from Table 8 of Hacker (2010). The paper argues that female data from this period are estimated with more error, so we use his measure for male mortality from 1880-89.

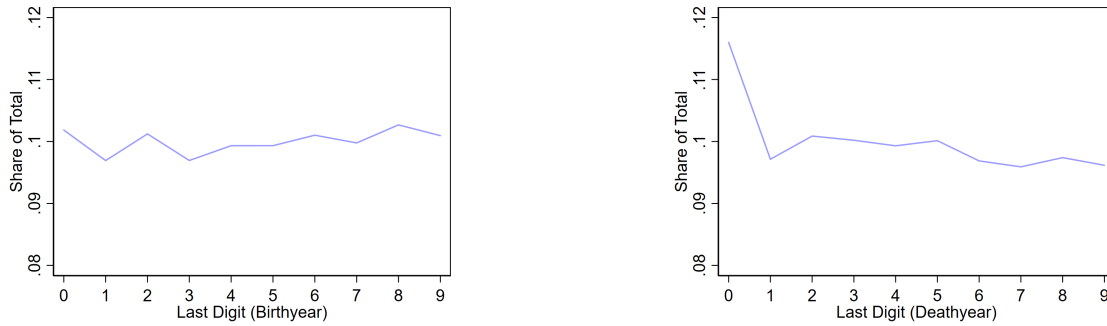
Figure A.10: Data quality: birth and death year heaping



(a) Distribution of vital years in analysis sample



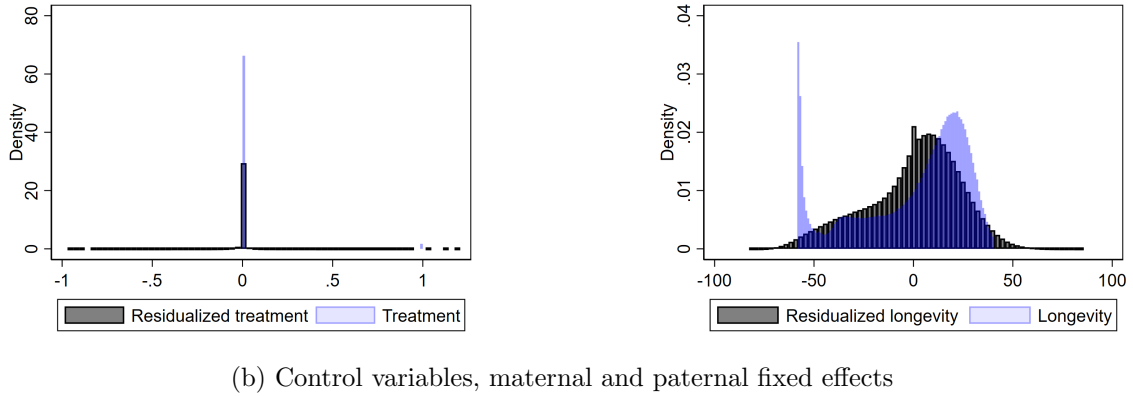
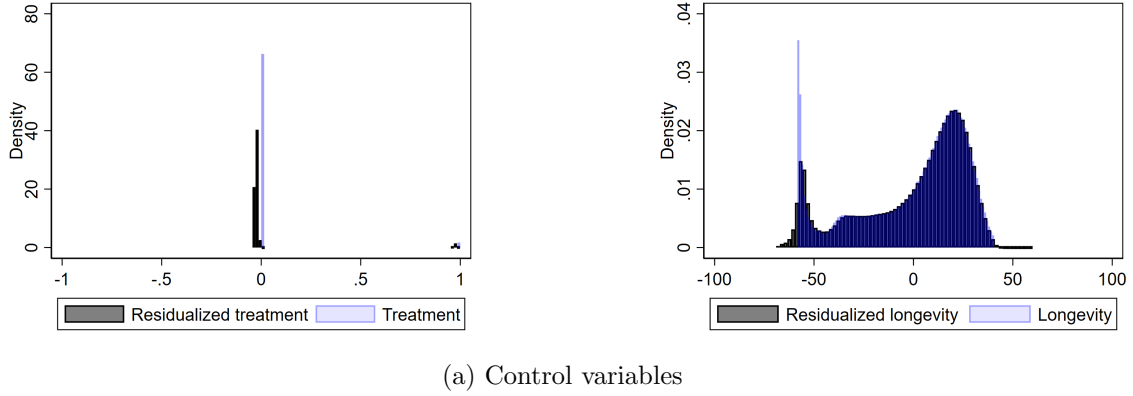
(b) Heaping in decadal census years



(c) Frequency of last digit in vital records

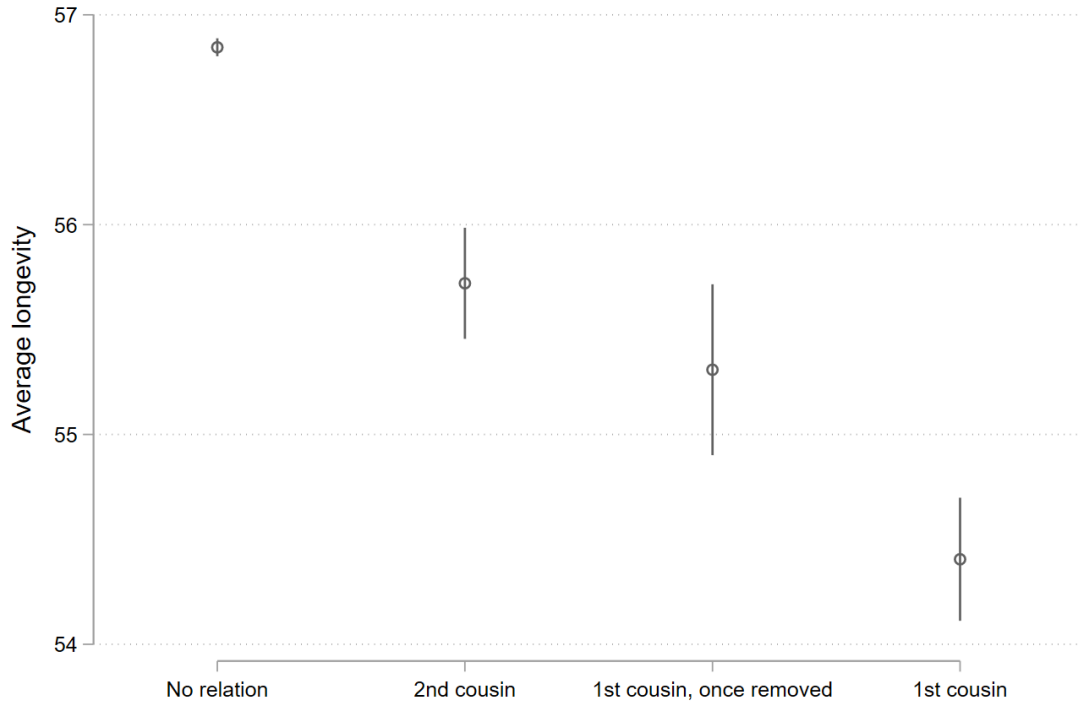
This figure describes age heaping in our analysis sample of 6.5 million individuals. Panels (a) and (b) depict the frequency (in thousands) of birth years and death years. Panels (c) and (d) depict particular segments of (a) and (b), respectively. Census years (ending in zero) are highlighted in a darker shade. Note that individual records for the 1890 census were lost in a fire and hence are not available. Panels (e) and (f) depict the percentage of birth years and death years ending in each digit.

Figure A.11: Residual variation in treatment and outcome of interest



This figure shows the distribution of the residuals in our analysis sample of 6.5 million individuals after regressing our treatment variable (an indicator for an individual's parents being first cousins or not), or our outcome variable (longevity) on a set of covariates. In panel (a) we regress these variables on our full set of controls (listed in table 2 and described in appendix C). In panel (b) we regress them on the full set of controls as well as maternal and paternal fixed effects.

Figure A.12: Average longevity by the degree of relatedness between spouses



Note: This figure shows how the degree of relatedness between spouses is correlated with the average longevity of their offspring. Parents in “No relation” group do not share any grandparents or great-grandparents; those in “2nd cousin” group share two great-grandparents; those in “1st cousin” group share two grandparents; and those in “1st cousin, once removed” are married to a child of their first cousin. We restrict our sample to those for whom we can determine whether they had a second-cousin marriage or not, i.e., those whose 30 ancestors (2 parents, 4 grandparents, 8 great-grandparents, and 16 great-great-grandparents) are identified in our dataset. We estimated our baseline model (1) without the sibling fixed effects but with three indicators for each type of marriages (i.e., 2nd cousin, 1st cousin once removed, and 1st cousin). We added to the coefficients on these indicators the average longevity of “No relation” group to make four groups comparable. N=1,884,966.

## B Online Appendix: Data

### B.1 Genealogical data coverage and selection

While our data includes a substantial proportion of the US population alive in the nineteenth century, selection into sample is not random or representative. This appendix describes how our dataset was compiled, the scope of coverage of our data, and characterizes selection into the sample.

