Health effects of cousin marriage: Evidence from US genealogical records

Sam Il Myoung Hwang
Deaglan Jakob
Munir Squires

Online Appendix

Table 1: Construction of analysis sample

	(1) Total Dropped	(2) Percent Dropped	(3) Remaining Observations
Nonmissing sex	60,502	0.15	40,514,197
Nonmissing birth year	581,969	1.44	39,932,228
Nonmissing death year	10,903,531	27.31	29,028,697
Nonmissing great-grandparents	21,618,457	74.47	7,410,240
Nonmissing maternal age at birth	29,439	0.40	7,380,801
Lifespan between 0 and 98	49,569	0.67	7,331,232
Birth year between 1750 and 1920	665,900	9.08	6,665,332
Survived to age 5	608,310	9.13	6,057,022
Singletons	142,785	2.36	5,914,237

This table shows how we create our final analysis sample of 5.9 million offspring from over 40 million genealogical records. Each row shows the number of observations remaining after we drop those for which a specific variable is missing. Singletons are groups with only one observation. These are relevant for specifications in which we include mother and father siblings fixed effects. See Correia (2015) for a more detailed description of singletons.

Table 2: Descriptive statistics - analysis sample

Analysis sample: Individuals with non-missing great-grandparents

	(1) Parents are first cousins	(2) Non-cousin	(3) Difference
Longevity conditional on surviving to age 5	60.97 [22.58]	63.73 [22.30]	-2.77 (0.06)
Parent Longevity	69.13 [11.90]	69.66 [11.59]	-0.53 (0.03)
Year of Birth	1,842.84 [32.35]	1,848.35 [34.25]	-5.51 (0.09)
Mother's Age at Birth	30.14 [7.02]	29.79 [6.97]	0.35 (0.02)
Female	0.47 [0.50]	0.47 [0.50]	-0.00 (0.00)
Number of brothers	4.21 [2.21]	4.16 [2.22]	0.05 (0.01)
Number of sisters	3.91 [2.14]	3.85 [2.15]	0.07 (0.01)
Birth order	4.39 [2.75]	4.33 [2.73]	0.06 (0.01)
Observations	148,682	5,765,555	5,914,237
Percent	2.51	97.49	100

Each observation is an offspring in the analysis sample. This table shows the mean of each variable we use in our preferred specification in table II of the main paper. We require children to survive until at least age 5 to be included in the analysis sample we describe here. 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). Parental longevity is substantially higher than child longevity since it only includes individuals who have children (and hence survived to reproductive age). Variable descriptions are in section II of the main paper. Standard deviations are in square brackets. Standard errors are in parentheses.

Table 3: Comparison of socio-demographic characteristics between key samples (1850)

	(1) 1% Census sample (IPUMS)	(2) Subset of (1) with genealogical profile	(3) Intersection of (2) and (4)	(4) Genealogical profiles in analysis sample
Panel A: Variables available in census records and genealogical profiles				
Female	.49	.49	.46	.46
Age in 1850	22.34	21.41	23.33	24.15
	[17.56]	[17.51]	[18.65]	[19.27]
Born in Northeast	.41	.43	.52	.52
Born in Midwest	.15	.18	.19	.18
Born in South	.32	.34	.28	.29
Born in West	0	0	0	0
Born in Foreign-born	.11	.04	0	.02
Panel B: Variables available only in census records				
Non-white	.02	0	0	
Related to head	.91	.98	.98	
Literate	.89	.9	.95	
White-collar	.1	.09	.11	
Farmer	.45	.59	.62	
Skilled	.26	.2	.18	
Unskilled	.19	.12	.09	
Live in urban area	.17	.1	.08	
Live on a farm	.53	.63	.69	
Value of real estate	249.92	296.26	470.27	
	[2978.47]	[3065.54]	[2867.36]	
Panel C: Variables available only in genealogical profiles				
Longevity conditional		65.11	65.65	66.09
on surviving to age 5		[19.99]	[19.83]	[19.65]
Mother's age at birth		29.02	29.22	29.33
		[7.03]	[6.94]	[6.95]
Number of brothers		4.08	4.37	4.34
		[2.25]	[2.22]	[2.25]
Number of sisters		3.76	3.95	3.98
		[2.13]	[2.13]	[2.16]
Birth order		3.94	4.16	4.2
		[2.71]	[2.69]	[2.72]
Sibling sex ratio		.48	.47	.48
		[.2]	[.19]	[.2]
Observations	197796	109825	15592	2257765

Note: This table compares the characteristics of individuals from four samples who were alive in 1850 (born pre-1850 and died post-1850). Column (1) corresponds to the 1850 U.S. Federal Census IPUMS 1% sample (Ruggles et al., 2024a); column (2) to the subsample of (1) linked to a genealogical profile on FamilySearch, as created in Hwang and Squires (2024); column (3) to the subsample of (2) that overlaps with our analysis sample; and (4) to our analysis sample. The variables in Panel A are available both in the census sample and our analysis sample, while Panels B and C contain variables available in one dataset but not the other. The numbers in the brackets represent standard deviations. We group occupations into white-collar, farmer, skilled, and unskilled following the categories used in Long and Ferrie (2013).

