Intergenerational Correlations in Longevity

Sandra E. Black
Department of Economics
Columbia University and NBER

Adriana Lleras-Muney
Department of Economics
UCLA and NBER

Nolan Pope
Department of Economics
University of Maryland

Joseph Price
Department of Economics
BYU and NBER

February 2020

Preliminary and incomplete—please do not circulate.

Abstract

While there is substantial research on the intergenerational persistence of economic outcomes such as income and wealth, much less is known about intergenerational persistence in health. We examine the correlation in longevity (an overall measure of health) across generations using a unique dataset containing information about more than eight million families obtained from Family Search family trees. The intergenerational persistence in longevity is much smaller than that of other outcomes, with a correlation of only about .10. This is surprising given the substantial education and income gradients in mortality. Moreover, this correlation remained stable throughout the 19th and early 20th centuries despite drastic changes in longevity and its determinants. This result is quite robust to a variety of specification tests. Consistent with work on the intergenerational correlation of income, we also find evidence of substantial geographic variation in this correlation, suggesting a possible role for institutions.

* This project was funded by the California Center for Population Research at UCLA (CCPR), which receives core support (P2C- HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). We are very grateful to Neil Duzett, Drew Horne, and Jacob Van Leeuwen for outstanding research assistance and to Hiram Beltran for comments. All errors are our own.

I. Introduction

A large literature has documented the substantial persistence in economic wellbeing across generations, focusing attention on outcomes such as education, income, and wealth. Much less is known about the persistence of health across generations. Previous work has documented substantial health inequality at a point in time: richer more educated individuals are healthier and live substantially longer lives in the US and most other developed countries today (Chetty et al. 2016, Cutler et al 2011, Marmot 2015, Galama et al 2018). However, there is much less research on the intergenerational persistence in health outcomes. Yet, as recent efforts in economics suggest (Jones and Klenow 2016, Becker, Philipson, and Soares 2005), health is an essential determinant of well-being, possibly as important as income. Moreover, inequality is health is also rising.[1]

In this paper, we use information from over 8 million individuals living in the United States from 1900 to 1920 and their parents to estimate the intergenerational correlations in longevity (IGCL). Correlations in longevity have several advantages over intergenerational correlations in socio-economic measures such as income or occupation: they can be computed for both genders, and are easily comparable across time and space—they do not require any adjustments for price levels or other factors. These estimates of the IGCL gives additional insight into how resources, genes, culture, and environment are transmitted from one generation to the next.

In order to estimate the IGCL, we combine United States Census records with data on family trees from FamilySearch. FamilySearch has a wiki-style family tree that includes over 1.2 billion people—the largest collection of its kind. We start with a base dataset of all individuals from the 1900, 1910, and 1920 US Census who were born between 1880 and 1920; we then match these individuals to the FamilySearch database to identify other family members. After restricting this sample to those who live to that age of 25, this data procedure results in 8.6 million individuals matched with their parents and other family members. This final matched sample is fairly representative of the US census as a whole, although blacks and immigrant are

[1] For example, the gaps between the more and the less educated have grown bigger in the US: see Meara et al. (2008), Hummer and Hernandez (2013, and Case and Deaton (2017). There is also research showing that the variance of longevity is rising (van Raalte et al. 2018).

under-represented.[2] We use data from the Social Security Administration to show that the survival rates in our sample are representative of the population survival rates.[3]

Using these data, we estimate the intergenerational correlation in longevity (IGCL) using a variety of different transformations of longevity, including lifespan, lifespan percentile, and log lifespan. We also estimate it nonparametrically. Regardless of our choice of functional form, our estimates for the IGCL for one parent and their child (e.g. father lifespan and child lifespan) range from 0.07 to 0.10. When we correlate children to both parents the estimates increase to 0.17 for both males and females. These estimates are still substantially lower than estimates of the intergenerational income elasticities in the literature that typically range between 0.30 and 0.40. When we break it by gender, we find that the IGCL tends to be largest between the same gendered parent and child, and higher for males. Our estimates are very robust to the inclusion of controls including race, cohort, location fixed effects, and individual and parent's immigration status.

We are also able to link these data to the 1940 US Census so that we have information or proxies for education, wealth, and income measures for individuals. This allows us to estimate sibling correlations not only for longevity, but also education, wealth, and income, to provide a comparison using the same dataset. Sibling correlations can be interpreted as a lower bound on the total importance of family background. When we do this, we find estimates consistent with the rest of the literature—higher correlations for education than income and much higher correlations for income than longevity. Also, interestingly the correlations in longevity among siblings are almost identical to the correlations across generations.

There are a number of early papers outside of economics that have documented the intergenerational transmission of longevity, beginning as early as 1899 with the classic work by Beeton and Pearson (1899, 1901). Compiling a variety of datasets, they find that, among those living to age 20, the intergenerational correlation in duration of life between father and son is between .12 and .14, and between brother and brother is about .26. It is striking that, despite the significantly smaller datasets (1000 father/son pairs) and the much earlier time period, the magnitudes we find are quite similar although our findings are lower for sibling correlations.

[2] We use weighted estimates for some of our analysis to make the sample representative along key gender, race, and spatial dimensions.

[3] To establish this, we produce cohort tables by state-of-birth, and compare these to those of the. We find that conditional on living to 25, our measure of longevity is fairly similar.

Other more recent work with also finds remarkably small variation over time and space in this correlation.[4] These studies' conclusions are usually limited by their small sample sizes and the fact that they are typically done with convenience samples that are not know to represent any specific population. Our study contains the largest data set ever used for this type of analysis (more than 8 million individuals), whose representativeness can be assessed because they are derived from complete census records.

There is also a burgeoning literature in economics on the intergenerational transmission of health, much of it taking advantage of recent access to administrative data. Earlier work by Currie and Moretti (2007) examined the intergenerational transmission of health at birth by correlating the birth weight of the mother and the child. Using birth records from California, they find that when a mother who is low birth weight, her child is significantly more likely to be low birth weight, even when comparing mothers who are sisters, and this correlation is stronger from mothers in high-poverty zip codes. More recently, work by Andersen (2019) uses administrative data from Denmark to document intergenerational correlations in health indices based on health conditions (around 0.1-0.15). |Using U.S. data, recent papers by Halliday et al. (2018, 2019) and on intergenerational transmission of self-reported health status or health indices and find somewhat larger correlations (around 0.2). Our conclusions are similar: mobility in health is higher than mobility in other SES measures in both the US and in Scandinavian countries.[5]

Our work builds on this foundation by incorporating new, large-scale data from the United States that enables us to look at both cross-sectional and time-series variation in this transmission. Our evidence is derived from longevity, an excellent, easy to interpret overall measure of health, which is continuous and does not require life cycle adjustments.[6] Importantly, we are also able to compare longevity correlations to those of other economic outcomes such as education and income for a constant sample.