Our dataset of 40 million genealogical profiles comes from the following four steps.

1. We obtained all US marriage records available on FamilySearch.org from the mid-eighteenth to mid-nineteenth centuries.<sup>25</sup> These married couples form the nucleus of our sample.
2. We scraped the family tree of every husband and wife in these marriage records who had a genealogical profile on FamilySearch.
3. We scraped the siblings of these married couples and their siblings' spouses.
4. We scraped the profiles of the parents of everyone so far in our sample (the original married couples, their siblings, and the spouses of their siblings).

The 40 million profiles in our final dataset includes the individuals whose genealogical profiles we explicitly scraped and also their close relatives: their parents, siblings, spouse, and children. The following variable are (usually) available for all such individuals: name, sex, year of birth, and year of death. See Appendix Figures A.1 for an example of a profile and what information is available on their close relatives. While we do not explicitly scrape the profiles of the offspring who are the focus of our analysis (our ‘analysis sample’), their names and vital dates are included in the profiles of their parents. Since vital dates of parents and children are included in this way in our data, we can study four linked

---

<sup>25</sup>Specifically, all marriage records from 1750 to 1858. The latter date was chosen in function of the identification strategy we used in Ghosh, Hwang and Squires (2023).

generations by scraping the profiles of only two generations (the spouses in our original dataset and their parents).

To visualize the coverage of our genealogical data, we show the number of records for which we have longevity information in Figure A.2. Panel (a) shows how many individuals in our dataset were alive in a given decade. This peaks at about 23 million in the late nineteenth century. Panel (b) shows the number of records we have relative to the US population in each decade (as per the census).<sup>26</sup> Our coverage as a percent of the population peaks at about 60% in the early-mid nineteenth century.

While broad in scope, our dataset of 40 million profiles is not representative of the US population. Crowd-sourced genealogical data typically under-represent non-whites, unmarried adults, and those with lower levels of education (see Hwang and Squires, 2023). Because our sample extends outwards from spouses for whom marriage records exist, our dataset is likely to be even more skewed towards married individuals and whites (marriage records from the mid-eighteenth to mid-nineteenth centuries include few black spouses).

However, given the relatively sparse information we have on each individual in our sample, we cannot speak more precisely to how representative our data are to the US population at the time. Genealogical profiles do not include race, occupation, education, or other similar characteristics.

For a limited subset of our data, however, we can link genealogical profiles to census individuals, and use these to speak to the differences between our sample and the broader US population. To do so, we link as many genealogical profiles as we can from our analysis sample to the IPUMS 1% 1850 census sample. (See Hwang and Squires (2023) for a description of how we do this.) We compare the census individuals who can link to those we cannot in Appendix Table B.12.

---

<sup>26</sup>This sub-figure starts in 1790, the year of the first Federal US census.



As expected, we find large differences across most characteristics. The census individuals we can link to our genealogical records (analysis sample) are more likely to be literate, to be white and native-born, and to be related to the head of the household. They are slightly more likely to be in a white-collar occupation, substantially more likely to be in a farming occupation.

Table B.12: Selection into sample: IPUMS 1% 1850 census sample linked to genealogical dataset

	(1) Census	(2) Linked	(3) Diff.	(4) Std. diff.
<b>Panel A: Socio-demographics</b>				
Can read & write	0.891 (0.312)	0.971 (0.167)	0.059*** (0.006)	0.226
Non-white	0.022 (0.146)	0.000 (0.010)	-0.021*** (0.001)	-0.149
Foreign-born	0.113 (0.316)	0.003 (0.052)	-0.087*** (0.003)	-0.343
Related to head	0.907 (0.290)	0.989 (0.103)	0.067*** (0.003)	0.266
<b>Panel B: Occupation</b>				
White-collar	0.102 (0.302)	0.114 (0.318)	0.014** (0.006)	0.030
Farmer	0.452 (0.498)	0.594 (0.491)	0.126*** (0.009)	0.203
Skilled	0.260 (0.438)	0.192 (0.394)	-0.061*** (0.008)	-0.114
Unskilled	0.187 (0.390)	0.099 (0.298)	-0.080*** (0.007)	-0.179
<b>Panel C: Census region of residence</b>				
Northeast	0.429 (0.495)	0.440 (0.496)	0.030*** (0.005)	0.017
Midwest	0.266 (0.442)	0.296 (0.457)	0.020*** (0.005)	0.048
South	0.297 (0.457)	0.259 (0.438)	-0.046*** (0.005)	-0.060
West	0.009 (0.093)	0.005 (0.068)	-0.003*** (0.001)	-0.036
Observations	197,796	9,932	207,728	

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . This table compares the 1% random sample of the 1850 census (column (1)) to its subsample that is linked to our analysis sample (column (2)). The 1850 census is chosen because the linkage rate is highest for that year. Column (3) presents differences, after adjusting for sex and birth year fixed effects. Column (4) presents standardized differences: a difference less than 0.1, the threshold recommended in Austin (2009), suggests balance.

## C Online Appendix: Variables used in analysis

### C.1 Offspring-level analysis (‘analysis sample’)

#### Treatment variable

- *Parents are first cousins*: This variable is equal to 1 if exactly one of the parents’ sets of grandparents (couples) match. It is equal to 0 if none of the parents’ sets of grandparents overlap. This variable is only defined for individuals where both parents have complete family trees up to the level of their grandparents.

#### Outcome variable

- *Longevity*: The year the individual died in minus the year the individual was born in. This number is changed to missing if it is below 0 or greater than 98 (the 99th percentile of longevity in our sample).

#### Control variables

- *Birth year*: The year in which individual  $i$  was born. In our baseline set of controls, we also include the square of this variable to capture non-linear effects.
- *Sex*: A binary variable equal to 1 if individual  $i$  is recorded as female in their genealogical profile, and equal to 0 otherwise.
- *Maternal age at birth*: The year in which the individual was born minus the year in which the individual’s mother was born. This number is considered missing if it is less than 12 or greater than 50. In our baseline set of controls, we also include the square of this variable to capture non-linear effects.
- *Number of sisters*: The total number of female children born to the two parents of individual  $i$ . If a set of parents has more than 10 female children, this number is

topcoded to 10. In our baseline set of controls, we also include the square of this variable to capture non-linear effects.

- *Number of brothers*: The total number of male children born to the two parents of individual  $i$ . If a set of parents has more than 10 male children, this number is topcoded to 10. In our baseline set of controls, we also include the square of this variable to capture non-linear effects.
- *Sex ratio*: The sex ratio of the children born to the two parents of individual  $i$ , constructed by taking the number of female children (“sisters”) divided by the total number of children (“sisters” + “brothers”). In our baseline set of controls, we also include the square of this variable to capture non-linear effects.
- *Birth order*: The order in which individual  $i$  was born relative to their siblings. If there are more than 10 children born to a set of two parents, this number is topcoded to 10. In our baseline set of controls, we also include the square of this variable to capture non-linear effects.
- *Parent longevity*: The mean of the longevity of the mother and father of individual  $i$ .

## C.2 Parent-level analysis

### Treatment variable

- *Married to first cousin*: This variable is equal to 1 if this individual and their spouse have two grandparents in common, equal to 0 if they have no grandparents in common, and is considered missing otherwise. This variable is only defined for individuals for whom we have information on all four of their grandparents and well as their spouse’s grandparents. Individuals with offspring from more than one spouse take on the max of this value across their spousal links.

### Outcome variable

- *Longevity*: The year the parent died in minus the year the individual was born in. This number is changed to missing if it is below 0 or greater than 98 (the 99th percentile of longevity in our sample).
- *Fertility*: The total number of children born to that parent while they themselves were between ages 12 and 50, inclusive.

### **Control variables**

These are identical to the control variables listed in the preceding sub-section, though they are defined relative to a parent rather than their offspring.

## D Online Appendix: Parent-level analysis

Our baseline specification uses offspring of cousin or non-cousin marriages as observations. Our parent-level analysis instead uses the spouses (the parents of the offspring) as units of observation. Each observation is one parent, and hence each couple is in the data twice. Table D.13 presents descriptive statistics on the parent-level observations which we use for the analysis described in this appendix.

As far as possible, this analysis replicates the offspring-level analysis. The key difference is that the fixed effects are now only for the parent’s own siblings, and not the siblings of their spouse.

We use the following specification:

$$Longevity_j^{Parental} = \gamma FirstCousinSpouse_i + \eta_s SameSexSiblings_s + \mathbf{W}_j' \boldsymbol{\phi} + \varepsilon_j, \quad (2)$$

Here,  $j$  refers to a parent of one of the individuals in our main analysis sample. Subscript  $s$  refers to a set of same-sex siblings (brothers for male parents, sisters for female). *FirstCousinSpouse* is equal to 1 if that parent’s spouse is their first cousin, and 0 otherwise. Parents who had children with more than one spouse are treated as having married a cousin if any one of those spouses was a cousin. The set of controls  $\mathbf{W}_j$  consists of birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in Appendix C. Since treatment is at the individual level, we use robust standard errors rather than clustering them. Standard errors change very little when clustering at the level of siblings.