Table 4: Descriptive statistics - parent sample

Parent sample: Parents of analysis sample

	(1) Married to first cousin	(2) Not married to first cousin	(3) Difference
Longevity	67.38	67.95	-0.57
	[16.96]	[16.89]	(0.09)
Year of Birth	1,812.84	1,817.55	-4.71
	[25.96]	[28.06]	(0.15)
Mother's Age at Birth	29.39	30.07	-0.68
	[6.87]	[7.24]	(0.04)
Number of Children	6.80	6.54	0.27
	[3.42]	[3.41]	(0.02)
Female	0.48	0.49	-0.01
	[0.50]	[0.50]	(0.00)
Number of brothers	2.62	2.44	0.19
	[1.74]	[1.67]	(0.01)
Number of sisters	2.35	2.28	0.07
	[1.57]	[1.53]	(0.01)
Birth order	2.87	2.87	-0.00
	[1.80]	[1.80]	(0.01)
Observations	35,363	1,219,212	1,254,575
Percent	2.82	97.18	100

Each observation is a parent of one of the offspring in our analysis sample. This table shows the mean of each variable we use in our preferred specification in table I of the main paper. 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 section II of the main paper. Standard deviations are in square brackets. Standard errors are in parentheses.

Table 5: Comparison of Socio-Demographic Characteristics Between First-Cousin Couples and Non-First-Cousin Couples

	Married to first cousin		Not married to first cousin		Difference	
	Obs.	Mean	Obs.	Mean	Raw	Sib. FE
Female	142	.51	6178	.48	.03	.02
					(.04)	(.04)
Born in Northeast	142	.24	6178	.21	.03	.02
					(.03)	(.01)
Born in Midwest	142	.11	6178	.32	-.21	-.02
					(.04)	(.02)
Born in South	142	.65	6178	.45	.2	0
					(.04)	(.01)
Born in West	142	.01	6178	.02	-.01	0
					(.01)	(0)
Foreign-born	142	0	6178	.01	-.01	0
					(.01)	(.01)
Non-white	142	0	6178	0	0	0
					(0)	(0)
Related to head	142	.99	6178	.99	0	-.01
					(.01)	(.01)
Literate	77	.84	3093	.92	-.08	-.04
					(.03)	(.04)
White-collar	45	.04	1820	.08	-.04	.12
					(.04)	(.08)
Farmer	45	.71	1820	.59	.12	-.18
					(.07)	(.12)
Skilled	45	.18	1820	.1	.08	0
					(.05)	(.09)
Unskilled	45	.07	1820	.22	-.15	.06
					(.06)	(.09)
Live in urban area	142	.05	6178	.05	0	-.01
					(.02)	(.02)
Live on a farm	142	.81	6178	.79	.02	.02
					(.03)	(.03)
Value of real estate	111	232.43	4747	336.93	-104.5	-39.98
		[688.89]		[2418.37]	(229.78)	(297.61)
Value of asset	51	123.8	2134	290.5	-166.7	11.78
		[371.5]		[1453.6]	(203.75)	(254.33)

This table presents the results of a t-test of differences in socio-demographic characteristics between couples married to first cousins and those who are not. We restrict our sample to those who satisfy the following two conditions: (1) those whom we can link to the 1850-1930 IPUMS 1% samples (Ruggles et al., 2024a), the linkage of which is created in Hwang and Squires (2024); and (2) those who have a sibling that is linked as well. When a person is linked to the IPUMS samples multiple times (2.3 to 4.6 percent, depending on the characteristics), we use the average of the socio-demographic characteristics. The number of observations differs across rows due to differences in the universe for the census questions or differences in the share of missing values. The column labeled "Raw" presents the raw difference in the sample means and the standard errors. The column labeled "Sib. FE" displays the differences in the sample mean between two groups after including sibling fixed effects.

Table 6: The effect of cousin marriage on longevity with observations reweighted to match sex and birth region in the U.S. Census

	(1) Raw	(2) Controls	(3) Parental fixed effects
Panel A: Life expectancy at age 5			
Parents are first cousins	-2.86	-2.18	-2.26
	(0.10)	(0.10)	(0.58)
Control mean	64.64	64.64	64.64
Observations (thousands)	3,209	3,209	3,209
Panel B: Life expectancy at age 20			
Parents are first cousins	-2.35	-1.83	-1.81
	(0.08)	(0.08)	(0.52)
Control mean	67.97	67.97	67.97
Observations (thousands)	3,000	3,000	3,000
Individual controls	No	Yes	Yes
Paternal fixed effects	No	No	Yes
Maternal fixed effects	No	No	Yes