---

[4] The most recent paper by Kaplanis et al. (2018) uses a large (N=130,000) but not clearly representative sample. Like us, it finds low correlations in longevity and no trend from 1650 to 1850. Kerber et al. (2001) and Gavrilov and Gavrilova (2001) are the two other most recent studies with relatively large sample sizes (N=78,994 and N=20,000 respectively). A complete list can be found in Appendix Table 1.

[5] There is also important new work by Bjorkegren et al. (2019) using data on adoptees in Sweden to decompose this intergenerational transmission into nature or nurture. Closely related is work by Hjelmborg et al (2006) that uses Danish, Finnish, and Sweden twins born between 1870 and 1910 to identify the genetic influences.

[6] Health increases and is overall excellent among young adults, but it then declines with age. As a result, the number of medical conditions individuals report or are diagnosed with increases steadily from age 20 onwards. As is the case with income, intergenerational correlations will be sensitive to *when* the health measure is observed. There are many imperfect ways to address this issue. Longevity is not subject to this problem.

II. Background

Assume that expected longevity follows an accelerated failure time process, so that the log of the age at death (or longevity) of a given individual *i* from family *j* born at time *t* can be expressed as

$$\mathrm{Log}\left(longevity_{ijt}\right) = \mathrm{c} + \mathrm{F}\left(\mathrm{I}_{ijt-1}, X_{jt}, longevity_{\mathrm{jt}-1}\right) + e_{ijt}$$

F is the production function of longevity which is determined by an individual's underlying health. It depends on the investment of the parents on the child I_{ijt} which include food, medicine, affection and investments in skills like education which allow the child to acquire resources that are health improving. It also depends on the home conditions and shared environment X_{jt}, which are shared by siblings in the same family. These conditions can include housing and neighborhood conditions, weather, pollution and other factors. It is also a function on the longevity of the parents because of common genetic components, and because parental longevity might directly improve on the health and wellbeing of individuals. Finally, there is an idiosyncratic error that is assumed to be orthogonal to all other components of longevity. This error term captures the fact that many deaths are accidental or hit individuals regardless of their health status or socio-economic status.

III. Data

We exploit a rich source of lifespan data that is available on a public wiki-style genealogical platform called FamilySearch. This platform includes over 13 million registered users and has profiles for over 1.2 billion deceased individuals. Most of these profiles are created by people who are doing research on their ancestors and they include information on various aspects of in the individual's life. This information includes dates and places of vital events (birth, marriage, and death); sources that are attached to the person's profile (vitals, censuses, etc.); and also links the person to the profiles of their immediate relatives (parents, siblings, spouses, and children). This platform provides a wiki-style open edit format so that anyone can make edits to any profiles, and individuals doing research on the same ancestor contribute to the same profile and have the ability to correct any errors that are made by others. The wiki nature of this platform has pros and cons but provides one of the largest collections of intergenerational lifespan data that is publicly available.

FamilySearch does not currently allow researchers to access the full corpus of profiles on the Family Tree. There is a public API that allows you to query the tree and gather information from the profiles of specific individuals. It is possible to query the tree based on information from a person's name or vital events but there is a cap on the number of profiles that match a specific query. If an API query includes more than 200 people, then no information will be shown for the query. It is also possible to use the public API to determine if an individual listed in a census record is attached to a profile on FamilySearch. This second approach, requires access to the FamilySearch version of the US census records since it requires knowing the unique FamilySearch identifier for the specific person-record observation.

We design a query of the Family Tree that uses a base sample of individuals that we gathered from the full-count US censuses for 1900, 1910, and 1920. We focused on everyone who was born between the years 1880 and 1920, which provides a base sample of 96.4 million person-year observations. Since someone born before 1900 could have appeared in all three census years, we are likely to have some individuals appear in multiple census records. We are able to find personal identifiers (PIDs) for 52.6 million person-year observations in our base sample, which is about 55% of our base sample.

After we have identified the PID for an individual from our census record, we can then use the information that is available on the profile page for that individual. We find that 16.6 million unique people from our original sample have a death date on their profile with this information being available for 15% of the person-year observations in our base sample. We also gather the PIDs for the individual's parents and siblings and then collect information about the birth and death dates of those family members. Of the people from our original sample that we have a death date for, we have a death date for at least one parent for 59% of them and a death date for both parents for 39% of them.

The Family Tree allows for multiple parental relationships. There is a marker that allows users to indicate the type of parental relationship which can include biological, step, or adopted, but this additional information is not always reported. Of the individuals in our base sample with parents, 98.4% have a single set of parents, 1.5% have two sets of parents, and less than 0.1% have three or more sets of parents. Because a large majority of individuals in our sample have only one set of parents, we use the first set of parents that is provided by the FamilySearch API. Individuals with multiple sets of parents comprise a small portion of our base sample, so using

this restriction does not have any significant effect on our sample. Additionally, the first set of parents provided by the FamilySearch API is most often the biological parents, so using the first set of parents is a reasonable restriction for individuals with multiple sets of parents.

Data Quality

Birth and death dates come from the Family Tree. An important issue is whether the longevity data derived from the Trees is accurate. The Social Security Administration (SSA) makes available cohort life tables for the social security area by year of birth and sex which we use to assess the quality of our data. Compared to the SSA data, our sample greatly underestimates mortality rates for infants and small children. This occurs either because individuals who are born between censuses and die young do not appear in a census, or alternatively, because their deaths are not noted in the Family Trees, a common problem in genealogical data (for example this is noted in Kaplanis et al. 2018). However, if we condition on individuals living to at least 25 years old, the distribution of lifespan in our data is much closer to the distribution in the SSA data. This is shown in Figure 1a, where we plot the raw lifespan histogram from our sample and for the SSA data, from age zero, and from age 25 onwards.

This is also clear from survival rates. Using the Kaplan-Meier estimate for survival analysis, we plot the survival functions of the 1900 and 1910 cohorts by gender for both the SSA data and for our sample data (Figure 1b). Our sample tends to have slightly higher life expectancy than the underlying population as reported by the SSA. Part of this difference is likely due to the fact that our final sample underrepresents African American and immigrants, two groups who may have experienced higher than average mortality rates. Another reason is that the SSA tables were limited to data from states that had Vital Registration Systems in place. They re-weight their sample in an effort to make it representative of the US, but the weighting might not be sufficient. Our data includes all states.