Table D.13: Descriptive statistics - parents

<b>Parent sample:</b> Parents of analysis sample			
	(1) Married to first cousins	(2) Not married to first cousins	(3) Difference
Longevity	67.14 [17.14]	67.75 [17.06]	-0.61 (0.00)
Year of birth	1,813.24 [25.63]	1,817.87 [27.65]	-4.63 (0.00)
Mother's age at birth	29.40 [6.87]	30.08 [7.24]	-0.69 (0.00)
Age at first child	23.52 [5.02]	23.33 [4.93]	0.19 (0.00)
Age at last child	36.39 [6.78]	36.02 [6.83]	0.37 (0.00)
Number of children	5.24 [2.99]	5.25 [3.02]	-0.01 (0.00)
Female	0.48 [0.50]	0.49 [0.50]	-0.01 (0.00)
Number of brothers	2.44 [1.67]	2.26 [1.60]	0.19 (0.00)
Number of sisters	2.21 [1.52]	2.14 [1.49]	0.07 (0.00)
Birth order	2.73 [1.72]	2.73 [1.71]	-0.00 (0.00)
Observations	35,992	1,251,994	1,287,986
Percent	2.79	97.21	100

Observations are at the level of parents. This table shows the mean of each variable we use in our preferred specification in table 2. Column (1) shows means for parents who are married to their first cousins. Column (2) shows means for parents who are not married to their first cousins. Column (3) shows the difference between columns (1) and (2). Variable descriptions are in appendix C. Standard deviations are in square brackets. Standard errors are in parentheses.

## E Online Appendix: Under-reporting of infant deaths

Our analysis sample consists of people who are registered on their parents’ genealogical profiles on FamilySearch. Among these people, we also restrict our sample to those whose birth and death years are non-missing, since we need these vital years to construct the main outcome variable of the study (i.e., longevity). This selection into sample driven by having a profile and having both birth and death years may bias our main results. In this section, we show that we are almost certainly underestimating the reduction in life expectancy from first-cousin marriage.

We first introduce some notation to simplify the exposition. Let  $\ell_i$  denote the longevity of person  $i$ .  $F_i \in \{0, 1\}$  denotes whether person  $i$  is included in our sample. That is,  $F_i = 1$  indicates that person  $i$  is registered on his/her parents’ FamilySearch profile and his/her birth and death years are not missing; and 0 otherwise.

With this notation, what we observe in our data can be represented as follows:

$$\mathbb{E}[\ell_i | F_i = 1]$$

This conditional mean may not be equal to the unconditional mean, i.e.,  $\mathbb{E}[\ell_i]$ . The unconditional mean corresponds to the mean longevity of a child born to a mother who has a FamilySearch profile, regardless of whether the child is included in our sample or not. The conditional mean and the unconditional mean may be different if those excluded from our sample are systematically different from those who are included. Based on previous works studying genealogical data (Hacker, 2010; Chong et al., 2022), we expect that those excluded from our sample are more likely to be infants or children who died too young to have records of their birth years and/or death years.

By the Law of Iterated Expectation, we obtain the following relationship between the



unconditional mean longevity and the conditional mean longevity.

$$\begin{aligned}
& \mathbb{E}[\ell_i] \\
&= \mathbb{E}[\mathbb{E}[\ell_i|F_i]] \\
&= \Pr(F_i = 1) \cdot \mathbb{E}[\ell_i|F_i = 1] + \Pr(F_i = 0) \cdot \mathbb{E}[\ell_i|F_i = 0]
\end{aligned} \tag{3}$$

Among the terms in the right-hand side of (3), we do not observe two of them: the probability that a child is included in our sample, i.e.,  $\Pr(F_i = 1)$ , and the mean longevity conditional on being excluded from our sample, i.e.,  $\mathbb{E}[\ell_i|F_i = 0]$ . We make an assumption about the former below.

**Assumption 1.** *A child whose parents have a FamilySearch profile is always registered on the parents' profile, but he/she may not have birth or death year recorded on the profile.*

Under Assumption 1, a person whose parent has a FamilySearch profile is excluded from our sample if and only if his/her birth or death year is missing. This allows us to estimate  $\Pr(F_i = 1)$  as follows:

$$\Pr(\widehat{F_i} = 1) = \frac{\# \text{ with non-missing birth and death yrs.}}{(\# \text{ with non-missing birth and death yrs.}) + (\# \text{ missing birth or death yrs.})} \tag{4}$$

To validate Assumption 1, we use a subsample that is linked to the one percent random sample of 1900 and 1910 censuses from our previous project (Hwang and Squires, 2023). In the 1900 and 1910 censuses, all women were asked about the number of children that they ever gave birth to, including those who died as infants. Therefore, for the mothers in the linked subsample, we can compare the number of children registered on their FamilySearch profile with the number of children ever born to them. If the two numbers agree, then that would suggest that Assumption 1 is a reasonable approximation to reality.

We indeed find support for Assumption 1 from this linked dataset. The total number of children that are registered on the FamilySearch profiles of mothers in the linked subsample is equal to 32,060, which is 99.6 percent of the total number of children ever born to them (32,177) according to the censuses.<sup>27</sup> A similar pattern is observed when we condition on the cousin-marriage status of mothers. The total number of children registered on the profiles of mothers who had a cousin marriage is 101.5 percent of the total number of children ever born to these mothers. The corresponding statistic for mothers who did not marry a cousin is 99.6 percent.

Under Assumption 1, we can compute the probability that a person is included in our sample, i.e.,  $\Pr(F_i = 1)$ , according to the formula (4). It is equal to 0.804, indicating that 19.6 percent of the people ever born to parents with a FamilySearch profile are excluded from our sample due to missing birth or death years. Among those who are missing vital years, the vast majority of them (95.4 percent) are missing death years only.

Now we can express the unconditional mean longevity ( $\mathbb{E}[\ell_i]$ ) as a linear function of the mean longevity conditional on being excluded from our sample ( $\mathbb{E}[\ell_i|F_i = 0]$ ) as follows:

$$\mathbb{E}[\ell_i] = \underbrace{\Pr(\widehat{F_i} = 0)}_{=0.196} \cdot \mathbb{E}[\ell_i|F_i = 0] + \underbrace{\Pr(\widehat{F_i} = 1)}_{=0.804} \cdot \underbrace{\mathbb{E}[\widehat{\ell_i}|\widehat{F_i} = 1]}_{=57.9} \quad (5)$$

$\approx 46.6$

where the hat ( $\wedge$ ) denotes an estimate obtained from our data.

We can also condition each term in the equality (5) on the cousin-marriage status of one's parents. Let  $C_i$  denote the cousin-marriage status of person  $i$ 's parents, with  $C_i = 1$  corresponding to a cousin-marriage, and  $C_i = 0$  otherwise. Then the mean longevity of

---

<sup>27</sup>At the individual level, the average ratio of the number of children registered on one's FamilySearch profile to the number of children ever born to her is 1.12. The fact that it is over 1 is partly driven by outliers. With the 90 percent Winsorization, the average ratio goes down to 1.01 with a standard deviation of 0.24. The discrepancy may also be driven by measurement errors in the census variables (Hwang and Squires, 2023).

those born to a cousin couple is as follows:

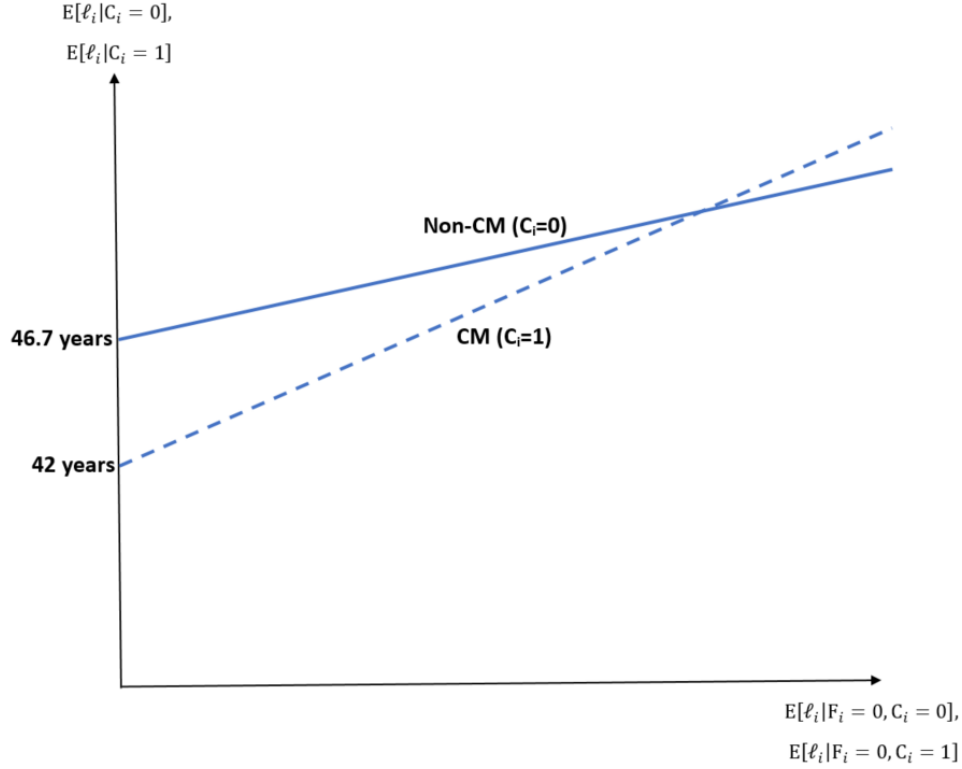
$$\begin{aligned}
& \mathbb{E}[\ell_i | C_i = 1] \\
&= \underbrace{\Pr(\widehat{F_i = 0} | C_i = 1)}_{\approx 0.235} \cdot \mathbb{E}[\ell_i | F_i = 0, C_i = 1] + \underbrace{\Pr(\widehat{F_i = 1} | C_i = 1)}_{\approx 0.765} \cdot \underbrace{\mathbb{E}[\ell_i | \widehat{F_i = 1}, C_i = 1]}_{\approx 54.9} \quad (6) \\
& \hspace{15em} \underbrace{\hspace{10em}}_{\approx 42.0}
\end{aligned}$$