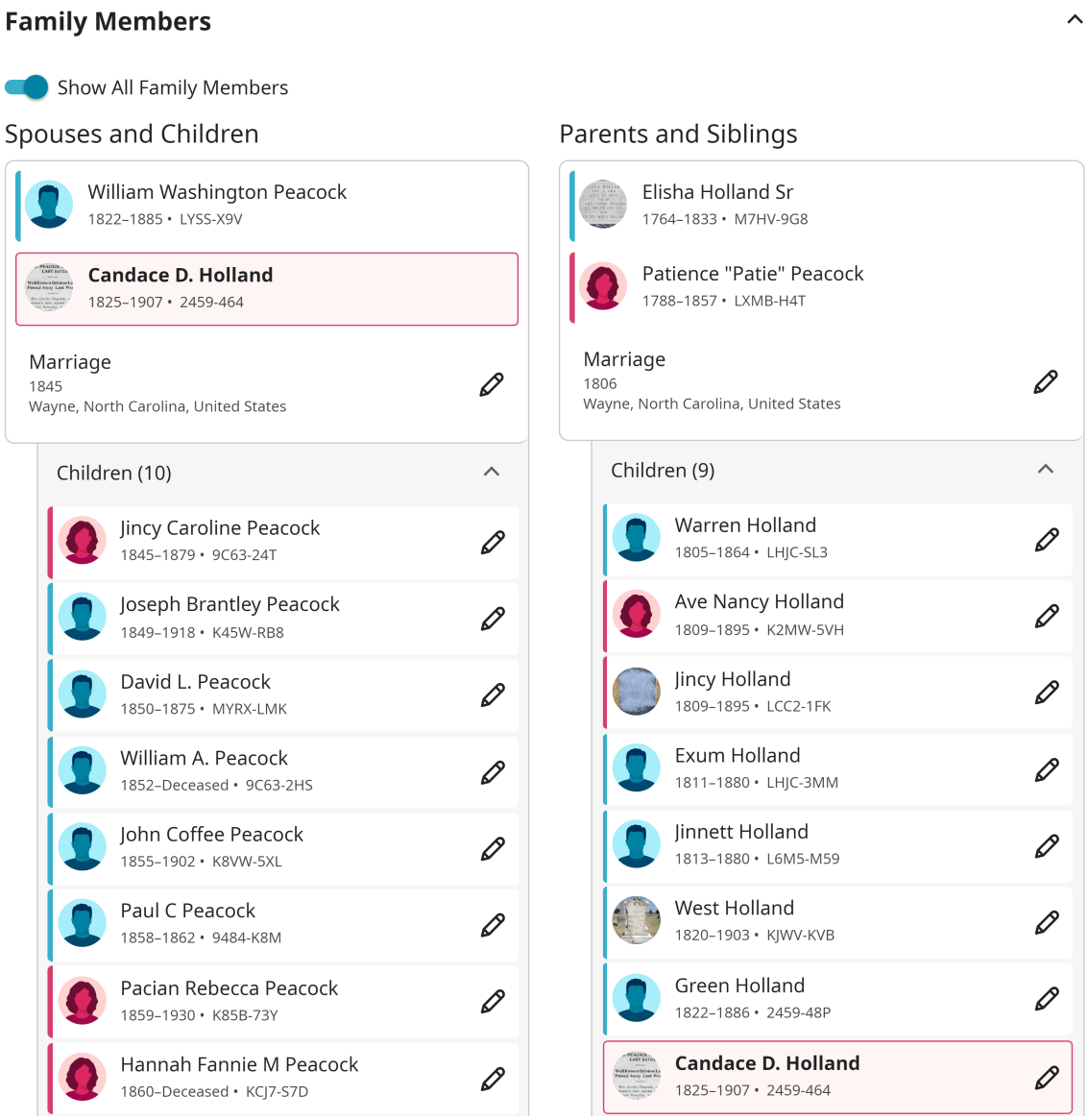
Note: This table shows estimates for the coefficient β from equation I of the main paper. Each observation is an offspring in the analysis sample. The outcome is the child's longevity (year of death minus year of birth), conditional on surviving to a specified age. Each regression weights members of each decadal birth cohort from 1840 to 1910 to match the sex and birth region of whites (Northeast, Midwest, South, and West) in the census closest to their birth (Ruggles et al., 2024b,c). For example, the 1840 birth cohort (those born between 1840 and 1849) is weighted so that the weighted share of each sex $\times$ birth region of whites matches the corresponding share in the full-count 1850 census. We exclude the 1880 birth cohort because the 1890 full-count census is not available. Column (1) coefficients are simply the weighted difference in mean longevity between the children of first cousins and the children of non-first cousins. Column (2) adds controls for birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order, as described in section II of the main paper. Column (3) adds mother's siblings fixed effects and father's siblings fixed effects. Standard errors are clustered at the level of the individual and their siblings. We restrict our sample to those born in the U.S.

Table 7: Life expectancy at birth

	Baseline specifications			Robustness checks				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Raw	Controls	Parental fixed effects	Flexible controls	Parent longevity	Age heaping	County-decade FE	Same surname
Life expectancy at birth								
Parents are first cousins	-3.10	-2.60	-3.18	-3.18	-3.08	-2.82	-3.09	-2.27
	(0.09)	(0.09)	(0.33)	(0.33)	(0.35)	(0.37)	(0.41)	(0.22)
Control mean	58.02	58.02	58.02	58.02	58.03	58.22	58.39	58.59
Observations (thousands)	6,539	6,539	6,539	6,539	6,460	5,744	5,087	13,978
Individual controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Paternal fixed effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Maternal fixed effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes

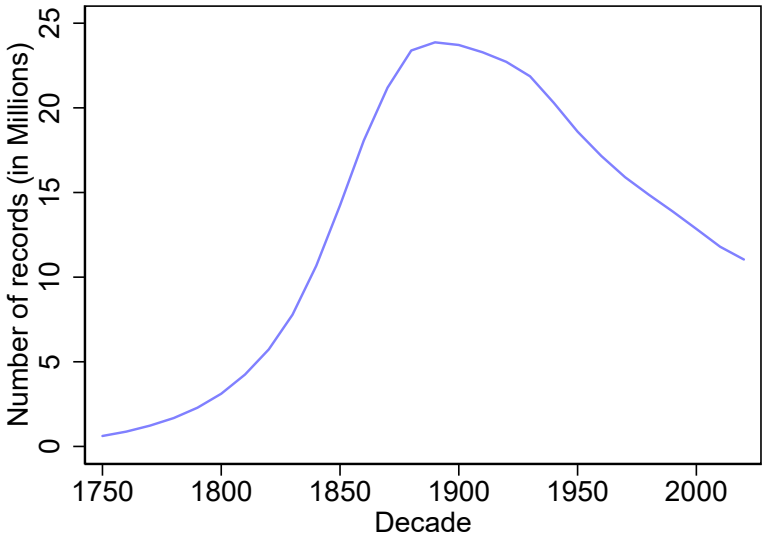
This table shows estimates for the coefficient β from equation I of the main paper estimated using OLS. Each observation is a child in the analysis sample. The outcome is the child's longevity (year of death minus year of birth). In this table we do not require the child to have survived to a certain age. Column (1) coefficients are simply the difference in mean longevity between the children of first cousins and the children of non-first cousins. Column (2) adds controls for birth year, sex, maternal age at birth, number of sisters, number of brothers, the sex ratio of siblings, and birth order, as described in section II of the main paper. Column (3) adds mother's siblings fixed effects and father's siblings fixed effects. Column (4) replaces the quadratic controls with sets of fixed effects for each integer value. Column (5) controls for parent longevity. Column (6) drops all individuals with death dates ending in 0. Column (7) adds county-by-decade-of-birth fixed effects and removes controls for birth year. In column (8), the treatment variable is equal to 1 if the child's parents have an identical surname, and 0 otherwise. We use the mother's father's surname instead of her own surname to account for the fact that she may have taken her husband's name in marriage. Standard errors are clustered at the level of the individual and their siblings.