Sample selection

There is a natural asymmetry in the lifespan distribution of our base sample and that of their parents. The asymmetry stems from the fact that individuals don't become parents until a particular age. As a result, we never observe what the correlation in lifespan would have been for

individuals who didn't live long enough to have children. To address this asymmetry and to address the issue of missing deaths among children, we require that everyone in our sample have a lifespan of at least 25 years. We impose this restriction on both our base sample and on their parents. This is also commonly done in the study of the heritability of longevity and it is standard in studies of the intergenerational correlation of SES measures.[7]

Finally, we make two more restrictions. We restrict our sample to people with death years earlier than 2019 to avoid possible death year input errors from FamilySearch. We also require that everyone in our sample have a lifespan less than or equal to 110—for the cohorts we consider, longer lifespans are likely a result of measurement/reporting error.

Summary Statistics

Our final sample includes 8.6 million parent-child pairs with non-missing lifespan information for the individual and for both parents, and it represents about 10% of the original census data. Our data includes information from the original census record, including race, place of birth of the parent and the child, and place of residence. We also obtained the total number of siblings and the individual's birth order from the Family Tree. Finally, we computed the parents' age at birth using information from the Family Tree. The summary statistics in Table 1 indicate that nearly all of our sample is white (99.4%), and we have a very small number of individuals who are immigrants (1.1%), though 13.8% of our sample have an immigrant father and 11.5% have an immigrant mother.

In Table 2, we provide a direct comparison between the characteristics of our analysis sample and the original base sample from the census records, which confirms that we have notable under-representation in our final sample of African Americans and immigrants. Past research has had similar issues with having equal representative of these groups and many previous studies have restricted their focus to native-born Whites.[8] Also worth noting is that our data is skewed towards the Midwest and under-represents the North relative to the full census data. Our data also represents younger individuals, most likely because birth and death dates are

[7] As Appendix Table 1 shows almost all the studies we reviewed make similar restrictions.
[8] We are continuing to explore options for increasing the representation of African Americans and immigrants in our sample.

more likely missing for cohorts born earlier. We do however represent women well. They are typically under-studied in analysis of intergenerational correlations.

IV. Empirical Strategy

Our main specification relates the lifespan of the child to the lifespan of the parent. We estimate the following equation:

$$L_i^C = \beta_0 + \beta_1 L_i^P + X\beta_3 + \epsilon_i \qquad (1)$$

where L_i, our main variable of interest, is longevity in years for individual *i* and L_i^P is the longevity of one of the parents. Alternatively, L_i^P will refer to the average of the parent's longevity which is frequently done in the heritability literature. *X* refers to the set of control variables; these include a dummy variable for gender, race dummies, birth order, maternal and paternal age at birth, dummies for immigrant status of the parent, and year-of-birth dummies for the child. We present results in levels, for logs on logs, and for percentiles. We further investigate whether this functional form is correctly specified as linear. The standard errors are clustered at the family level.

The coefficient of interest is β_1. In levels, β_1 represents the increase in longevity associated with a one year increase in the longevity of the parent. When the model is estimated in logs, β_1 represents the *intergenerational elasticity* of longevity. The intergenerational correlation in longevity can then be expressed then as $\gamma = \frac{\sigma_p}{\sigma_c} * \beta_1$, where σ_p represents the standard deviation in longevity for the parents and σ_c is the standard deviation for the children. Alternatively, when *L* is converted into a percentile, β_1 represents the spearman correlation in longevity.

We do not have exogenous variation in parental longevity, so our estimates are only associations, they do not reflect causal effects. However, we will show results with and without covariates to assess whether the relationship appears to be subject to substantial omitted variable bias. We will also explore heterogeneity by interacting β_1 with other covariates and by estimating models separately by gender, cohort and state of birth to understand which factors might lead to greater or lower correlations.

We also estimate sibling correlations in longevity. To do so, we replace the parents' longevity with that of the sibling. Because we are able to observe a variety of other outcomes for this sample, we are also able to estimate sibling correlations in education, and, for those who are working, occupation and earnings. This enables us to make direct comparisons of the correlations

across outcomes on a constant sample.

Finally, we can also estimate correlations among twins. Sibling correlations are usually interpreted as providing a lower bound on the contribution of family background (Solon 1999). Identical twin correlations on the other hand put an upper bound on family background (Björklund, and Salvanes 2011). While we cannot identify identical twins, this is still a useful exercise.

IV. Results

Intergenerational Correlations

We start by estimating the correlation between the lifespans of sons and fathers and test the robustness of this relationship to the inclusion of various controls (Table 3). We find that the correlations between the lifespan of a son and his father is about 0.105, and this relationship is nearly the same regardless of whether we include additional controls for birth year, race, birth order, parent age at birth, region of the country where they live, birth state and immigrant status. Across the eight specifications that we estimate, the correlation ranges from 0.101 to 0.109 with the largest coefficient occurring when we include the most controls.[9] We get similarly robust estimates of the father-son correlation if we use a log-log specification or estimate the correlation of percentiles of lifespan.

One assumption in Table 3 is that the relationship between a son and father's lifespan is linear. To check that our linear specification is appropriate, we also visually plot the relationship between the lifespans of parents and children (Figure 2). We do this both for the lifespan and the lifespan percentile. We find that for both the lifespan and lifespan percentile, the relationship is fairly flat for those whose parents died prior to 40 years old and then becomes steeper afterwards, although remains roughly linear. We hypothesize that the flat relationship for parents who died young could be due to many of these early deaths being accidental deaths which result in the parent's lifespan having less meaningful information about the health of the parent.

In Table 4, we use the same specification as column 8 in Table 3 and extend the analysis from father/son correlations to include all of the other combinations of parent/child pairings. We also estimate these relationships using a log-log specification. For sons, we find the correlation

[9] The other coefficients in this table show the expected signs and magnitudes. Whites live longer and blacks live shorter lives than other ethnic groups. First born live longer. Conditional on making to age 25, individuals born to older mothers have a small longevity disadvantage and those born to older fathers have a very small advantage. Immigrants and the sons of immigrants live longer lives relative to natives.

with father's lifespan is larger than for the correlation with mother's lifespan (0.109 vs. 0.074). When we average the two parent's lifespans together, we find this correlation with the child's lifespan is even larger (0.169).

When we look at daughters, we find that the correlation in lifespan with fathers and mothers are about the same (0.089) and the correlation with the average of the parent's lifespans that is also much larger (0.167). When averaging across parents, there is no meaningful difference in the parental transmission of longevity to sons and daughters.

The results suggest that there is some level of transmission of lifespan from mothers that operates separately from the transmission from fathers; when we include the average of the two, we get a higher estimated coefficient than including either one alone. To this end, we also estimate the correlation in lifespan between the fathers and mothers in our sample and find that this correlation, while positive, is only 0.041. Thus, there is some amount of positive assortative mating based on health, but it is small compared to, for example, correlations in education across spouses today. Given this low correlation it is not surprising the both parents have substantial and almost equal contributions to their children's longevity.