Similarly, the mean longevity of those born to a non-cousin couple is equal to the following:

$$\begin{aligned}
& \mathbb{E}[\ell_i | C_i = 0] \\
&= \underbrace{\Pr(\widehat{F_i = 0} | C_i = 0)}_{\approx 0.195} \cdot \mathbb{E}[\ell_i | F_i = 0, C_i = 0] + \underbrace{\Pr(\widehat{F_i = 1} | C_i = 0)}_{\approx 0.805} \cdot \underbrace{\mathbb{E}[\ell_i | \widehat{F_i = 1}, C_i = 0]}_{\approx 58.0} \quad (7) \\
& \hspace{15em} \underbrace{\hspace{10em}}_{\approx 46.7}
\end{aligned}$$

We graphically represent the linear relationship between  $\mathbb{E}[\ell_i | C_i]$  and  $\mathbb{E}[\ell_i | F_i = 0, C_i]$  in Figure E.13. The solid line represents  $\mathbb{E}[\ell_i | C_i = 0]$  and the dashed line corresponds to  $\mathbb{E}[\ell_i | C_i = 1]$ , respectively. The two lines are upward-sloping because the mean longevity increases as the conditional mean longevity, i.e.,  $\mathbb{E}[\ell_i | F_i = 0, C_i]$ , rises. The dashed line (corresponding to  $\mathbb{E}[\ell_i | C_i = 1]$ ) is steeper than the solid line (corresponding to  $\mathbb{E}[\ell_i | C_i = 0]$ ). The difference in the slope is approximately 0.04, and it is statistically significant at the 1 percent level.

Figure E.13: The linear relationship between the mean longevity conditional on being excluded from our sample and the unconditional mean longevity



Note: This figure depicts the linear relationship between the mean longevity conditional on being excluded from our sample and the unconditional mean longevity. The horizontal axis measures the conditional mean longevity and the vertical axis measures the unconditional mean. The line labeled “Non-CM ( $C_i = 0$ )” corresponds to the offspring of non-cousin couples, and the one labeled “CM ( $C_i = 1$ )” corresponds to the offspring of cousin couples. The slope of the former is 0.195, and that of the latter is 0.235, with the difference being statistically significant at the 1 percent level.

The effect of cousin marriage on the offspring longevity depends on the mean longevity of those excluded from our sample, i.e.,  $\mathbb{E}[\ell_i | F_i = 0, C_i = 0]$  and  $\mathbb{E}[\ell_i | F_i = 0, C_i = 1]$ . By construction, these two conditional means are unobservable to researcher. Whatever the values of these unobservable moments are, there can be three cases:

**Case 1.**  $\mathbb{E}[\ell_i | F_i = 0, C_i = 0] = \mathbb{E}[\ell_i | F_i = 0, C_i = 1]$

**Case 2.**  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] > \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$

**Case 3.**  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] < \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$

We discuss each of these three cases in turn below.

**Case 1.** Suppose that the two conditional means are equal, that is,

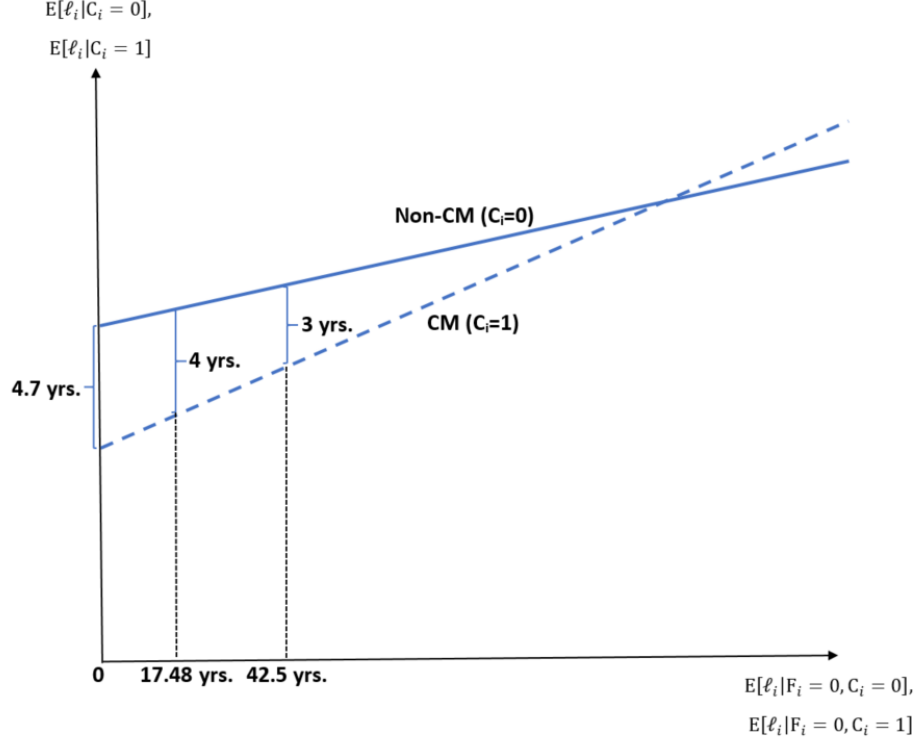
$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1].$$

Then, geometrically, the effect of cousin marriage on the offspring longevity is equal to the vertical distance between the two lines evaluated at the value of the conditional mean (see Appendix Figure E.14). For example, if everyone who is excluded from our sample died at age zero, i.e.,

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1] = 0,$$

then the effect of cousin marriage is equal to the distance between the two lines evaluated at zero, i.e. 4.7 years.

Figure E.14: The effect of cousin marriage on the offspring longevity in **Case 1**, i.e.,  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$



Note: This figure illustrates how the magnitude of the treatment effect is underestimated in **Case 1**, i.e., when  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$ . Our results in the main text indicate that the treatment effect is approximately 3 years. However, when we take into account the sample selection, the true effect is greater than 3 years under **Case 1**, as long as the unobservable conditional means, i.e.,  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0]$  and  $\mathbb{E}[\ell_i|F_i = 0, C_i = 1]$ , are less than 42.5 years. We argue in the text that 42.5 years is a conservative upper bound for these means. When we calibrate the unobservable conditional means at 17.48 years (see text for how they are calibrated), then the true effect is estimated to be 4 years.

On the other hand, the effect of cousin marriage would equal the effect observed in our sample (i.e.,  $\mathbb{E}[\ell_i|F_i = 1, C_i = 1] - \mathbb{E}[\ell_i|F_i = 1, C_i = 0] \approx -3$  years) if the two unobservable conditional means are equal to 42.5 years, i.e.,

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1] = 42.5.$$

See Appendix Figure E.14. In **Case 1**, therefore, we underestimate the magnitude of the effect of cousin marriage, as long as the two unobservable conditional means are below 42.5 years.

