The Healthy Migrant Hypothesis in the USA at the Turn of the 20th Century:

Investigating Origin and Destination Effects Using Crowdsourced Online Genealogies

Saverio Minardi*, Paul Puschmann+, Nicola Barban*
*University of Bologna, + Radboud University Nijmegen

Starting from the 1980s, several studies have documented that migrants tend to experience a survival advantage compared to non-migrant populations. This advantage—also known as "healthy immigrant paradox", "migrant mortality advantage", or "healthy migrant effect"—has been observed in diverse contexts and among various migrant groups. Researchers have primarily attributed this health advantage to a selection process, implying that more healthy individuals are more likely to migrate. However, the precise underlying mechanisms of this selection remain uncertain, partly due to the difficulties of collecting appropriate data on migrant populations across nations.

This paper investigates the presence and extent of a migrant mortality advantage in the US for cohorts born between 1850 and 1890, comparing migrants to 1) the general population in the receiving country, 2) the population in the origin country, and 3) their non-migrant siblings in the origin country. Analyses leverage recently crowdsourced online genealogies, which allow to reconstruct family histories beyond national borders and locate individuals' birth and death places as well as those of their relatives. Using genealogical data allows to investigate novel and underexplored aspects of the migration-mortality nexus.

First, most studies compare migrants' health to that of the receiving population, and few others compare them to the origin country. However, due to data limitations, few studies simultaneously compare them to both groups. Nevertheless, both contexts influence the health of migrants, and both populations are relevant for comparison. Migrants share early life conditions and lifestyles with the origin population but later life conditions and environmental opportunities and constraints with the destination population. Genealogy data allows to investigate migrants' mortality effects with respect to both origin and destination.

Second, very few studies have explored what characteristics drive migrants' selection. There has been limited research examining the healthy migrant effect from a family-oriented perspective since it is usually hard or even impossible to collect international data on migrants' families from the sending region. The most valuable insights into familial influences on migrants' mortality advantage come from historical studies on internal migration (Mourits and Puschmann, 2023). Here, we investigate whether international migrants have an advantage compared to their non-migrating siblings, exploring whether selection occurs *within* rather than across families.

Third, most studies examine migrant populations as a whole, underestimating the important heterogeneity that may exist between migration groups based on countries of origin. Indeed, nationality of origin can play a crucial role in many ways. First, migrants from different countries are exposed to very different early life conditions and customs influencing their health outcomes later in life. Second, migrants from different nationalities undergo different treatment in the receiving country, different levels of discrimination, occupational segregation and access to labour markets, and locate in different geographical areas. Clearly distinguishing between nationalities of origins allows us to shed light on the possible mechanisms leading to a migrant advantage.

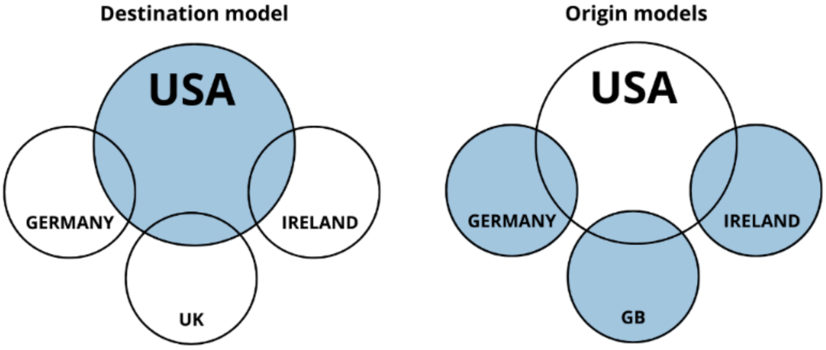
Data and methods

We use online crowdsourced genealogies from the Familinx project that originated from the website Geni.com and subsequently tested and cleaned for consistency (Kaplanis et al., 2018). We limit our sample to individuals with non-missing values on places of birth and death, birth year, and death year. We restrict the birth cohort from 1850 until 1890 to include individuals more likely to have migrated before more restrictive immigration laws were installed.