Sibling Correlations

One challenge in comparing our intergenerational correlation of lifespan with IGC estimates for education or income is that the lower estimates that we find may stem from the specific sample that we are using which does not cover the same cohort or may not represent the same groups. In order, to estimate a more direct comparison between the correlation of lifespan and the correlation of other economic outcomes, we construct a sample of sibling pairs whose education, occupation, and income outcomes are observed in the 1940 census records at the same age. (We could compute correlations in education or income across generations bit only using parents who survived to 1940, the first year when education and income are observed).

We construct our sample of siblings by using people in our data who have the same two parents who shared the same parents and for whom both siblings are attached to the 1940 census on their profile on the Family Tree at FamilySearch. We divide our analysis here to compare siblings regardless of gender, brothers with brothers, sisters with sisters, and sisters with brothers.

The results for this sibling sample are shown in Table 5. Several conclusions emerge. First, the IGC in longevity are somewhat larger among siblings: on average they are 0.12. Similar to child-parent correlation, we see a stronger correlation between males' lifespans (0.142) than females' lifespans (0.107). And we see a rather weak correlation between brothers and sisters (0.034). Second, the correlations in income we estimate for males are comparable to those estimated in the literature which range from 0.3 to 0.4.

Third, in the same sample, we see much larger correlations for education, income or occupation for all sub-samples. The correlations for occupation or income are roughly twice as large as those for longevity. For income, they range from 0.10 for brother-sisters to 0.295 among brothers. For occupation, they range from 0.11 for brother-sisters to 0.335 for sisters. And the correlations for education are even larger ranging from 0.5 for brother-sister to 0.65 for sisters, more than four times larger than the correlations for longevity. Notably, the sister-sister coefficient for education (0.649) is higher than the brother-brother coefficient (0.539). Same gender coefficients are larger than differing gender coefficients across the board.

Due to a lack of women with an occupation (or income) score, our sister-sister pair sample is significantly smaller than our brother-brother sample. The last two columns in Table 5 do not restrict the sample to only include observations with both income and occupation scores, requiring only education and lifespan. The IGC in both education and longevity are very similar in this larger sample (though a bit smaller for sister-sister education correlations) suggesting our sample restrictions do not affect our results much.

The last rows of the table show the IGCL among twins. We identify twins as individuals that are born to the same parents in the same year. Unfortunately, we cannot differentiate between identical and fraternal twins. But we have a large sample of twins of more than 6,000 individuals with all no-missing data, and more than 15,000 with education and longevity information. Lifespan correlations are larger among twins are larger than among siblings, as are education and occupation correlations. Only income correlations are not. Nevertheless the patterns we describe for siblings persist. Lifespan correlations are low: they range between 0.15 and 0.18, and they are smaller than correlations in all other SES measures.

Heterogeneity

In Table 6, we test whether the IGCL differs based on some characteristics of the individual including their gender, the age of their father, the number of siblings, whether they are an immigrant and whether their parent in an immigrant. For this specification we combine son's and daughter's in a pooled sample and include a female dummy and control for both the father and mother's lifespan.

In the first column, where we examine differences based on the child's gender, we find that, compared to sons, daughters have a correlation with their father's lifespan that is 0.026 smaller and a correlation with their mother's lifespan that is 0.014 larger. Not surprisingly the coefficient on gender is about 6.6 showing that women live much longer lives than men. Relative to the main effects in this regression for the correlation for sons, these gaps by gender indicate that the correlation between daughters and fathers is about 26% lower than for sons and fathers and the correlation between daughters and fathers is about 19% larger than for sons and mothers. This means that there is some gender-specificity in the way that parents transmit lifespan to their children with stronger correlations for children and parents of the same gender.

Next we investigate whether the IGCL varies based on the age of the dad which is said to affect the quality of his chromosomes, though it is also correlated with greater income and wealth. We find no significant effects. In the third column we look at whether the correlation varies based on the number of children. Economic theories suggest there is a quantity-quality trade-off when investing in children. Indeed we do find that individuals in larger families live short lives, though the effect is modest, about a third of a year. The the IGCL is only very modestly affected by sib-ship size: although the interactions are positive and significant, the coefficients are very small. In the fourth and fifth columns, we investigate the effects of migration. To the extent that migrants are exposed to different environments from their children, we might expect that the IGCL are lower for migrants. We find that being an immigrant lowers longevity by 4 years, and having an immigrant parents is associated with lower longevity (about a year lower). This is not surprising since immigrants were poorer than natives. But interestingly the correlations in longevity are higher among immigrants than natives. However, these effects are again very small, though our results should be taken with a grain of salt since our sample of immigrants is likely to be very selected. Finally, we investigate if these correlations vary by birth order. Birth order has a small positive effect on longevity. But it has a large effect on the IGCL, particularly with fathers. Overall the main conclusion from this exercise is that the IGCL are

rather stable and exhibit much less variation that one would expect, with the exception of birth order.

We also investigate potential changes in the IGCL over time by estimating IGCL by birth cohort and gender. (see Figure 3). The IGCL between children and their parents' average lifespan and between daughters and mothers declines slightly to about 0.07 when comparing children born between 1880 and 1920, while the IGCL between sons and fathers remains remarkably consistent across this same time period with a slight downward trend as well. However, we cannot reject the null that there is no significant trend. To further probe this, we make use of a sample of the full family tree that spans the entire 19th century. While this sample has many limitations (we cannot assess who it represents and it is likely that the quality of the data declines as we go back in time) the results, shown in Figure 4 are remarkable: over the entire century the correlations are almost stable, again with very little change overtime.

These results are surprising. Again, they suggest little variation in the IGCL over a long period of time. Second if anything the trend is downward. One might have expected that given the decrease in external causes of death and the decline in infectious diseases, one would observe an increase in the IGCL. It is possible this decline is due to increases in technological change which extend the lives of the children beyond what would be predicted based on their parents' longevity.

Heterogeneity by place of birth

An important literature today documents the importance of place of birth and place of residence in the determination of education and mortality. Figure 4 shows the IGCL for fathers and sons by state for every decade from 1880 to 1920. As show above the correlations decline in time. But there is rather limited variation across space at a point in time.

V. Conclusion

While there is a robust literature examining the intergenerational correlation in a broad range of economic outcomes such as income and wealth, much less attention has been paid to the correlation in health, primarily due to an absence of data. In this paper, we use information from over 8 million individuals living in the United States from 1900 to 1920 and their parents to estimate the intergenerational correlations in longevity. Consistent with research in other fields,

we find very small correlations in longevity, with father son correlations in the range of .10. Taking advantage of data on the family structure, we use estimates of sibling correlations to show that these low correlations in longevity are in stark contrast to the correlations of other outcomes such as education, occupation, and income using the same sample and methods. Importantly, we find no trend in this correlation over time, suggesting little evidence of aggregate effects of technological and medical advances over this time period. Future work will use spatial variation in the intergenerational transmission of longevity to try to tease out the underlying causes driving this relationship.

