Internet crowd-sourced genealogy and intergenerational transmission of demographic behaviours.

Nicola Barban, *University of Oxford, Nuffield College*

September 2016

Paper presented for the annual meeting of the Population Association of America, 2017.

1 Introduction

Online genealogy databases contain large-scale information about millions of people and their past and present family relations (Fire and Elovici, 2015). Most of these datasets are based on collaborative work of genealogy enthusiasts around the world who documented and shared their family connections by filling the information on different websites. To our knowledge, these data have rarely been used in a scientific context to study the evolution of demographic behaviours across time.

In this paper, we used data from one of the largest genealogy project: the FamiLinx initiative (`www.familinx.org`), carried out by the Erlich lab at Columbia University. FamiLinx is a scientific resource of organized genealogical data from tens of millions of people, mostly from the last 500 years. Based on a crowd-source approach, FamiLinx combines the public information available on `Geni.com`, a genealogy-driven social network that is operated by MyHeritage. Geni.com is a website that allow genealogists to enter their family trees into the database and to create profiles of family members with basic demographic information such as sex, birth date, marital status, and location. Users can decide whether they want the profiles to be public or private. New or modified family tree profiles are

constantly compared to all existing profiles, and if there is high similarity to existing ones, the website offers the users the option to merge the profiles and connect the trees. FamiLinx is a database created for scientific purposes that contains public profiles of individuals from Geni.com.

During the Population of America Association presidential address of 2010, Robert Mare stressed the importance of taking a multi-generational approach in the study of intergenerational transmission of demographic behaviour and social stratification (Mare, 2011). Most of the studies that look at the resemblance of demographic behaviours among generations use a "Markovian model" in which children characteristics depends on the same characteristics of the previous generation. However, recent studies show long term influence of distant ancestors and kin (Song et al., 2015) on intergenerational transmission of social class. Grandparents' characteristics or even more remote generations may have direct effect on grandchildren. Genealogy datasets represent a unique source of information on multigenerational demographic characteristics such as birth, age of death, marriages and migration. The obvious drawback of genealogy data is their retrospective nature and the non-representativity of their information. Information on distant kin may be underreported, genealogists may have very few information on distant ancestors and presumably more information are available for higher social status families. Nevertheless, differently from historical archives that provide information on family relationships on specific geographical levels (for instance parish records), genealogy data provide transnational information as well as migration events. Last, genealogy data provide information on many members of the same family tree, making possible to derive genetic relatedness to study heritability of demographic characteristics, such as longevity (Gavrilov et al., 2002) and fertility (Tropf et al., 2015).

In this paper, we use this novel dataset to address two key questions in demographic research. First, we describe the degree of intergenerational transmission of demographic characteristics such as longevity and examine whether this has changed with time. Second,

we study the long-term effects of migration on demographic characteristics. Do descendent of migrant family live longer than descendent of non-migrant families? Do migrants have more descendents of non-migrants? To address these questions compare families for which some siblings migrated and others did not, in order to account for specific family effect. In particular, I will use family fixed effect models (grandparents fixed effect and great-grandparent) to study differences in longevity and family size among cousins and great-cousins.

2 Data

The FamiLinx database contains basic demographic information on 43,589,566 individuals, derived by 12,080,102 "founders", that is, individuals who have no parents in the database. For 4,351,044 individuals, FamiLinx contains information on year of death and year of birth. Burial location (or birth location for those alive) is available for 5,404,864 individuals. This information has been transformed by the FamiLinx team into geographical coordinates and converted into country (based on current political borders) and continent. In total, the database has information on more than 51 million parent-child relationships that spans over multiple generations.

To investigate the role of migration, we define a migrant as an individual whose burial location (or birthplace) is in a different country with respect to the location of her/his parents. Using the relationship informations provided by the study, family trees are recursively reconstructed in order to identify multiple generation relationships.

3 Preliminary analysis

Figure 1 shows the mean age of death reported in the FamiLinx dataset by individual year of death (individual died after year 1700). The figure describes an almost linear increase in mean age at death starting from 1900, with the exception of the two world wars of the

last century. It is possible also to notice that female mean age at death, became higher than male starting approximately from twentieth century.

FIGURE 1 HERE

3.1 Intergenerational transmission of longevity and fertility

Figures 2 and 3 show the correlation between fathers and sons on age at death and number of offsprings. Preliminary results show overall positive relationship between parental age at death and children' longevity. One interesting results is that correlation on longevity is quite irregular for most of the historical period of interest and increases only in the last century. On the other hand, the correlation on number of offsprings decreases almost monotonically until the last part of twentieth century. For this initial analysis, we adopted a patrilinear model and restricted the analysis to men only.

FIGURE 2 HERE

3.2 Long term effects of Migration

Table 1 and 2 show the results of OLS and Fixed-Effects regression models in which we examine the effects of migration events on longevity and number of descendants. We defined individuals as migrants if their country location is different from the location of their parents. We run two sets of regression on a male-only sample. We first run OLS regression with clustered standard error where we control for year of birth. Then, we use a fixed effect model in which we compare siblings (Father fixed effect model). In this model we analyze the differences in longevity and number of offsprings of siblings who migrated compared to siblings who did not. In the third set of models, we compare cousins (Grandfather fixed effect model). This last models allow us to compare different outcomes of children of immigrants, compared to individuals whose parents did not migrate. Using fixed-effect

model make it possible to take into account possibile bias and selection issues relative to the family of origin.

The preliminary results show that migrants live longer than non-migrant. However, once we control for family fixed effect (both at father and grandfather level), differences in longevity are not statistically different.