Figure 1: Example FamilySearch Profile

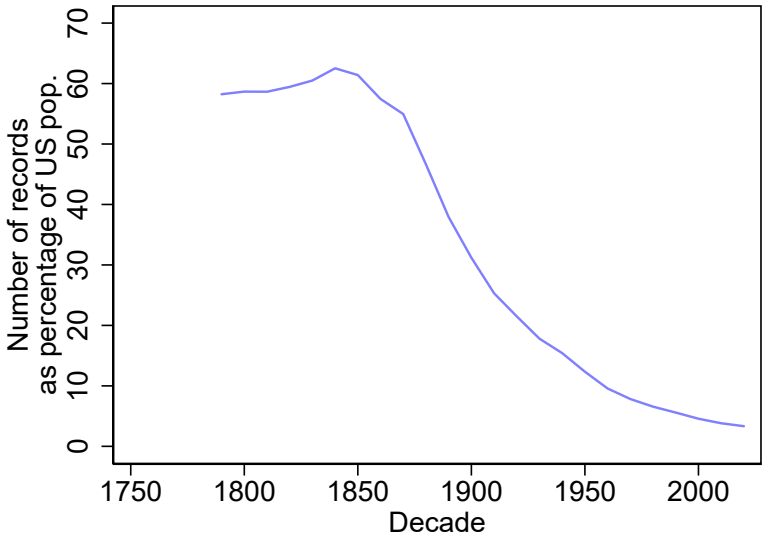


Note: This figure depicts a typical FamilySearch profile (that of Candace D. Holland). Place of birth is not shown in this example.

Figure 2: Record coverage



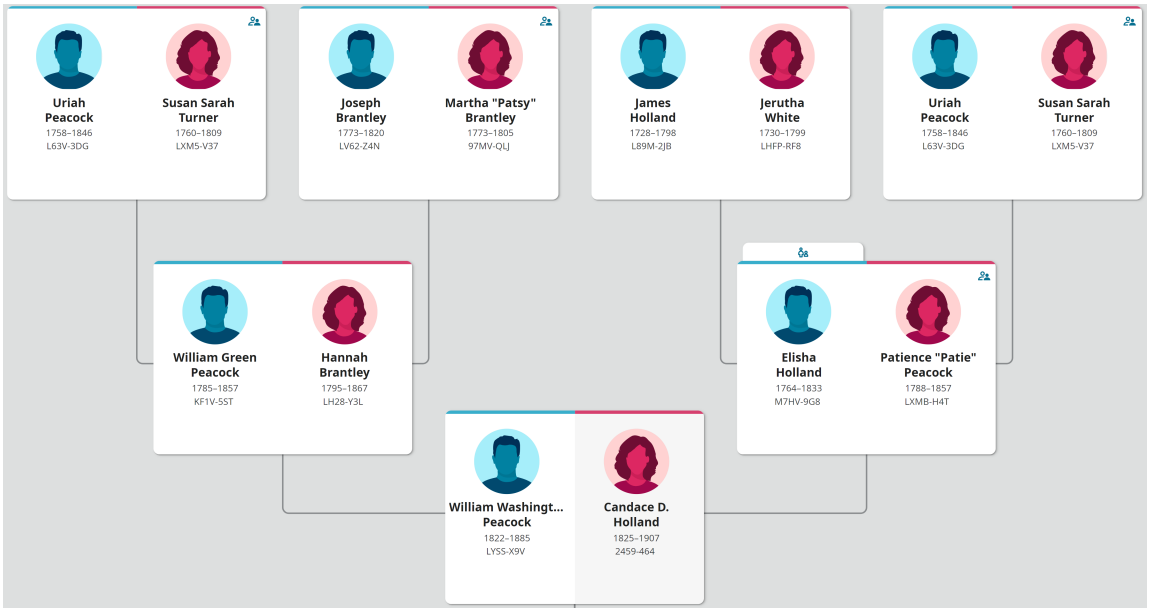
(a) Number of individuals in our dataset alive per decade