References

Andersen, C. (2019). Intergenerational health mobility: Evidence from Danish registers. *Economics Working Papers 19-04, Aarhus University*.

Beeton, M., Pearson, K., 1899, "Data for the problem of evolution in man, II: A first study of the inheritance of longevity and the selective death rate in man", Proceedings of the Royal Society of London. 65: 290-305.

Becker, Gary S., Scott Duke Kominers, Kevin M. Murphy, and Jörg L. Spenkuch, "A Theory of Intergenerational Mobility," Journal of Political Economy 126, no. S1 (October 2018): S7-S25.

Beeton, M., & Pearson, K. (1901). On the inheritance of the duration of life, and on the intensity of natural selection in man. *Biometrika*, *1*(1), 50-89.

Björklund, A., & Salvanes, K. G. (2011). Education and family background: Mechanisms and policies. In *Handbook of the Economics of Education* (Vol. 3, pp. 201-247). Elsevier.

Björkegren, E., Lindahl, M., Palme, M., & Simeonova, E. (2019). Pre-and Post-Birth Components of Intergenerational Persistence in Health and Longevity Lessons from a Large Sample of Adoptees. *Journal of Human Resources*, 0318-9421R1.

Case, A., & Deaton, A. (2017). Mortality and morbidity in the 21st century. *Brookings papers on economic activity*, *2017*(1), 397-476.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., ... & Cutler, D. (2016). The association between income and life expectancy in the United States, 2001-2014. *Jama*, *315*(16), 1750-1766.

Currie, J., & Moretti, E. (2007). Biology as destiny? Short-and long-run determinants of intergenerational transmission of birth weight. *Journal of Labor economics*, *25*(2), 231-264.

Cutler, D. M., Lleras-Muney, A., & Vogl, T. (2011). *Socioeconomic status and health: dimensions and mechanisms*. The Oxford Handbook of Health Economics, Glied, S. and Smith P. (ed.) Oxford University Press 2011: pp. 124-163

Galama, T. J., Lleras-Muney, A., & van Kippersluis, H. (2018). *The Effect of Education on Health and Mortality: A Review of Experimental and Quasi-Experimental Evidence* (No. w24225). National Bureau of Economic Research.

Gavrilov, L., & Gavrilova, N. (2001). Biodemographic Study of Familial Determinants of Human Longevity. Population: An English Selection, 13(1), 197-221.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., ... & Bhatia, G. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, *360*(6385), 171-175.

Kerber, R. A., Obrien, E., Smith, K. R., & Cawthon, R. M. (2001). Familial Excess Longevity in Utah Genealogies. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences,56(3).

Hummer, Robert A. and Elaine M. Hernandez (2013) "The effect of educational attainment on adult mortality in the United States," Population Bulletin, 68(1), 1-16.

Jones, C. I., & Klenow, P. J. (2016). Beyond GDP? Welfare across countries and time. *American Economic Review*, *106*(9), 2426-57.

Halliday, T. J., Mazumder, B., & Wong, A. (2019). The intergenerational transmission of health in the United States: A latent variables analysis. *Health Economics*.

Halliday, T., Mazumder, B., & Wong, A. (2018). Intergenerational health mobility in the US.

Lindahl, Mikael; Evelina Lundberg; Marten Palme; and Emilia Simeonova. (2016). Parental Influences on Health and Longevity: Lessons from a Large Sample of Adoptees. NBER working paper.

Marmot, M. (2015). The health gap: the challenge of an unequal world. *The Lancet*, *386*(10011), 2442-2444.

Meara, Ellen R., Seth Richards, and David M. Cutler. 2008. "The Gap Gets Bigger: Changes in Mortality and Life Expectancy, by Education, 1981–2000." Health Affairs 27(2):350–60.

Piraino, P., Muller, S., Cilliers, J., & Fourie, J. (2014). The transmission of longevity across generations: The case of the settler Cape Colony. Research in Social Stratification and Mobility, 35, 105-119

Solon, G. (2014). Theoretical models of inequality transmission across multiple generations. *Research in Social Stratification and Mobility*, *35*, 13-18.

Solon, G. (1999). Intergenerational mobility in the labor market. In *Handbook of labor economics* (Vol. 3, pp. 1761-1800). Elsevier.

van Raalte, A. A., Sasson, I., & Martikainen, P. (2018). The case for monitoring life-span inequality. *Science*, *362*(6418), 1002-1004.

Table 1. Summary Statistics of Analysis Sample

	Census based sample matched to FS, cohorts born 1880-1920		
	Full Sample	Sons	Daughters
Average Lifespan	72.729	70.096	75.84
	(16.898)	(16.18)	(17.198)
Father's Average Lifespan	72.954	72.928	72.985
	(12.429)	(12.424)	(12.435)
Mother's Average Lifespan	73.379	73.357	73.404
	(14.701)	(14.686)	(14.718)
Birth Year	1900	1900	1900
White	0.994	0.994	0.994
Black	0.004	0.004	0.004
Place of birth and ancestry			
Northeast	0.119	0.118	0.119
Midwest	0.411	0.411	0.41
South	0.351	0.353	0.349
West	0.118	0.116	0.12
Immigrant Status	0.011	0.011	0.011
Father's Immigrant Status	0.138	0.137	0.138
Mother's Immigrant Status	0.115	0.114	0.115
Family characteristics			
Siblings	2.664	2.632	2.701
	(2.216)	(2.204)	(2.230)
Birth order	2.31	2.301	2.322
	(1.610)	(1.602)	(1.620)
Age of mother at birth	29.025	29.029	29.02
	(6.640)	(6.632)	(6.649)
Age of father at birth	33.8	33.799	33.802
	(7.786)	(7.782)	(7.791)
Observations	8,623,844	4,670,496	3,953,348

Notes: The estimates in this table compare individuals who were age 25 or older in one of the US censuses from 1900-1920 for whom we have information about their own lifespan and the lifespan of both of their parents. Standard deviation in parentheses.