We further limit our sample to individuals who lived at least until the age of 40 and not over 110. The lower bound is meant to limit the inner bias of genealogical data, which tends to exclude childless individuals since they are less likely to have living descendants. As a result, people who died before reproductive age tend to be underrepresented, and mortality levels from genealogies are generally lower. However, recent research has shown that mortality rates above age 30 tend to be fairly consistent with available official statistics for the same periods (Chong et al., 2022). We have compared the genealogical data to the existing official statistics with respect to our research question and conclude the same: the genealogical data are fairly representative for the purposes for which we use them.

We focus on the five most numerous migrant groups according to the US censuses: Germans, Canadians, Irish, British, Italian and Scandinavians (from Sweden, Norway, Finland, and Denmark), and a residual category containing all other nationalities. We build different samples depending on the reference population. A stylized representation of the baseline samples is presented in Figure 1. The sample for the *destination model* includes migrants to the US and US-born individuals. The *origin models* are tested on six samples, one for each origin category, including the same migrants reported in the destination sample and non-migrants in the corresponding origin country. To allow comparison, migrants of the same nationality to countries other than the US are excluded. The third sample compares migrants to the US to their non-migrant siblings in their home country. This is a subset of the origin-model sample, which only includes migrants to the US with at least one sibling and their non-migrant siblings.

Figure 1 Stylized representation of samples for origin and destination models



We estimate three sets of models depending on the reference population and distinguishing by gender. For each sample, we regress age at death (lifespan) on dichotomous variables indicating migrant status or nationality of origin controlling for birth cohort. In this way, we obtain a double estimate for each nationality group, one for origin and one for destination, in which the migrants are the same, but the reference population changes. Finally, using the sibling samples, we regress migration status on lifespan separately for each origin group, including siblings' fixed effects, therefore controlling for all unobserved characteristics at the family level.

Preliminary results

Figure 2 descriptively explores variation in lifespan conditional on reaching age 40 by birth cohort (1950-1900) and for each migrant group disaggregated by migration status and reference population. The first relevant observation is that for more than one nationality, migrants did not experience a survival advantage and, in some cases, experienced a disadvantage compared to the US-born. These relationships changed over time. This is confirmed by looking at the evolution of period life expectancy computed from the same data showing the absence and reversal of the healthy migrant advantage during the first half of the twentieth century. The same pattern has recently been documented using available official statistics (Bakhtiari, 2022).

Heterogeneity across groups is more clearly visible in Figure 3, which reports the average difference of each migrant group relative to the origin and destination populations, disaggregated by gender. A mortality disadvantage relative to the US-born was present for the Irish, British, and Other non-specified groups. On the other hand, when comparing to the origin population, all groups had a mortality advantage with two exceptions: male Canadians, which show a limited disadvantage, and male Irish, which show a substantial one.

Figure 4 reports results, including siblings fixed effects. Overall, the direction of the coefficients is consistent with the baseline results, suggesting that the selection process is not based on family-level characteristics—such as family geographical origin, family social status, or family health—but rather on individual factors—such as personality or individual health. Nevertheless, results from these models should be taken cautiously as only a limited number of non-migrant siblings could currently be included in the sample.

Figure 2 Trends in average lifespan by birth cohort, migration status, and reference population (1850-1900)