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

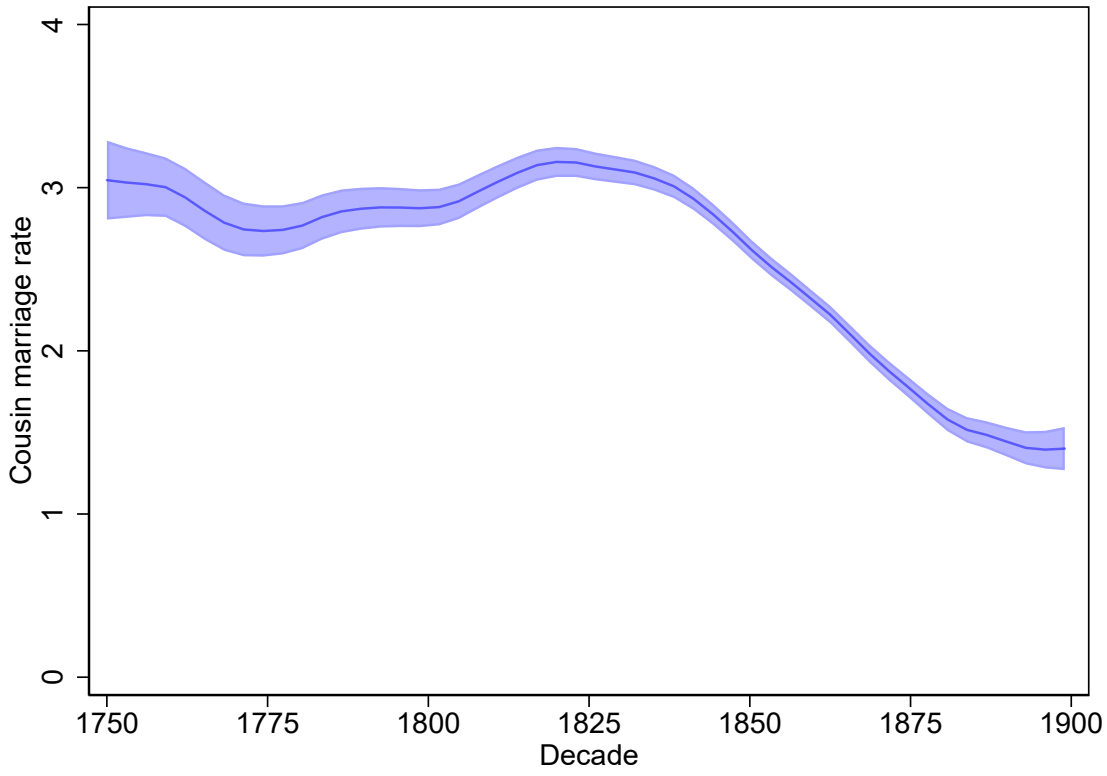
Note: Panel (a) shows the total number of records in our full dataset of 40 million individuals. An individual is counted if they were alive at any point in a given decade. Panel (b) shows these records as a percentage of the US population at the time. US population estimates come from US Census Bureau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

Figure 3: Genealogical profile of first cousin spouses



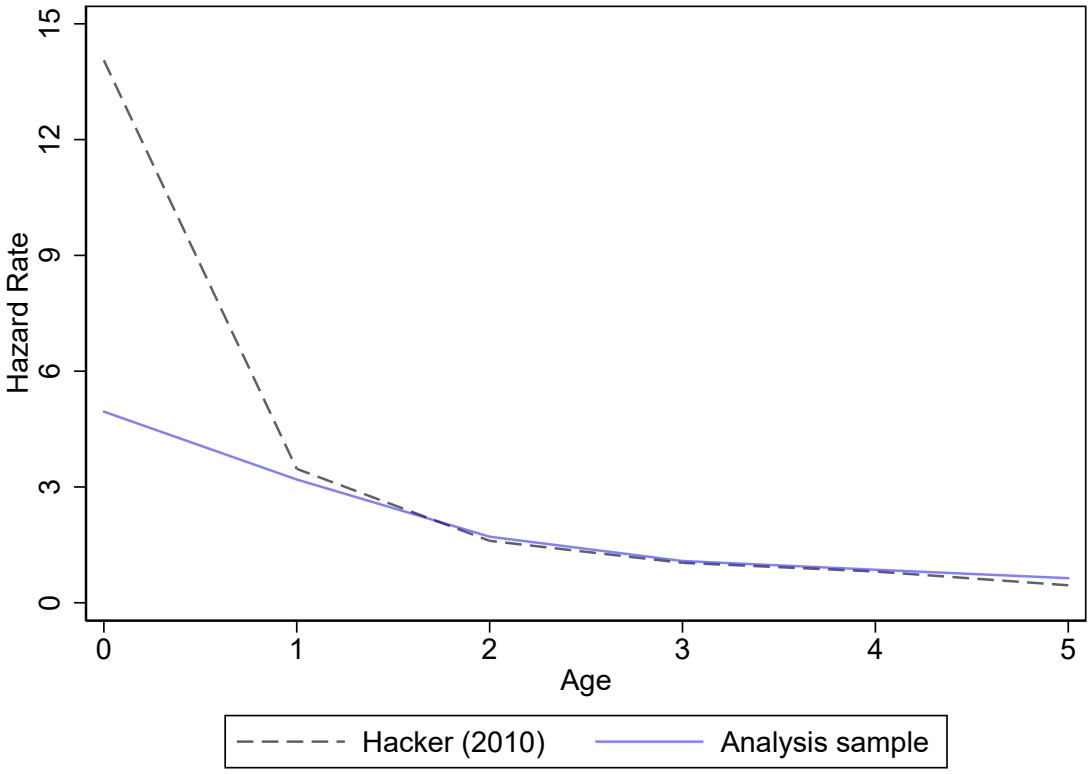
Note: This figure taken from FamilySearch shows the parents and grandparents of spouses (William and Candace) whose names and vital dates are in the bottom row. The husband's father and the wife's mother are siblings. This can be seen by observing the overlapping set of grandparents in the top row of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

Figure 4: Cousin marriage rates over time



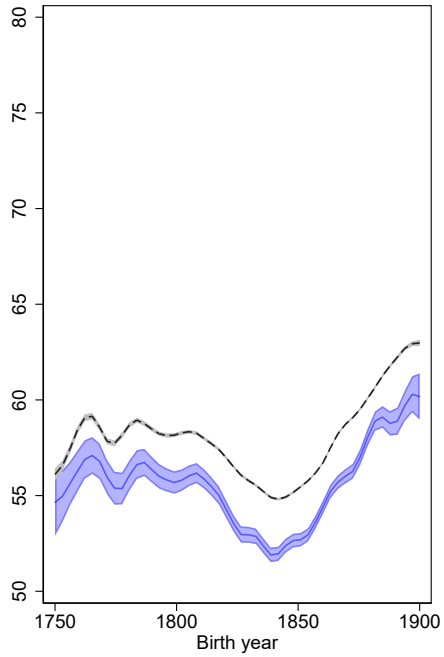
This figure depicts the share of marriages in our analysis sample of 5.9 million offspring that are between first cousins. As a proxy for year of marriage, this figure uses the year of birth of the first child born of a given union. The rate is computed by taking the number of first-born children with first-cousin parents in a given decade divided by the total number of first-born children born in that decade.