Table 2. Who is represented in the matched sample? Comparison of Full Census Sample and Matched Sample

Variable	Matched Analysis Data	Full Census Sample	Difference*
Female	0.463	0.458	0.05
Birth Year	1900	1875	25
White	0.994	0.859	0.135
Black	0.004	0.128	-0.124
Northeast Census Region	0.119	0.308	-0.189
Midwest Census Region	0.411	0.304	0.107
South Census Region	0.351	0.29	0.061
West Census Region	0.118	0.086	0.032
Immigrant Status	0.011	0.246	-0.235
Father's Immigrant Status	0.138	0.407	-0.269
Mother's Immigrant Status	0.115	0.389	-0.274
Occupation Score 1900	1.349	9.887	-8.538
Occupation Score 1910	3.235	10.298	-7.054
Occupation Score 1920	4.746	9.083	-4.337
Observations	8,623,844	87,700,136	

Notes: The estimates in this table compare individuals who were age 25 or older in one of the US censuses from 1900-1920 based on whether or not we were able to match the individual to information on their lifespan and the lifespan of both of their parents.
*Each difference in statistically significant at the 5% level.

Table 3. Estimate of Intergenerational Correlation of Lifespan for Sons and Fathers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Father's Lifespan	0.104	0.108	0.103	0.105	0.102	0.103	0.101	0.109
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Birth Year		0.096						0.096
		(0.001)						(0.001)
White			0.571					1.094
			(0.155)					(0.156)
Black			-1.615					-0.915
			(0.206)					(0.208)
Birth order				0.228				0.104
				(0.005)				(0.006)
Maternal age				-0.039				-0.039
at birth				(0.001)				(0.001)
Paternal age				0.005				0.004
at birth				(0.002)				(0.002)
Northeast Region					-1.197			-0.892
					(0.031)			(0.031)
Midwest Region					-0.446			-0.155
					(0.025)			(0.025)
South Region					-1.482			-1.19
					(0.025)			(0.026)
Region Missing					2.225			1.852
					(0.228)			(0.266)
Immigrant Status						0.124		0.191
						(0.0724)		(0.0763)
Father's Immigrant Status						0.454		0.41
						(0.0224)		(0.0232)
State of Birth FE							X	X
R-squared	0.006	0.011	0.006	0.007	0.008	0.006	0.006	0.012

Notes: The same includes 4,670,496 observations in all regression specifications. Standard errors in parentheses. Column 7 includes dummies for state of birth if the person was born in the US, or for country of birth if the person was born abroad.

Table 4. IGCL for Varying Child and Parent Pairings and Measures

	Outcome		
Model	Lifespan (Years)	Log Lifespan	Observations
Son/Father	0.109	0.101	4,670,496
	(0.001)	(0.001)	
Son/Mother	0.074	0.063	4,670,496
	(0.001)	(0.001)	
Son/Parents' Average	0.169	0.167	4,670,496
	(0.001)	(0.001)	
Daughter/Father	0.089	0.078	3,953,348
	(0.001)	(0.001)	
Daughter/Mother	0.086	0.073	3,953,348
	(0.001)	(0.001)	
Daughter/Parents' Average	0.167	0.160	3,953,348
	(0.001)	(0.001)	
Father/Mother	0.041	0.041	8,623,844
	(0.001)	(0.001)	

Notes: Each cell is a separate regression that provides the estimated correlation in lifespan or log lifespan between the two individuals indicated in the row. Each regression includes all of the same controls as column 8 in Table 3.

Table 5: Sibling Correlations for outcomes in the 1940 census compared to lifespan

Model	Education	Income	Occupation	Lifespan	Education	Lifespan
Panel a: siblings						
Sibling/Sibling	0.544	0.238	0.267	0.126	0.526	0.105
	(0.001)	(0.019)	(0.001)	(0.001)	(0.001)	(0.001)
Obs	999,922	999,922	999,922	999,922	2,794,151	2,794,151
Sister/Sister	0.649	0.214	0.335	0.107	0.565	0.105
	(0.005)	(0.084)	(0.006)	(0.006)	(0.001)	(0.001)
Obs	28,866	28,866	28,866	28,866	598,536	598,536
Brother/Brother	0.539	0.295	0.302	0.142	0.534	0.141
	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
Obs	748,328	748,328	748,328	748,328	834,084	834,084
Brother/Sister	0.507	0.107	0.113	0.034	0.504	0.0411
	(0.002)	(0.004)	(0.002)	(0.002)	(0.001)	(0.001)
Obs	222,728	222,728	222,728	222,728	1,361,531	1,361,531
Panel b: Twins						
Twin/Twin	0.684	0.233	0.366	0.182	0.638	0.153
	(0.0116)	(0.092)	(0.0123)	(0.0137)	(0.00742)	(0.00834)
Obs	6,143	6,143	6,143	6,143	15,776	15,776

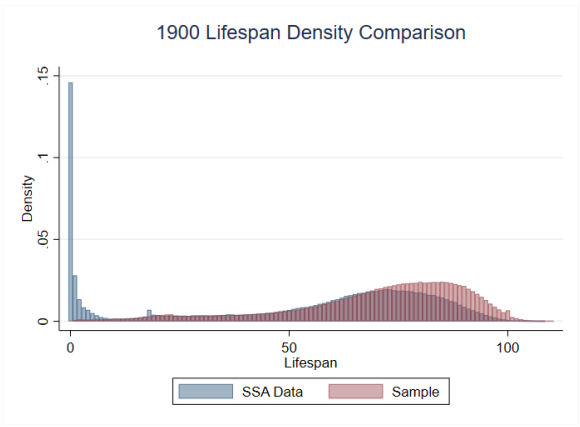
Notes: Each cell in this table is a separate regression that provides an estimate for the correlation on the outcome between different types of siblings. In the first four columns, we only use sibling pairs for which information on all four outcomes is available for both siblings. Since occupation and income is often missing for women in the 1940 census, we include the last two columns and we restrict the sample to just those sibling pairs for which both education and income is available. Twins are determined by existing sibling pairs that share birth year.

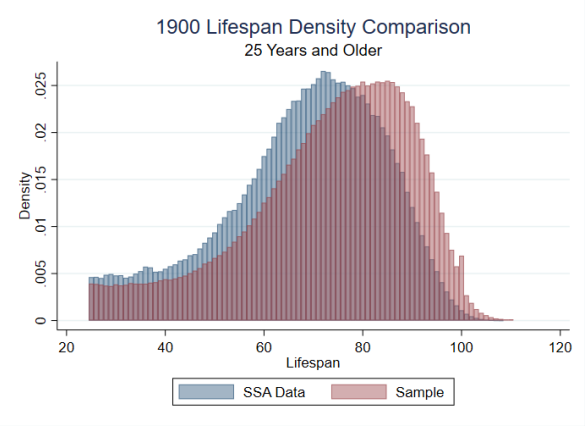
Table 6: Heterogeneity by gender, SES and immigrant status