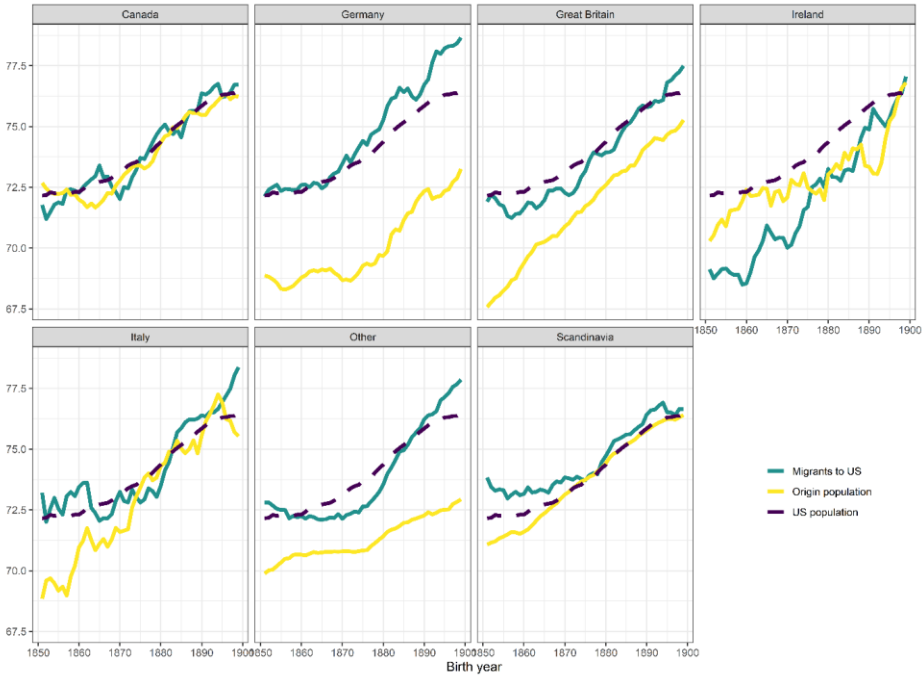
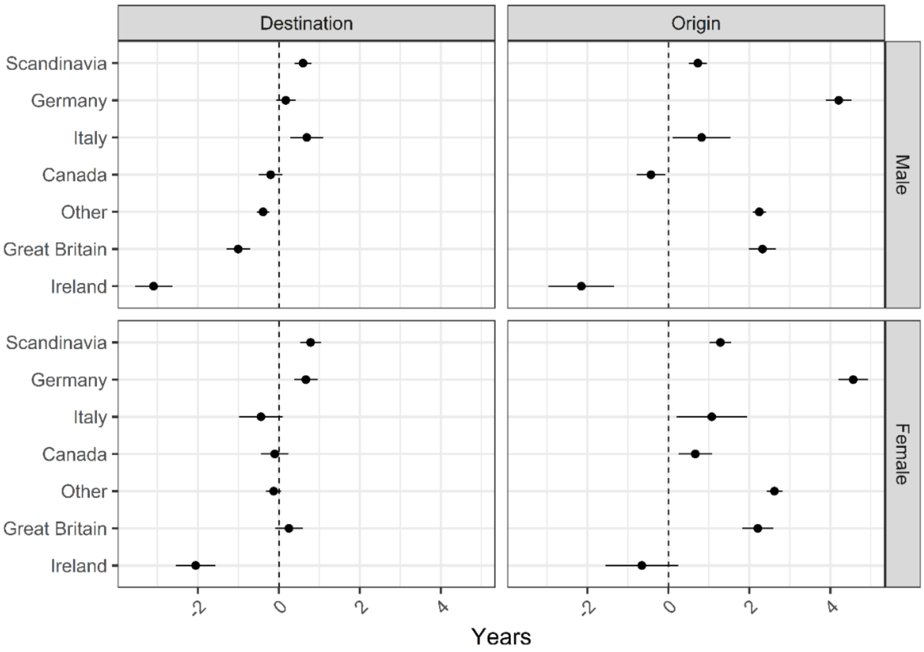