Figure 5: Mortality rates, ages 0-5

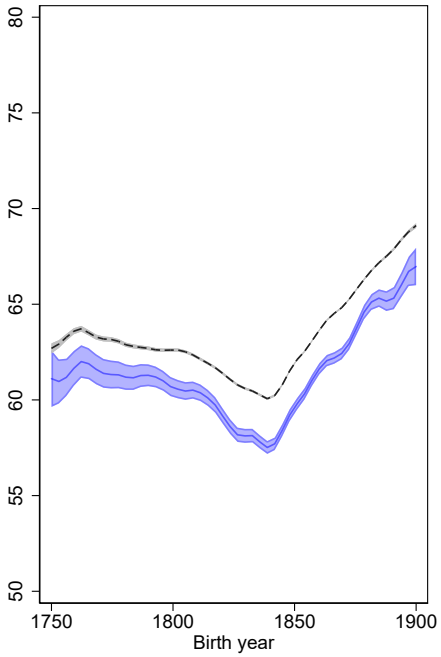


This figure depicts hazard rates for the male offspring in our analysis sample, including those who died before age 5, who died between 1880-1889. We define hazard as the percentage of individuals who die at a given age, conditional on surviving to that age. Historical longevity estimates depicted by 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.

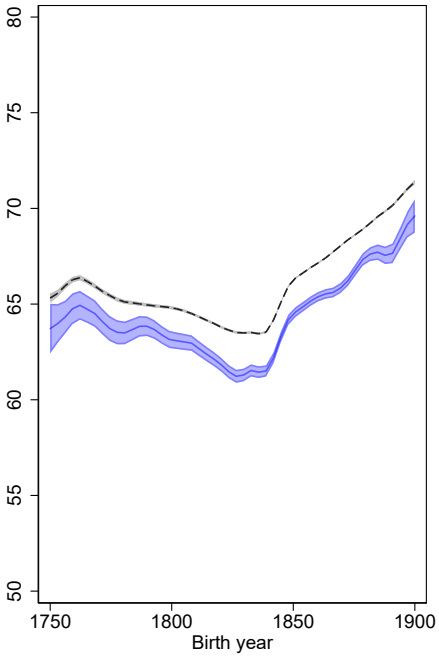
Figure 6: Life expectancy by birth cohort (at birth and at age 5, 20, and 60)