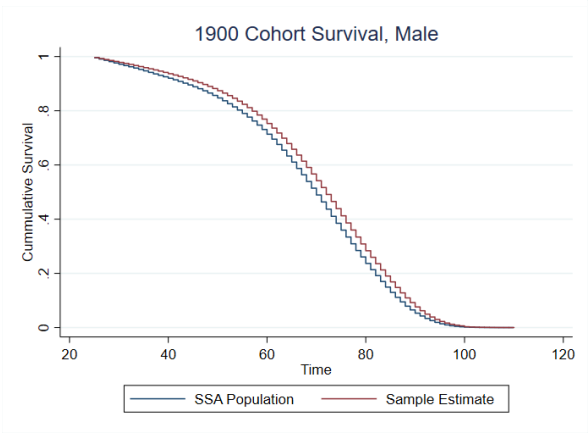
	Child's Lifespan					
Variable	X=female	X=age of dad	X=number of siblings	X=immigrant	X=immigrant parent	X=birth order
Father's Lifespan	0.099	0.102	0.077	0.088	0.087	0.08398
	(0.001)	(0.002)	(0.001	(0.000)	(0.001)	(0.00081)
Mother's Lifespan	0.072	0.078	0.073	0.078	0.077	0.07903
	(0.001)	(0.002)	(0.001)	(0.000)	(0.000)	(0.00069)
X	6.59	0.018	-0.329	-4.01	-0.865	0.18084
	(0.088)	(0.006)	(0.021)	(0.434)	(0.124)	(0.02879)
Father's Lifespan*X	-0.026	-0.000	0.004	0.023	0.007	0.02879
	(0.001)	(0.000)	(0.000)	(0.005)	(0.001)	(0.00130)
Mother's Lifespan*X	0.014	0.000	0.002	0.029	0.009	-0.00037
	(0.001)	(0.000)	(0.000)	(0.004)	(0.001)	0.00025
Observations	8,623,844	8,623,844	8,623,844	8,623,844	8,623,844	8,623,844

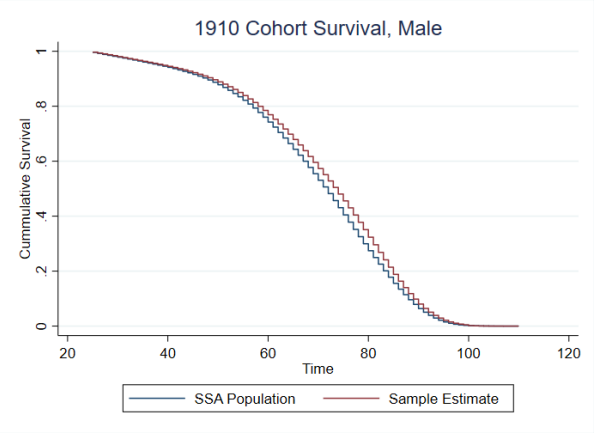
Notes: Each column is a separate regression. The heading for each column provides the variable that we include in place of X and also interact with the father's and mother's lifespan measures.

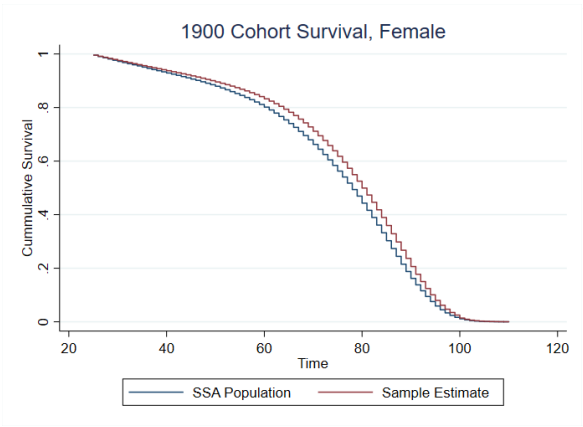
Figure 1. Comparison with SSA Longevity Data