Table 2 shows the results of different regression models on number of offspring. OLS models indicate that migrant families have less offspring. However, the results are completely reversed once we control for family fixed effect. Second generation immigrants have more offspring than their cousins whose parents did not migrated.

TABLES 1 AND 2 HERE

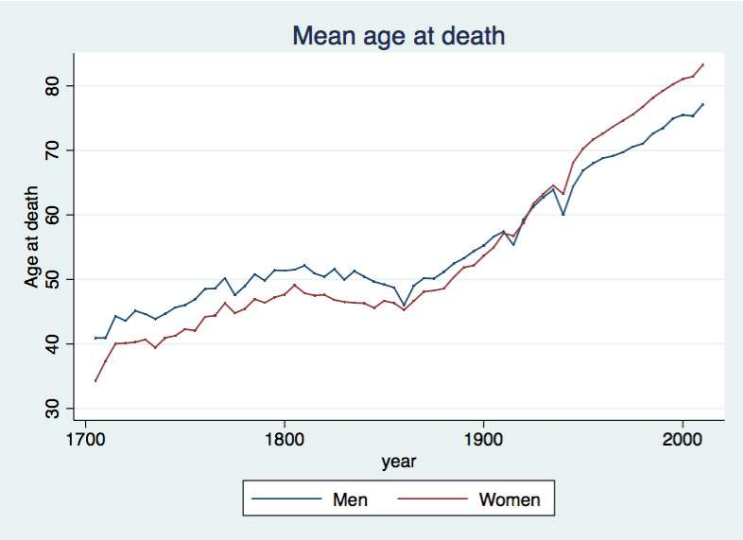

Figure 1: Mean age at death by death year

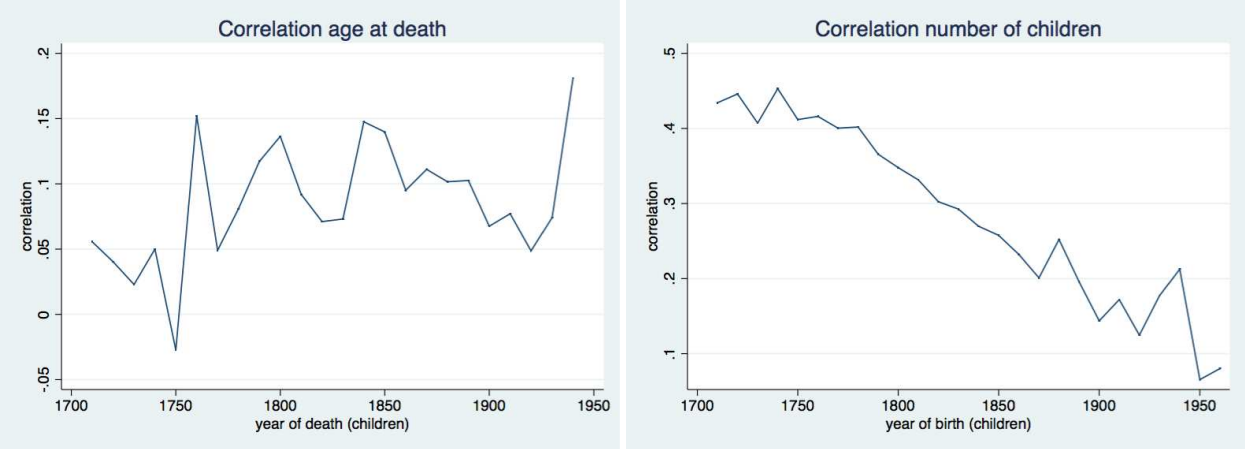

Figure 2: Correlation of age at death and number of descendants among fathers and sons.

Table 1: Regression table on longevity

Age at Death	(1) OLS	(2) Father FE	(3) Grandfather FE
Parental age at death	0.190***		0.123***
	(0.0164)		(0.0241)
Own migration	1.558	1.414	-1.907
	(0.933)	(1.232)	(1.491)
Parental migration	2.428**		2.953
	(0.785)		(2.968)
Constant	62.59***	58.32***	48.86***
	(3.931)	(0.104)	(1.687)
Observations	33819	73062	33820

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

OLS clustered SEs at family level. Controls: 10-years dummies for birth-year

Table 2: Regression table on number of descendants

Number of children	(1) OLS	(2) Father FE	(3) Grandfather FE
Parental number of children	0.302***		0.0402***
	(0.00648)		(0.00790)
Own migration	-0.0132	0.220	0.0840
	(0.0922)	(0.127)	(0.127)
Parental migration	-0.0342		0.709***
	(0.0681)		(0.197)
Constant	0.711***	3.666***	3.418***
	(0.0923)	(0.0104)	(0.0398)
Observations	51741	81939	56373

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

OLS clustered SEs at family level. Controls: 10-years dummies for birth-year

References

Fire, M. and Y. Elovici, 2015. Data mining of online genealogy datasets for revealing lifespan patterns in human population. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(2):28.

Gavrilov, L. A., N. S. Gavrilova, S. J. Olshansky, and B. A. Carnes, 2002. Genealogical data and the biodemography of human longevity. *Social Biology*, 49(3-4):160–173.

Mare, R. D., 2011. A multigenerational view of inequality. *Demography*, 48(1):1–23.

Song, X., C. D. Campbell, and J. Z. Lee, 2015. Ancestry Matters Patrilineage Growth and Extinction. *American Sociological Review*, page 0003122415576516.

Tropf, F. C., N. Barban, M. C. Mills, H. Snieder, and J. J. Mandemakers, 2015. Genetic influence on age at first birth of female twins born in the UK, 1919–68. *Population studies*, 69(2):129–145.