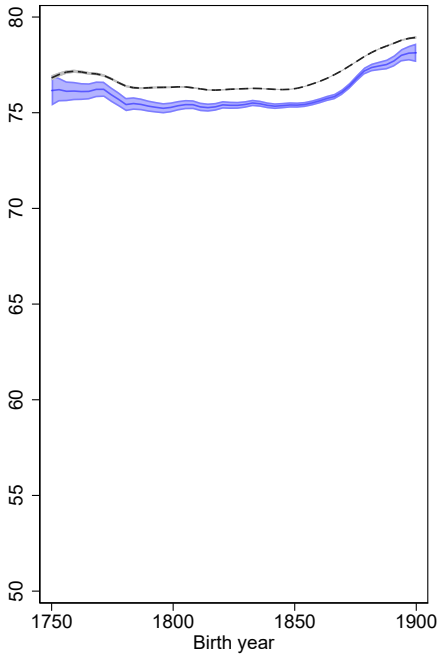
(a) At birth



(b) Life expectancy at age 5



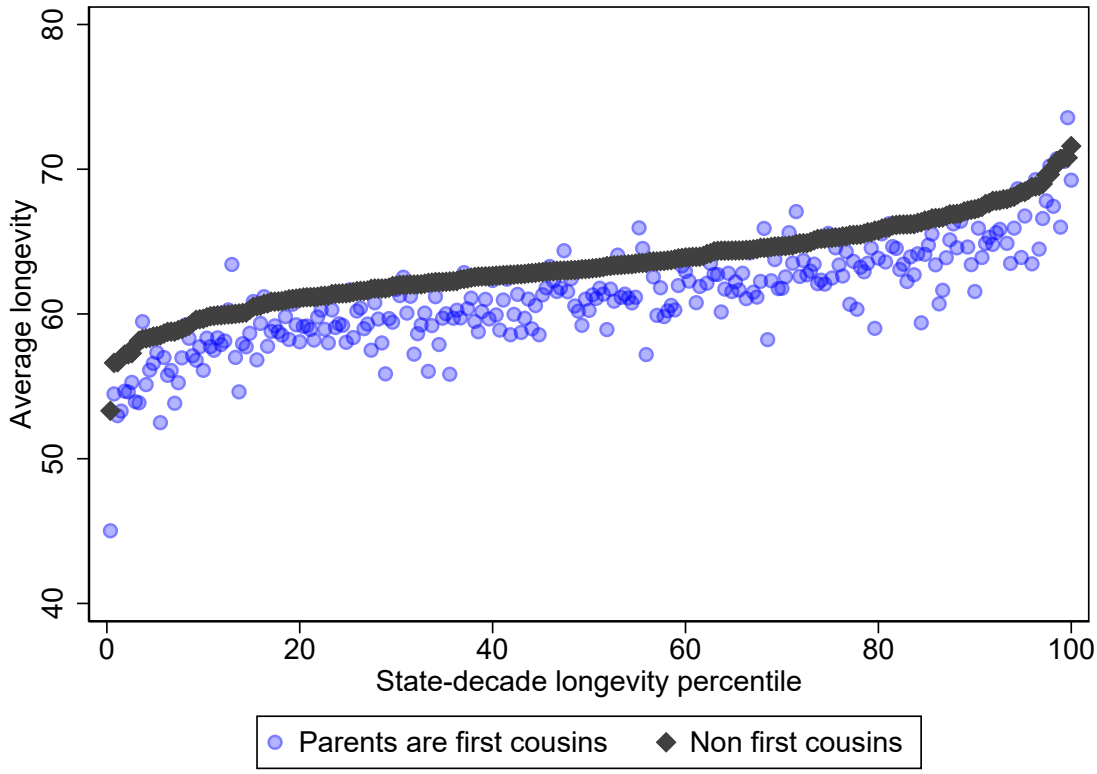
(c) Life expectancy at age 20



(d) Conditional on surviving to age 60

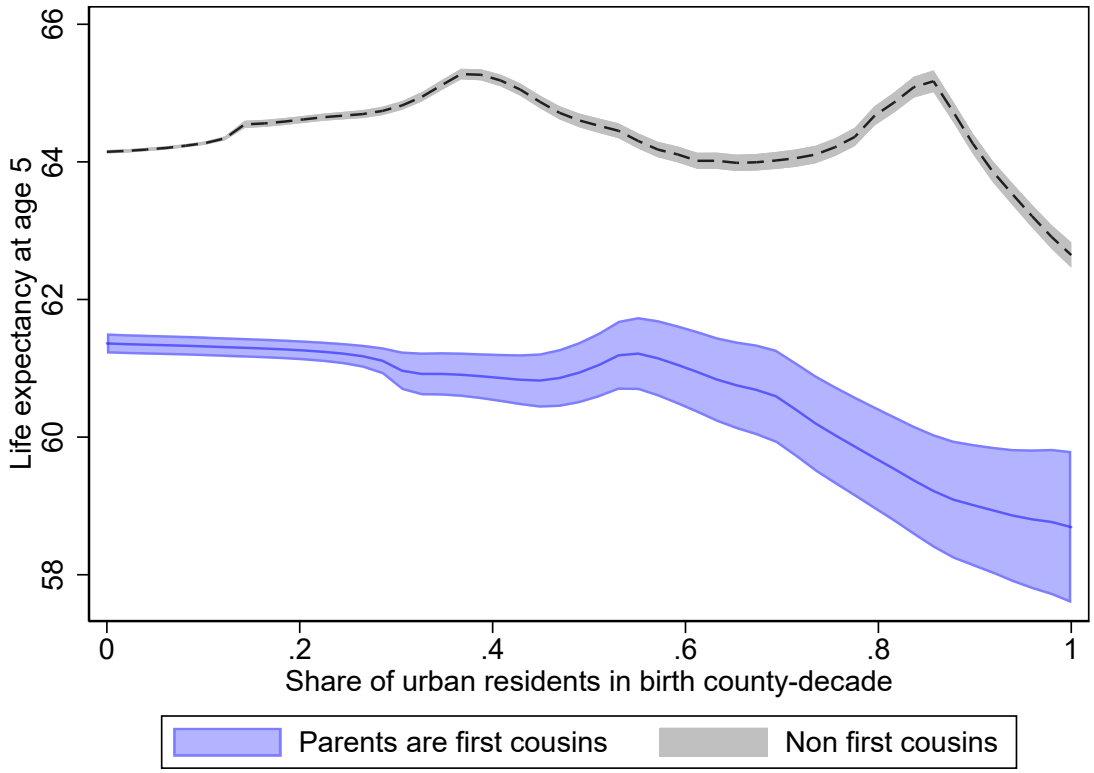
This figure depicts life expectancy of offspring in our analysis sample conditional on surviving to a specified age. Offspring of first cousins are represented by solid blue lines and offspring of non first cousins are represented by dashed gray lines. Panel (a) is a local polynomial regression of life 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. The shaded areas represent 95% confidence intervals.

Figure 7: Life expectancy at age 5 by state-decade of birth



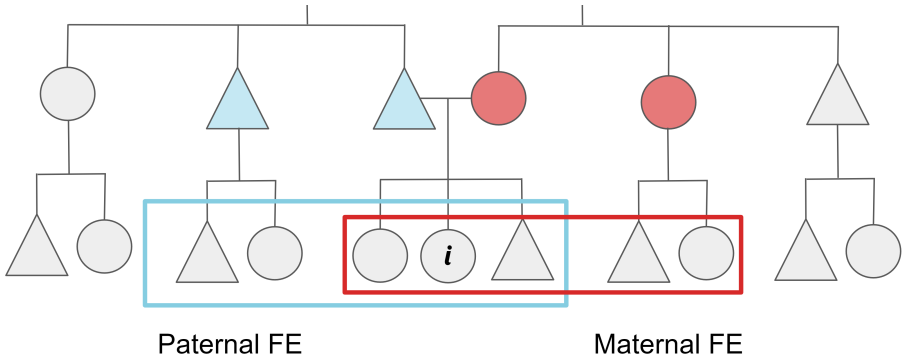
This figure depicts the average longevity at age 5 by state of birth and decade (without controls) for the 5.7 million offspring in our analysis sample for which state of birth is available. Each point is a state-decade pair. Data are sorted by the mean longevity of individuals whose parents are not first cousins in a state-decade pair.