In a back-of-the-envelope calculation, we show below that the conditional means are unlikely to be greater than 42.5 years. According to the life tables in Hacker (2010), the mean longevity for whites<sup>28</sup> in the nineteenth-century U.S.<sup>29</sup> ranges from 35.6 to 49.5 years for males, and from 37.1 to 51.2 years for females, depending on the decade. Taking into account the selection of long-lived families into online genealogical data (Stelter and Alburez-Gutierrez, 2022), the unconditional mean longevity of children whose parents are registered on FamilySearch ( $\mathbb{E}[\ell_i]$ ) is likely to be close to the upper limit of the ranges in the life tables. We assume that the unconditional mean longevity is 50 years, i.e.,

$$\mathbb{E}[\ell_i] = 50.$$

Then we can obtain (8) below, which is a linear constraint on the two conditional mean longevitys:

$$\begin{aligned}
50 &= \mathbb{E}[\ell_i] = \mathbb{E}[\mathbb{E}[\ell_i|C_i]] \\
&= \underbrace{\Pr(C_i = 1)}_{=0.027 \text{ by Assumption 1}} \cdot \overbrace{\mathbb{E}[\ell_i|C_i = 1]}^{\approx 0.235 \cdot \mathbb{E}[\ell_i|F_i=0, C_i=1] + 42 \text{ by (6)}} + \underbrace{\Pr(C_i = 0)}_{=0.973 \text{ by Assumption 1}} \cdot \overbrace{\mathbb{E}[\ell_i|C_i = 0]}^{\approx 0.195 \cdot \mathbb{E}[\ell_i|F_i=0, C_i=0] + 46.7 \text{ by (7)}} \\
&\approx 0.006345 \cdot \mathbb{E}[\ell_i|F_i = 0, C_i = 1] + 0.189735 \cdot \mathbb{E}[\ell_i|F_i = 0, C_i = 0] + 46.5731 \tag{8}
\end{aligned}$$

---

<sup>28</sup>While we do not observe race in our analysis sample, it is likely that most of the people in it are whites. This is because people in our analysis sample are offspring of parents who registered their marriage pre-1858, which must have been an unlikely event for most blacks at that time. The fact that few blacks exist in our subsample linked to 1900 and 1910 censuses supports our conjecture (see Table B.12).

<sup>29</sup>95% of people in our sample are born on or before 1900.

Combining the constraint (8) with the assumption of **Case 1** that

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1],$$

we obtain the following value for the conditional mean longevity:

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = \mathbb{E}[\ell_i|F_i = 0, C_i = 1] \approx 17.48$$

The effect of cousin marriage on offspring longevity when the conditional means are equal to 17.48 is approximately 4 years. Under these assumptions, we underestimate the magnitude of the effect of cousin marriage on offspring longevity by approximately one year.

**Case 2.** It is unlikely that the two conditional means are exactly equal. Instead, we suppose in **Case 2** that the mean longevity of cousin couple's offspring who are excluded from our sample is smaller than that of non-cousin couple's offspring, i.e.,

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] > \mathbb{E}[\ell_i|F_i = 0, C_i = 1] \quad (9)$$

In this scenario, we underestimate the magnitude of the effect of cousin marriage on the offspring longevity as long as the constraint (8) holds. This can be easily seen geometrically.<sup>30</sup>

Let  $\widetilde{\mathbb{E}[\ell_i|F_i = 0, C_i = 0]}$  and  $\widetilde{\mathbb{E}[\ell_i|F_i = 0, C_i = 1]}$  denote specific values of the two conditional

---

<sup>30</sup>Mathematically, recall from the constraint (8) that the two unobserved conditional means, i.e.,  $\mathbb{E}[\ell_i|F_i = 0, C_i = 1]$  and  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0]$ , satisfy the following equality:

$$0.006345 \cdot \mathbb{E}[\ell_i|F_i = 0, C_i = 1] + 0.189735 \cdot \mathbb{E}[\ell_i|F_i = 0, C_i = 0] \approx 3.4269$$

If we substitute  $\mathbb{E}[\ell_i|F_i = 0, C_i = 1]$  with  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] - k$  with  $k > 0$ , then we obtain expressions for the two conditional means as follows:

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = 17.48 + 0.032k \quad (10)$$

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 1] = 17.48 - 0.968k. \quad (11)$$



means that satisfy the following conditions:

**Condition 1.**  $\mathbb{E}[\widetilde{\ell_i|F_i=0, C_i=0}] > \mathbb{E}[\widetilde{\ell_i|F_i=0, C_i=1}]$ ;

**Condition 2.** The constraint (8) is satisfied, i.e.,

$$0.006345 \cdot \mathbb{E}[\ell_i|F_i=0, C_i=1] + 0.189735 \cdot \mathbb{E}[\ell_i|F_i=0, C_i=0] + 46.5731 \approx 50$$

Notice that  $k$  cannot be too large, since  $\ell_i$  is non-negative, and hence,

$$\mathbb{E}[\ell_i|F_i=0, C_i=1] = 17.48 - 0.968k \geq 0$$

which implies that  $0 < k \leq 18.057$ .

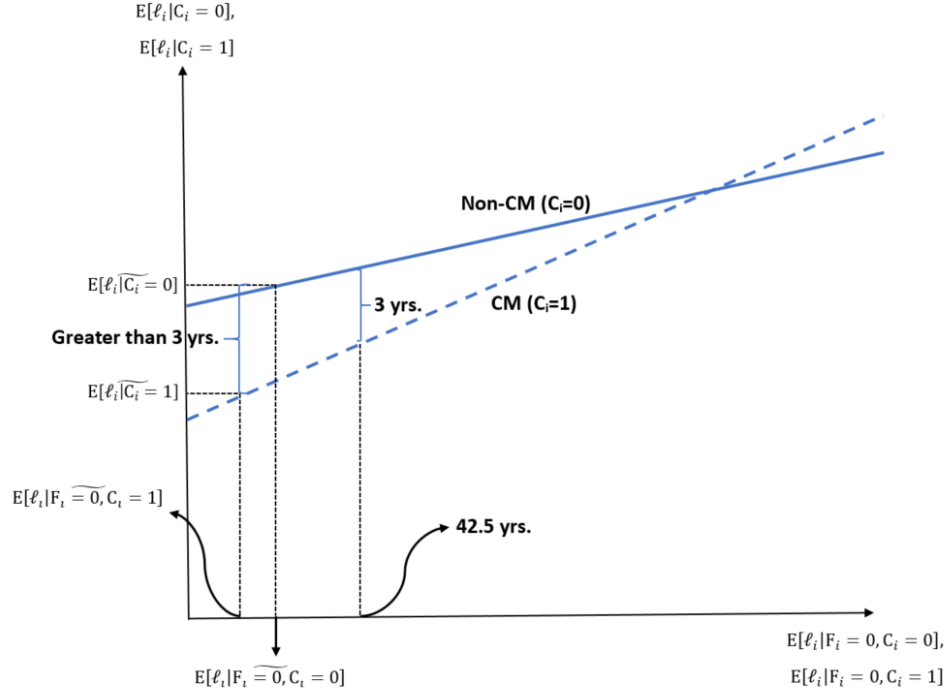
This also implies that neither of the conditional means can be greater than 42.5. This is because as  $k$  increases,  $\mathbb{E}[\ell_i|F_i=0, C_i=0]$  becomes larger but  $\mathbb{E}[\ell_i|F_i=0, C_i=1]$  becomes smaller to satisfy the constraint (8). However  $k$  cannot be too large since both conditional means have to be non-negative, and hence  $\mathbb{E}[\ell_i|F_i=0, C_i=0]$  cannot be too large. In fact, it will be less than or equal to 18.058.

Plugging (10) and (11) into the equality (6) and (7), respectively, and taking the difference between the two equalities, we obtain the effect of cousin marriage on the offspring longevity as a function of  $k$ :

$$\mathbb{E}[\ell_i|C_i=1] - \mathbb{E}[\ell_i|C_i=0] = -4.0008 - 0.23372k \text{ with } k \in (0, 18.057]$$

whose magnitude is greater than what we observe in our sample ( $\approx 3$  years) for any  $k$  in the range  $(0, 18.057]$ .

Figure E.15: The effect of cousin marriage on the offspring longevity in **Case 2**, i.e.,  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] > \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$



Note: This figure illustrates how the magnitude of the treatment effect is underestimated in **Case 2**, i.e., when  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] > \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$ . Our results in the main text indicate that the treatment effect is approximately 3 years. However, when we take into account the sample selection, the true effect is greater than 3 years under **Case 2** as long as the constraint (8) is satisfied.