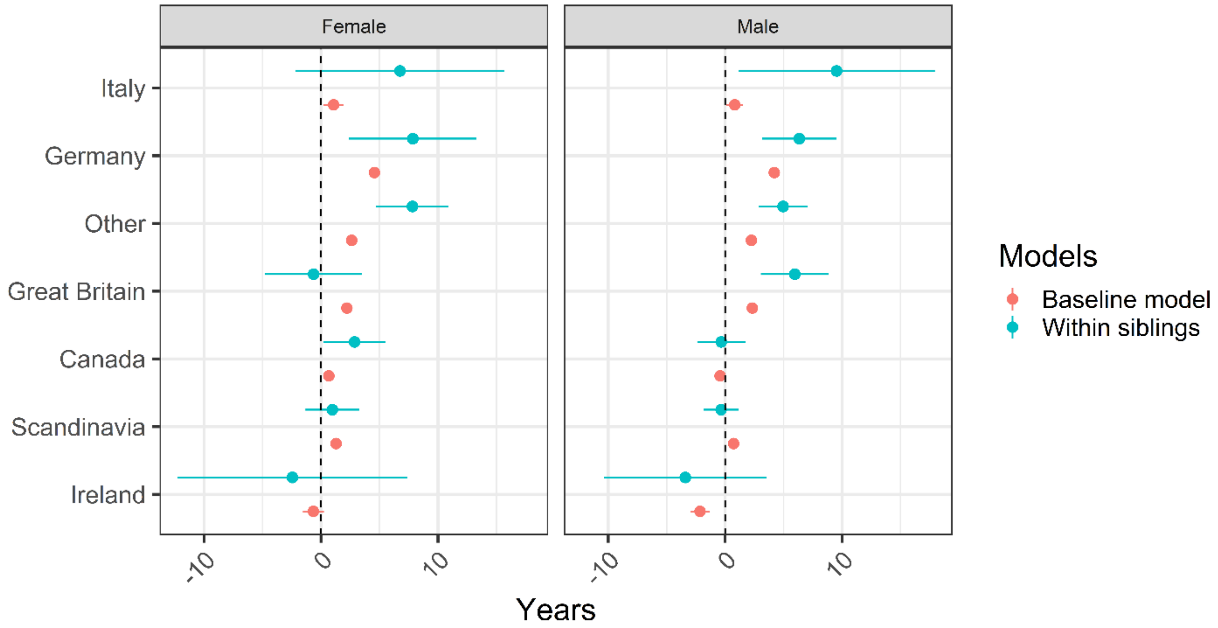
Figure 3 Migrants' lifespan difference from non-migrant population in origin and destination countries by gender and migrant nationality



Discussion

Findings highlight that migrant mortality advantage is not a historical constant but shows high heterogeneity across periods, nationalities, and genders relative to origin, destination population, and even family members. Moreover, results show that migrants' health advantage or disadvantage is not necessarily consistent when comparing them to their origins or destination population. Finally, results shed light on possible mechanisms driving selection from origin countries, highlighting that migrants experience a health advantage also compared to their siblings. In other words, migrants are not selected based on family-level characteristics but on individual ones.

Figure 4 Average marginal effect of migration by country of origin estimated from siblings fixed-effects model



Even though ascendant genealogies have frequently been used for historical demographic research, they are not free of limitations. Given that genealogies are collected backwards from descendants, they underrepresent childless individuals. At the same time, genealogies are more likely to include individuals of higher socio-economic status and education since they are more likely to be remembered and to have preserved their family tree and history. These concerns are certainly important but less relevant as far as biases are similar between migrants and non-migrants. In these respects, comparisons with existing statistics from official sources are consistent with genealogies and similarly show an absence of a healthy migrant effect in the early 20th century. We further compared trends in the share of migrants by nationality to the US censuses. The comparison showed that statistics from the genealogies are consistent with the census for all origin groups except for the Irish, who were highly underrepresented. We further compare life expectancy from the genealogies to estimates from the Human Mortality Database and Hacker (2010). Comparisons show that genealogies tend to overestimate life expectancy for the total population, but they are consistent with other sources regarding overall historical time trends.

Next steps

The next steps will involve more precisely identifying the mechanisms driving differences between origin groups with the aid of other data sources, most importantly the US censuses. We will investigate the differences between origins groups regarding geographical location, occupation, education, and other characteristics within the US. We further plan to extend the analysis to broader periods and test different model specifications. We will further validate the data by more detailed comparison with official statistics from origin and destination countries.

References

Bakhtiari, E. (2022). The missing mortality advantage for European immigrants to the United States in the early twentieth century. Demography, 59(4), 1517-1539.

Chong, M., Alburez-Gutierrez, D., Del Fava, E., Alexander, M., & Zagheni, E. (2022). Identifying and correcting bias in big crowdsourced online genealogies. Max Planck Institute for Demographic Research.

Hacker, J. D. (2010). Decennial life tables for the white population of the United States, 1790–1900. Historical methods, 43(2), 45-79.

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

Mourits, R. J., & Puschmann, P. (2023). Exploring familial factors in the migrant mortality advantage among domestic migrants in later life: Zeeland, the Netherlands, 1812–1962. SSM-Population Health, 22, 101359.