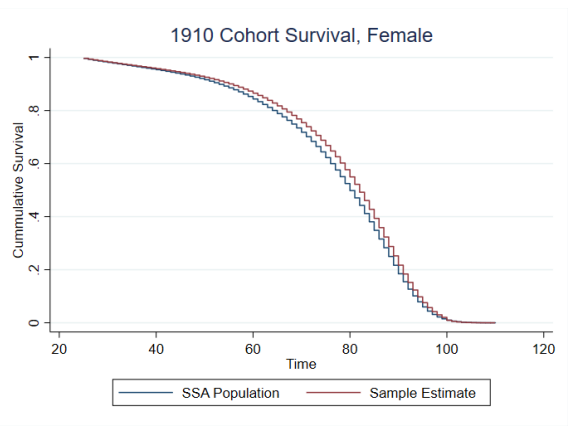






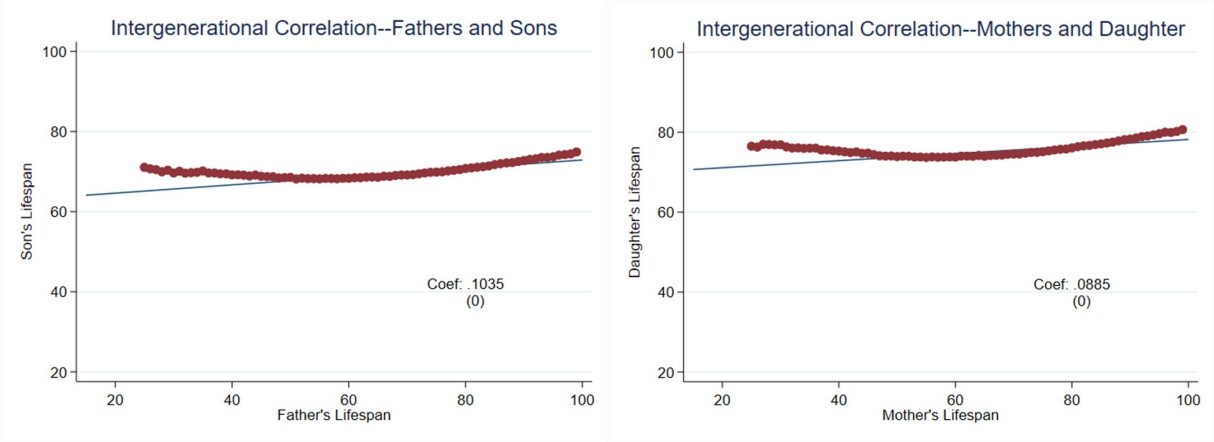






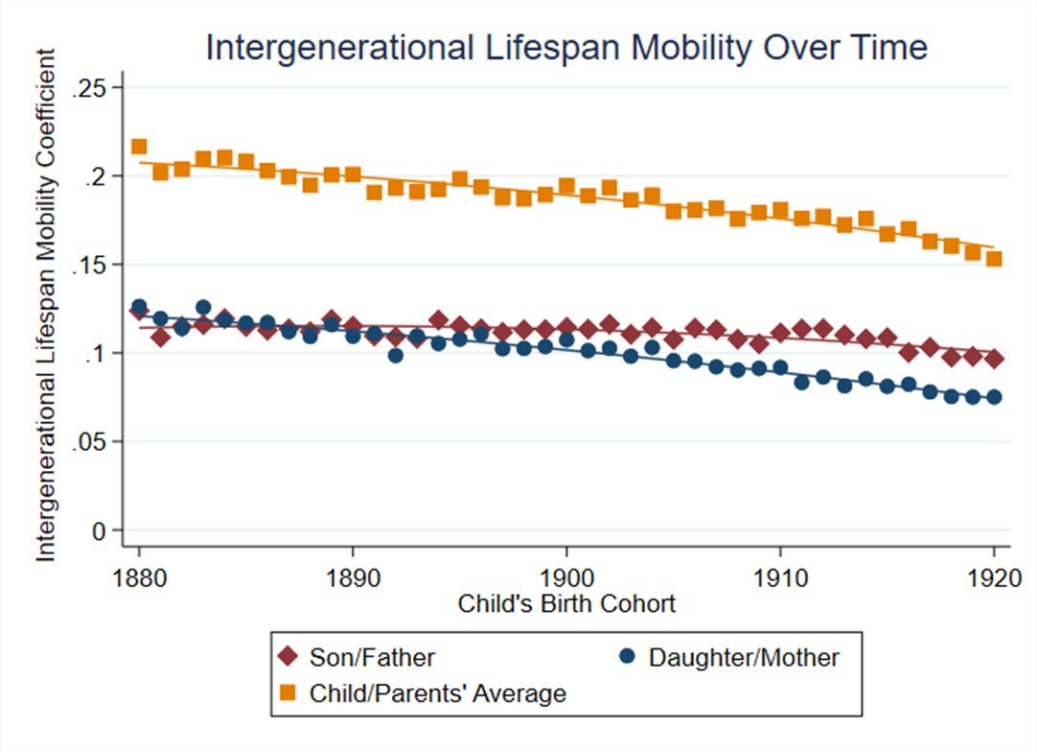
Notes: these figures use our sample derived from the Family Tree (see text for details) and cohort life tables produced by the Social Security Administration (SSA), available here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/. Kaplan-Meier estimates are produced using the methods described here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

Figure 2. Test for Linearity of Correlation (Father/Son and Mother/Daughter Lifespan)



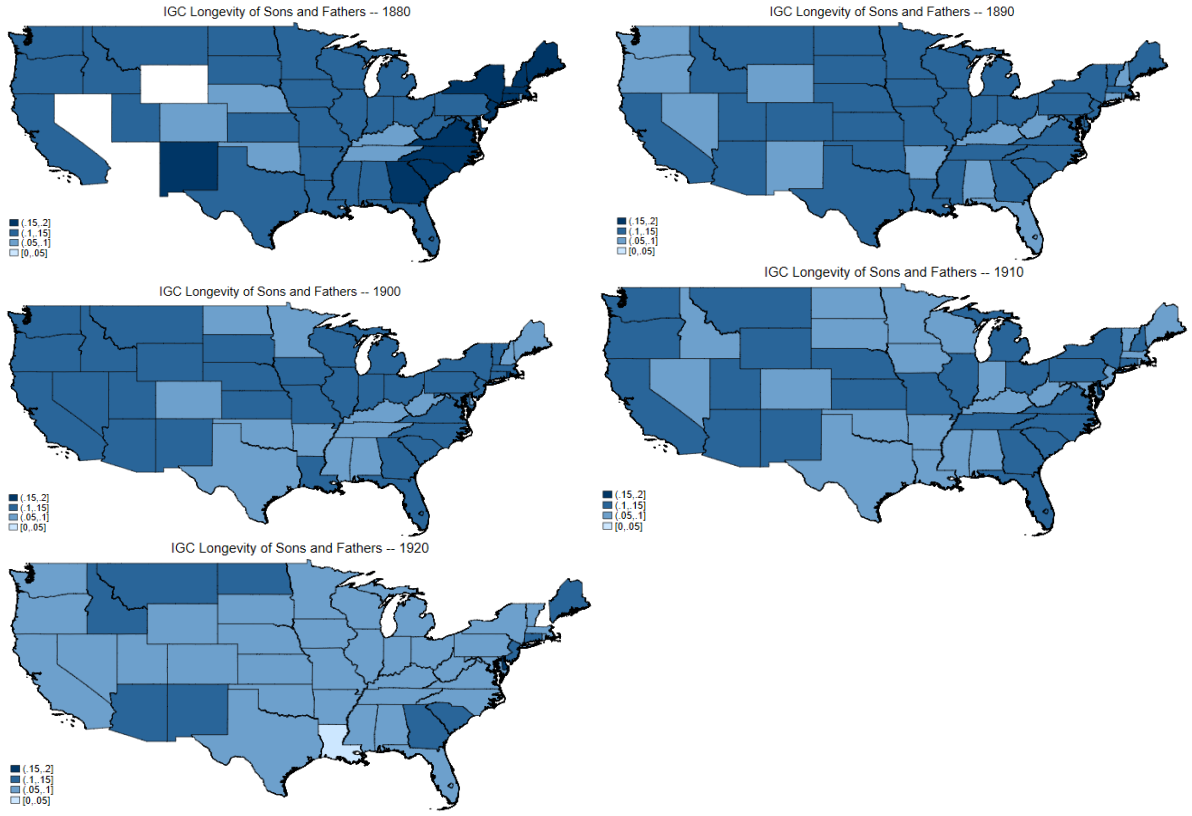
Notes: This figure provides average of the son's lifespan in one-year bins based on the father's age. The relationship between father and son lifespan is nearly linear starting at age 40 for fathers. This might reflect the fact that deaths before 40 are more likely to be driven by factors that are less in the control of the individual.

Figure 3. Changes in the IGCL over time



Notes: In this figure, we split our sample based on the birth cohort of the child and estimate the intergenerational correlation of lifespan for son/father pairs, daughter/mother pairs, and the correlation between the child's lifespan and the average lifespan of both parents.

Figure 4. Heterogeneity by birthplace



Notes: The years in the titles are determined by birth year, with 10 year bins centered on the year in the title. 1880 and 1920 have 5 year bins. For 1880, Arizona and Wyoming are missing, presumably from lack of statehood. Nevada and Vermont are missing because their coefficients had a p value greater than 0.05. For 1920, New Hampshire is missing because its coefficient had a p value greater than 0.05.