Figure E.15 illustrates the geometry behind our results. In this figure, we have marked on the horizontal axis the values of the two conditional means, i.e., “ $\mathbb{E}[\ell_i|\widetilde{F_i = 0}, C_i = 0]$ ” and “ $\mathbb{E}[\ell_i|\widetilde{F_i = 0}, C_i = 1]$ ”, that satisfy Conditions 1 and 2, and the corresponding mean longevity on the vertical axis (“ $\mathbb{E}[\ell_i|\widetilde{C_i = 0}]$ ” and “ $\mathbb{E}[\ell_i|\widetilde{C_i = 1}]$ ”, respectively). As can be seen in the figure, the difference in the mean longevity between offspring of cousin couples and non-cousin couples is greater than 3 years. This is because of two reasons: first, 3 years is the shortest vertical distance between the two lines (labeled “Non-CM ( $C_i = 0$ )” and “CM

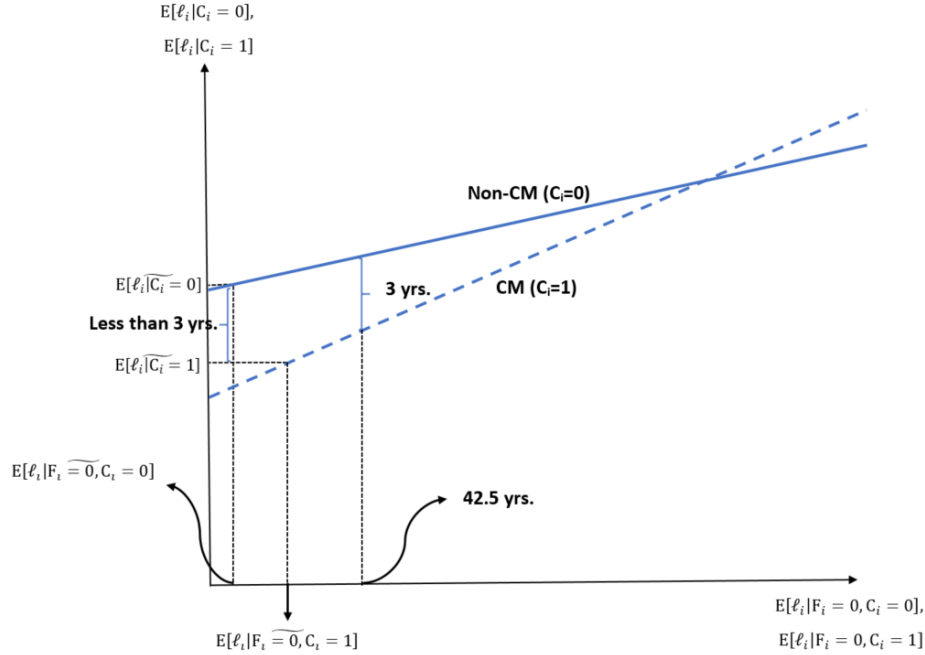
( $C_i = 1$ )” as long as the conditional means are less than 42.5, and the two conditional means are indeed less than 42.5 if they were to satisfy constraint (8) and non-negativity constraint simultaneously (see Footnote 30 for details). Second, the “Non-CM” line is flatter than the “CM” line. Recall that the difference in slopes is statistically significant at the 1 percent level.

**Case 3.** Suppose that the following inequality holds between the two unobservable conditional means:

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] < \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$$

In this case, our sample selection may actually lead us to overestimate the magnitude of the effect of cousin marriage on the offspring’s longevity, depending on the difference between the two conditional means. Figure E.16 presents an example. In this figure, the difference between the two conditional means is sufficiently large so that the magnitude of the effect, i.e.,  $\mathbb{E}[\widetilde{\ell_i|C_i = 0}] - \mathbb{E}[\widetilde{\ell_i|C_i = 1}]$  is less than 3 years.

Figure E.16: The effect of cousin marriage on the offspring longevity in **Case 3**, i.e.,  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] < \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$



Note: This figure illustrates how the magnitude of the treatment effect can be overestimated in **Case 3**, i.e., when  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] < \mathbb{E}[\ell_i|F_i = 0, C_i = 1]$ . Our results in the main text indicate that the treatment effect is approximately 3 years. However, when we take into account the sample selection, the true effect can be less than 3 years under **Case 3**, if the difference between the two conditional means is greater than 4.28 years.

We can mathematically<sup>31</sup> solve for the minimum difference between the two conditional

<sup>31</sup>Recall from (8) that the two unobserved conditional means, i.e.,  $\mathbb{E}[\ell_i|F_i = 0, C_i = 1]$  and  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0]$ , satisfy the following equality:

$$0.006345 \cdot \mathbb{E}[\ell_i|F_i = 0, C_i = 1] + 0.189735 \cdot \mathbb{E}[\ell_i|F_i = 0, C_i = 0] \approx 3.4269$$

If we substitute  $\mathbb{E}[\ell_i|F_i = 0, C_i = 1]$  with  $\mathbb{E}[\ell_i|F_i = 0, C_i = 0] + k$  with  $k > 0$ , then we obtain expressions for the two conditional means as follows:

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] = 17.48 - 0.032k \quad (12)$$

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 1] = 17.48 + 0.968k. \quad (13)$$

Notice that the non-negativity constraints for the conditional means are not binding, since  $k$  has to be greater than  $546.25 = \frac{17.48}{0.032}$  in order for the non-negativity constraint to be violated.

Plugging (12) and (13) into the equality (6) and (7), respectively, and taking the difference between the

means that would lead to over-estimation. We find that the difference larger than 4.28 years, i.e.,

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 1] - \mathbb{E}[\ell_i|F_i = 0, C_i = 0] \geq 4.28$$

would lead us to overestimate the effect of cousin marriage on the offspring's longevity (see Footnote 31 for details).

However, we believe that such a difference is unrealistic, considering what is known about the effects of cousin marriage on offspring mortality. Almost all previous studies of cousin marriage and offspring mortality find that parents' cousin marriage increases the mortality of their young children Bittles (2012). Importantly, these studies use survey or administrative data rather than genealogical data, and hence do not have the same selection issues as ours (e.g., Grant and Bittles (1997)). Intuitively, if selection into our sample is independent of parents' cousin marriage status, then the offspring of cousin couples excluded from our sample should have died at a younger age (on average) than the offspring of non-cousin couples excluded from our sample.

To formalize this intuition, we present a proposition below:

**Proposition 1.** *Let  $\Pr(\ell_i|C_i)$  denote the probability mass function for longevity conditional on parents' cousin-marriage status. Let  $\mathcal{L} = \{0, 1, \dots, L\}$  denote the support of the probability mass function. Suppose the following three conditions hold.*

**Condition 1.**  $\Pr(\ell_i|C_i = 0)$  first-order stochastically dominates  $\Pr(\ell_i|C_i = 1)$ ;

**Condition 2.** For each  $\ell \in \mathcal{L}$ ,  $\Pr(F_i = 0|\ell_i = \ell, C_i)$  is independent of  $C_i$ ; and

---

two equalities, we obtain the effect of cousin marriage on the offspring longevity as a function of  $k$ :

$$\mathbb{E}[\ell_i|C_i = 1] - \mathbb{E}[\ell_i|C_i = 0] = -4.0008 + 0.23372k \text{ with } k > 0$$

whose magnitude is greater than what we observe in our sample ( $\approx 3$  years) if  $k$  is greater than 4.282.

**Condition 3.**  $\Pr(F_i = 0|\ell_i = \ell)$  weakly decreases in  $\ell$ , and  $\ell \cdot \Pr(F_i = 0|\ell_i = \ell)$  weakly increases in  $\ell$  over  $\mathcal{L}$

Then, the following inequality holds:

$$\mathbb{E}[\ell_i|F_i = 0, C_i = 0] \geq \mathbb{E}[\ell_i|F_i = 0, C_i = 1] \quad (14)$$

The proof of Proposition 1 is collected at the end of this section. This proposition states that under three conditions, the inequality (14) holds between the two unobservable conditional means. The implications of the proposition are: (a) to the extent that the sufficient conditions are realistic, **Case 3** is not empirically relevant; and (b) we are underestimating the magnitude of the effect of cousin marriage on the offspring longevity, as shown in **Case 1** or **Case 2**, where we assume that the inequality (14) holds.

Discussions about the three sufficient conditions are in order. The first condition can be interpreted as parents' cousin marriage negatively affecting children's health throughout their lives. This is evidenced by panel (a) of Figure 2, where we show that the gap between the survival function of offspring of non-cousin couples and that of offspring of cousin couples continues to widen over the lifespan. We model this fact in Condition 1 by assuming that the distribution of longevity of cousin couples' offspring is first-order stochastically dominated.

Condition 2 states that whether a child is registered on his/her parent's FamilySearch profile or not does not depend on the parents' cousin marriage status, conditional on longevity. We cannot test this condition directly with our data, since we do not observe the longevity of those excluded from our sample.

This condition would be violated if descendants of non-cousin couples are better at recovering the vital information of their ancestors than their cousin-couple counterparts (or

the other way around). That would be possible if, for example, the offspring of non-cousin couples were more likely to be believers of The Church of Jesus Christ of Latter-day Saints (LDS henceforth), who are known to seek information about their family members for religious reasons.<sup>32</sup> While we do not observe the religious affiliation of people in our sample, we do observe the birth state of a majority of them (92.3 percent). Since a large share of believers of the LDS have lived in the state of Utah, we use being born in Utah as a proxy for being a believer of the LDS. Those born in Utah indeed show characteristics that we expect to see from the believers of the LDS: they are much less likely to miss birth or death years than non-Utah-borns. Only 1.5 percent of the former are missing vital years, whereas 18.7 percent of the latter are.

We find that Utah-borns are more likely to be the offspring of non-cousin-couples. However, they account for such a small fraction of our sample that they are unlikely to affect our results. Utah-borns account for 0.9 percent of the non-cousin-couple offspring and 0.3 percent of their cousin-couple counterparts. Unsurprisingly, when we drop all Utah-borns, our main results hardly change (see Table E.14).