Figure 8: Life expectancy at age 5 and share of urban residents in birth county-decade



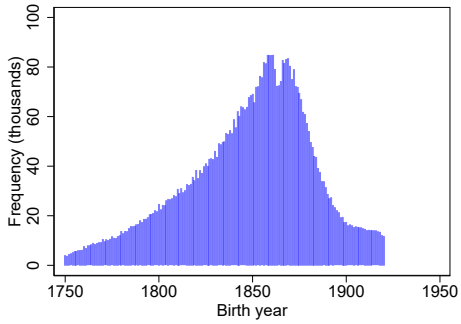
Note: This figure describes the correlation between life expectancy and the share of urban residents in one's birth county-decade. The sample for this figure consists of 4.2 million offspring in our analysis sample whose birth county is observed. The two curves shown in the figure are local polynomial regressions of life expectancy at age 5 on the share of urban residents in one's birth county-decade. The shaded areas represent 95% confidence intervals. The data on county-decade-level shares of urban residents come from Haines and Inter-university Consortium for Political and Social Research (2010).

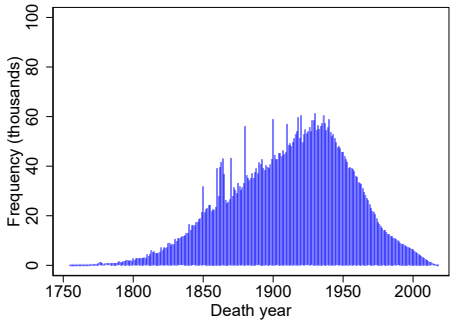
Figure 9: Empirical design



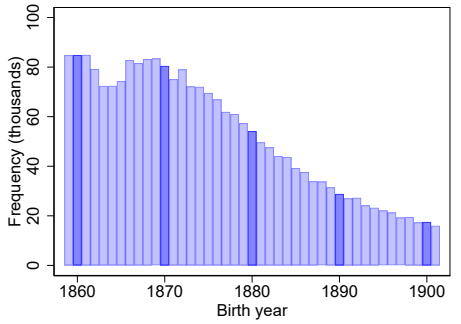
Notes: This figure visualizes the *Maternal* and *Paternal* fixed effects through two generations of related males (triangles) and females (circles). The bottom row represents the 'offspring' of married cousins or non-cousins, and represent the observations in our analysis. The blue and red rectangles 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).

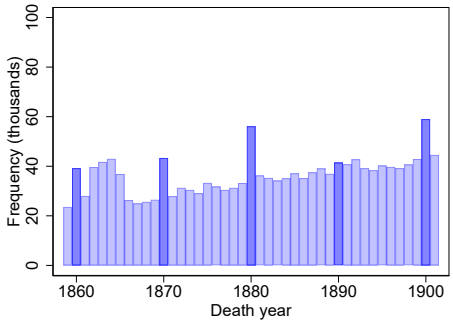
Figure 10: Data quality: birth and death year heaping



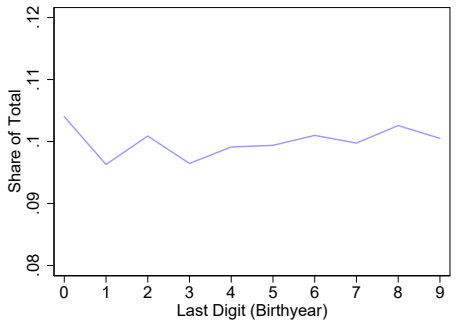


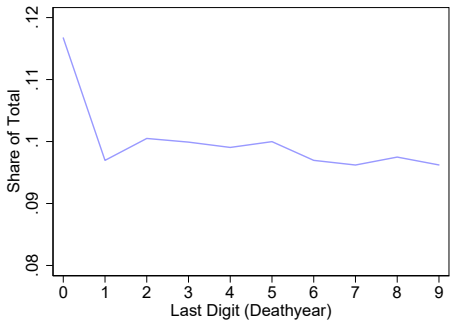
(a) Distribution of vital years in analysis sample





(b) Heaping in decadal census years





(c) Frequency of last digit in vital records

This figure describes age heaping in our analysis sample of 5.9 million offspring. 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.

References

Correia, Sergio (2015) "Singletons, Cluster-Robust Standard Errors and Fixed Effects: A Bad Mix," working paper, Duke University, http://scorreia.com/research/singletons.pdf.

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, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c821042fb1921ee37edf6ac1e591a4f300978257.

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.

Hwang, Sam Il Myoung and Munir Squires (2024) "Linked samples and measurement error in historical US census data," *Explorations in Economic History*, 101579.

Long, Jason and Joseph Ferrie (2013) "Intergenerational occupational mobility in Great Britain and the United States since 1850," *American Economic Review*, 103 (4), 1109–1137.

Ruggles, Steven, Sarah Flood, Matthew Sobek et al. (2024a) "IPUMS USA: Version 15.0, 1% Random Sample of U.S. Federal Census 1850-1930," https://doi.org/10.18128/D010.V15.0.

——— (2024b) "IPUMS USA: Version 15.0, IPUMS Ancestry Full Count Data," https://doi.org/10.18128/D010.V15.0.

Ruggles, Steven, Matt A. Nelson, Matthew Sobek, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Evan Roberts, and J. Robert Warren (2024c) "IPUMS Ancestry Full Count Data: Version 4.0," https://doi.org/10.18128/D014.V4.0.

US Census Bureau (2021) "Change in Resident Population of the 50 States, the District of Columbia, and Puerto Rico: 1910 to 2020," April, https:

//www2.census.gov/programs-surveys/decennial/2020/data/apportionment/population-change-data-table.pdf, Accessed: 2023-10-13.