---

<sup>32</sup>To the believers of LDS, seeking information about their ancestors is important because they believe that they can be connected to their family members forever, even after death. According to the official website of the religion (The Church of Jesus Christ of Latter-Day Saints, 2023), “because Church members believe that families are forever, they seek to identify generations of relatives through family history work. By discovering where your family came from, who your ancestors were, and what motivated them, you learn about your core family values, and you build a bridge connecting you to your ancestors.”

Table E.14: The effect of cousin marriage on longevity, with or without Utah-borns

	(1) With	(2) Without
Parents are first cousins	-3.14*** (0.35)	-3.15*** (0.35)
Control mean	58.54	58.55
Observations	6,080,885	6,016,798
Individual controls	Yes	Yes
Paternal FE	Yes	Yes
Maternal FE	Yes	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.010$ . Observations are at the level of children. This table shows estimates for the coefficient  $\beta$  from equation (1) with or without those born in the state of Utah. In both specifications in this table, we control for the parents' sibling fixed effects and the following covariates: birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order. These are described in appendix C. Standard errors are clustered at the level of the individual and their siblings. For this table, we use the subsample of our analysis sample whose birth state is observed.

Lastly, the first part of Condition 3 implies that you are more likely to be included in our sample if you died at a more advanced age (i.e.,  $\Pr(F_i = 0 | \ell_i = \ell)$  weakly decreases in  $\ell$ , or equivalently,  $\Pr(F_i = 1 | \ell_i = \ell)$  weakly increases in  $\ell$ ). This condition is justifiable as longer-lived people tend to have more descendants who can potentially create their FamilySearch profiles. Longer-lived people also tend to have more official records of birth and death years, as the coverage of the U.S. birth and death registration system gradually expanded over time.

On the other hand, the second part of Condition 3 imposes an upper bound on how fast  $\Pr(F_i = 0 | \ell_i = \ell)$  can decrease. If  $\ell \cdot \Pr(F_i = 0 | \ell_i = \ell)$  is increasing in  $\ell$ , then the implied upper bound on the magnitude of the percent change from  $\Pr(F_i = 0 | \ell_i = \ell)$  to



$\Pr(F_i = 0|\ell_i = \ell + 1)$  is equal to  $\frac{1}{\ell+1}$ .<sup>33</sup>

Taken together, Condition 3 imposes a mild restriction on the shape of the function  $\Pr(F_i = 0|\ell_i = \ell)$ . Most importantly, Condition 3 does not impose any upper bound on the probability that an infant death is not included in our data, i.e.,  $\Pr(F_i = 0|\ell_i = 0)$ . This is important for two reasons: first, we prefer not to restrict the probability because we do not have a very good sense of how large the probability should be. Second, a restriction on  $\Pr(F_i = 0|\ell_i = 0)$  could affect other values of the function  $\Pr(F_i = 0|\ell_i = \ell)$ .<sup>34</sup>

Condition 3 may appear to restrict the shape of the function  $\Pr(F_i = 0|\ell_i = \ell)$  for large values of  $\ell$ 's. Since the upper bound on the magnitude of the percent changes from  $\Pr(F_i = 0|\ell_i = \ell)$  to  $\Pr(F_i = 0|\ell_i = \ell + 1)$  approaches 0 as  $\ell$  increases, the function  $\Pr(F_i = 0|\ell_i = \ell)$  should be nearly constant for large values of  $\ell$ .

However, this shape restriction is justifiably realistic. The probability that a person is included in our sample should, for the most part, depends on factors that are not sensitive to age at death, as long as the age at death is large enough. For example, the number of offspring is arguably one of the most important factors that determine whether one is included in our sample or not: if one has more offspring, then the probability that one of them registers you on FamilySearch would be larger, everything else equal. Since one's death is preceded by the end of his/her reproduction, the probability of being included in our sample should not be sensitive to age at death.

---

<sup>33</sup>To see why, suppose that  $\ell \cdot \Pr(F_i = 0|\ell_i = \ell)$  is increasing in  $\ell$ , i.e.,

$$(\ell + 1) \cdot \Pr(F_i = 0|\ell_i = \ell + 1) \geq \ell \cdot \Pr(F_i = 0|\ell_i = \ell)$$

This inequality implies the following:

$$-\left(\frac{\Pr(F_i = 0|\ell_i = \ell + 1) - \Pr(F_i = 0|\ell_i = \ell)}{\Pr(F_i = 0|\ell_i = \ell)}\right) \leq \frac{1}{\ell + 1}$$

where we take the negative of the percent change to obtain the magnitude of the percent change. The percent change is negative since Condition 3 requires that  $\Pr(F_i = 0|\ell_i = \ell)$  is decreasing in  $\ell$ .

<sup>34</sup>To see why, consider an extreme case that satisfies Condition 3: the function  $\Pr(F_i = 0|\ell_i = \ell)$  is constant for all  $\ell \in \mathcal{L}$ . Then any restriction on  $\Pr(F_i = 0|\ell_i = 0)$  will affect the value of the function at other  $\ell$ 's. Clearly, a similar logic would apply to less extreme cases.

*Proof of Proposition 1.* Note first that the mean longevity conditional on being excluded from our sample and on parents' cousin marriage status is equal to:

$$\begin{aligned} & \mathbb{E}[\ell_i | F_i = 0, C_i] \\ &= \sum_{\ell \in \mathcal{L}} \ell \cdot \Pr(\ell_i = \ell | F_i = 0, C_i) \end{aligned} \tag{15}$$

By Bayes' Theorem, each probability in (15) can be rewritten as follows:

$$\Pr(\ell_i = \ell | F_i = 0, C_i) = \frac{\Pr(F_i = 0 | \ell_i = \ell, C_i) \cdot \Pr(\ell_i = \ell | C_i)}{\Pr(F_i = 0 | C_i)} \tag{16}$$

Plugging in (16) into (15), we obtain the following:

$$\begin{aligned} & \mathbb{E}[\ell_i | F_i = 0, C_i] \\ &= \sum_{\ell \in \mathcal{L}} \ell \cdot \frac{\Pr(F_i = 0 | \ell_i = \ell, C_i) \cdot \Pr(\ell_i = \ell | C_i)}{\Pr(F_i = 0 | C_i)} \\ &= \frac{1}{\Pr(F_i = 0 | C_i)} \cdot \left( \sum_{\ell \in \mathcal{L}} \ell \cdot \Pr(F_i = 0 | \ell_i = \ell, C_i) \cdot \Pr(\ell_i = \ell | C_i) \right) \\ &= \frac{1}{\Pr(F_i = 0 | C_i)} \cdot \left( \sum_{\ell \in \mathcal{L}} \ell \cdot \Pr(F_i = 0 | \ell_i = \ell) \cdot \Pr(\ell_i = \ell | C_i) \right) \end{aligned} \tag{17}$$

where the last equality follows from the independence assumption in Condition 2.

Let us first focus on the terms in the parenthesis in (17). Since  $\ell \cdot \Pr(F_i = 0 | \ell_i = \ell)$  is weakly increasing in  $\ell$  by Condition 3, and  $\Pr(\ell_i | C_i = 0)$  first-order stochastically dominates  $\Pr(\ell_i | C_i = 1)$  by Condition 1, the following holds by the definition of the first-order stochastic dominance (Courtault et al., 2006):

$$\left( \sum_{\ell \in \mathcal{L}} \ell \cdot \Pr(F_i = 0 | \ell_i = \ell) \cdot \Pr(\ell_i = \ell | C_i = 0) \right) \geq \left( \sum_{\ell \in \mathcal{L}} \ell \cdot \Pr(F_i = 0 | \ell_i = \ell) \cdot \Pr(\ell_i = \ell | C_i = 1) \right)$$

Then we only need to show the following inequality holds in order to close the proof:

$$\frac{1}{\Pr(F_i = 0|C_i = 0)} \geq \frac{1}{\Pr(F_i = 0|C_i = 1)}$$

or equivalently,

$$\Pr(F_i = 0|C_i = 1) \geq \Pr(F_i = 0|C_i = 0) \quad (18)$$

To see why (18) holds, the Law of Total Probability requires that:

$$\begin{aligned} \Pr(F_i = 0|C_i) &= \sum_{\ell \in \mathcal{L}} \Pr(F_i = 0|\ell_i = \ell, C_i) \cdot \Pr(\ell_i = \ell|C_i) \\ &= \sum_{\ell \in \mathcal{L}} \Pr(F_i = 0|\ell_i = \ell) \cdot \Pr(\ell_i = \ell|C_i) \end{aligned}$$

where the second equality follows from the independence assumption in Condition 2. Since  $\Pr(F_i = 0|\ell_i = \ell)$  is weakly decreasing by Condition 3, it follows that its negative, i.e.,  $-\Pr(F_i = 0|\ell_i = \ell)$ , is weakly increasing. Therefore, by the definition of first-order stochastic dominance, we have:

$$\begin{aligned} \sum_{\ell \in \mathcal{L}} (-\Pr(F_i = 0|\ell_i = \ell)) \cdot \Pr(\ell_i = \ell|C_i = 0) &\geq \sum_{\ell \in \mathcal{L}} (-\Pr(F_i = 0|\ell_i = \ell)) \cdot \Pr(\ell_i = \ell|C_i = 1) \\ \Rightarrow \sum_{\ell \in \mathcal{L}} \Pr(F_i = 0|\ell_i = \ell) \cdot \Pr(\ell_i = \ell|C_i = 0) &\leq \sum_{\ell \in \mathcal{L}} \Pr(F_i = 0|\ell_i = \ell) \cdot \Pr(\ell_i = \ell|C_i = 1) \\ \Rightarrow \Pr(F_i = 0|C_i = 0) &\leq \Pr(F_i = 0|C_i = 1) \end{aligned}$$

as desired. □

## F Online Appendix: Survival model (Cox proportional hazards)

In this section, we check the robustness of our results in the main section against an alternative model of lifespan. In the main text, our model specifies longevity as a linear function of a vector of covariates including the maternal and paternal fixed effects. In this section, we estimate the Cox proportional hazards model, which is widely used to estimate time-to-death, and check if the estimates of this alternative model are consistent with those obtained from our linear model.

For readers who are unfamiliar with the Cox proportional hazards model, we briefly describe the model specification. The Cox proportional hazards model specifies a person’s hazard rate at a given age as a product of two terms: the baseline hazard rate and the relative risk. The baseline hazard rate can depend on which “group” one belongs to. In our empirical context, we posit that there are two groups based on the cousin marriage status of one’s parents. The baseline hazard rate is common to all members of a group but varies across groups. The relative risk, on the other hand, varies within a group and depends on person-specific covariates. In our context, a person’s hazard rate for death may depend on, for example, maternal age at birth.

The hazard rate at age  $t$  of a person  $i$  in group  $g$  is specified as follows:

$$h(t|X_i, g) = h_{0g}(t) \exp(X_i' \delta) \quad (19)$$

where  $h_{0g}(t)$  is the group- $g$  baseline hazard rate at age  $t$  and  $X_i$  is a vector of covariates for person  $i$ .  $\exp(X_i' \delta)$  is referred to as the relative risk.

The popularity of the Cox proportional hazards model stems from the fact that no parametric form is imposed on the baseline hazard rate function  $h_{0g}(t)$ , unlike other

parametric hazards models that assume a particular form for  $h_{0g}(t)$ .

As discussed, we investigate if the estimated hazard rate function is consistent with our estimates in the main text. Recall that we did not estimate hazard rate functions in the main text. What we did estimate within our linear model are: a) the effect of cousin marriage on offspring longevity; and b) age-specific difference in survival probabilities (see Figure 2b). It is straightforward to show that these quantities can be derived from the hazard rate function estimates, which makes our comparison viable.<sup>35</sup>

---

<sup>35</sup>For example, the survival function, denoted by  $S(t)$ , is equal to the following:

$$S(t) = \exp\{-H(t)\}$$

where  $H(t)$  is the cumulative hazard function, defined as follows:

$$H(t) \equiv \int_0^t h(u) du.$$

where  $h(\cdot)$  is the hazard rate function.

The density of age at death, denoted by  $f(t)$ , can be derived from the hazard rate function as follows:

$$f(t) = h(t) \exp\{-H(t)\}.$$

The effect of one's parents' cousin marriage can be derived from the density of the age at death as follows:

$$\mathbb{E}[t_i | C_i = 1, X_i] - \mathbb{E}[t_i | C_i = 0, X_i]$$

where  $t_i$  denotes  $i$ 's age at death, and the expectation is taken with respect to the density  $f(t)$ .  $C_i$  denotes  $i$ 's parents' cousin marriage status.

Table F.15: Robustness of our results against alternative model: the treatment effect

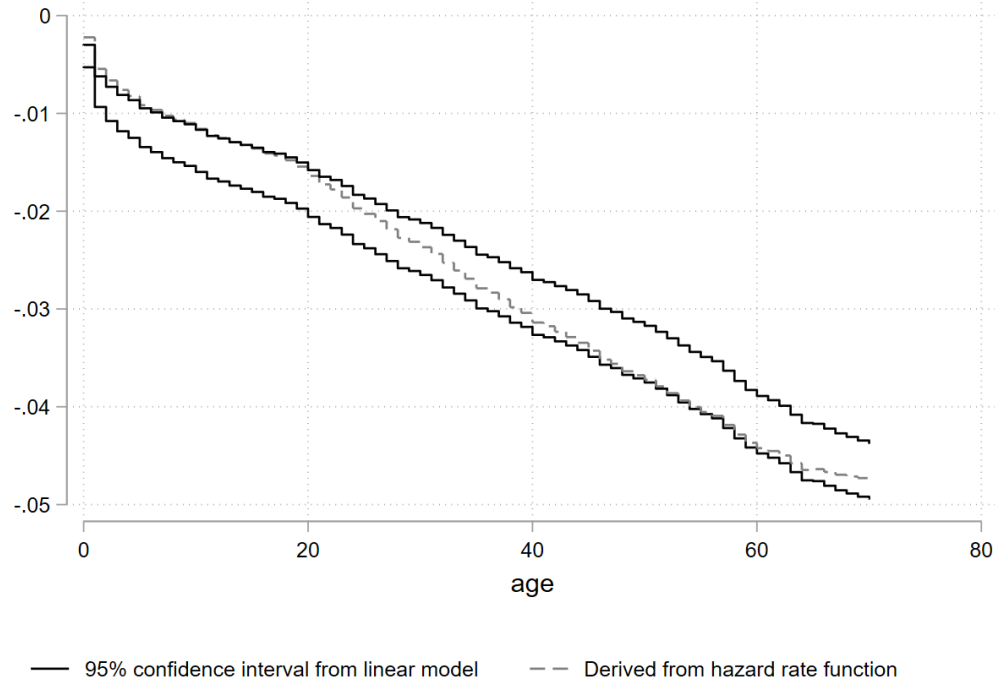
Effects of cousin marriage on longevity		
Hazard function	Linear model	95% C.I.
-2.614	-2.628	[-2.809, -2.448]

Note: In this table, we present the effect of cousin marriage on longevity estimated within two different models. The estimates in the column labeled “Hazard function” are derived from our estimates of the hazard rate function. For derivation of the estimates, see footnote 35. The columns labeled “Linear model” and “95% C.I.” contain the estimates and the 95% confidence interval from a model similar to (1) in the main text. The covariates that are controlled in either model are the same as those for the model (1), except for the father/mother sibling fixed effects.

In Table F.15, we compare the treatment effects derived from the estimated hazard rate function with the estimates from our linear model. The treatment effects derived from the hazard rate function indicate that parents’ cousin marriage reduces the mean longevity of their children, as do the estimates from the linear model. In addition, the magnitude of these effects is similar to what we obtain from the linear model, and they lie within the 95 percent confidence interval estimated with the linear model.

Note that treatment effects in Table F.15 are obtained from a model with the same set of controls as the one in the main text but without the father and mother sibling fixed effects. The reason that we do not include the fixed effects is computational feasibility: unlike our linear model, estimating a hazard rate function with over a million fixed effects via the maximum likelihood estimator is not computationally feasible. Thus the estimates in the column labeled “Linear model” in Table F.15 are comparable to column 2, rather than column 3, of Table 2. They do not include the maternal and paternal fixed effects.

Figure F.17: The magnitude of shift of survival functions due to parents' cousin marriage



Note: This figure presents the magnitude of shift of children's survival function due to parents' cousin marriages. We estimate the shift in two different models: within a linear model and using hazard function estimates, respectively. The former is obtained from a linear probability model estimated at each age, where the dependent variable is equal to 1 if one survives up to a given age and 0 otherwise. The solid lines in the figure correspond to the bounds of the 95 percent confidence intervals for the parents' cousin marriage indicator. The dashed line, on the other hand, corresponds to the estimates derived from the hazard rate function. In either model, we control the same set of covariates as our baseline model (1), except for the father/mother sibling fixed effects.

Turning to survival functions, we plot in Figure F.17 how much parents' cousin marriages shift the survival function of their children. This figure is essentially the same as panel (b) in Figure 2 in the main text, except that here we use the model without the father and mother sibling fixed effects for computational feasibility.

The two solid lines in Figure F.17 correspond to the bounds of the 95% confidence interval of the shift in survival function due to parents' cousin marriage. These bounds

are obtained from our linear model. The dashed line comes from our hazard rate function estimates. As is clear from the figure, the two models yield similar estimates: the dashed line (derived from the hazard rate function) largely stays within the bounds of the confidence interval (obtained from the linear model), except for 13-year-olds or younger.<sup>36</sup> However, even when the dashed line is outside of the confidence interval, the magnitude of the deviation is negligible, with a maximum deviation equal to .00125.

Our results in this section suggest that our baseline estimates, obtained with a linear model of longevity, are robust to a popular alternative approach to modeling lifespans.

---

<sup>36</sup>Some of the discrepancies between the two models may originate from the fact that we are using linear probability model, whose fit to the data deteriorates as the modeled probability approaches zero or one (Long (1997)).